T.C.
EGE ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

DERİN ÖĞRENME TEKNİKLERİ KULLANILARAK KÜÇÜK VERİ SETLERİ İÇİN GÖRÜNTÜ ALTYAZISI ÜRETME

İlayda YILDIZ

Danışman: Prof. Dr. Aybars UĞUR

Bilgisayar Mühendisliği Anabilim Dalı
Bilgisayar Mühendisliği Doktora Programı

İzmir
2025

İlayda YILDIZ tarafından DOKTORA tezi olarak sunulan "DERİN ÖĞRENME TEKNİKLERİ KULLANILARAK KÜÇÜK VERİ SETLERİ İÇİN GÖRÜNTÜ ALTYAZISI ÜRETME" başlıklı bu çalışma EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliği ile EÜ Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi'nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve .....tarihinde yapılan tez savunma sınavında aday oybirliği/oyçokluğu ile başarılı bulunmuştur.

Jüri üyeleri		**İmza**
Jüri Başkanı	: Prof. Dr. Aybars UĞUR	
Raportör Üye	: Doç. Dr. Şebnem BORA	
Üye	: Doç. Dr. Korhan KARABULUT	
Üye	: Prof. Dr. Aylin KANTARCI	
Üye	: Dr. Öğr. Üyesi Osman GÖKALP	

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

ETİK KURALLARA UYGUNLUK BEYANI

EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliğinin ilgili hükümleri uyarınca Doktora Tezi olarak sunduğum "DERİN ÖĞRENME TEKNİKLERİ KULLANILARAK KÜÇÜK VERİ SETLERİ İÇİN GÖRÜNTÜ ALTYAZISI ÜRETME" başlıklı bu tezin kendi çalışmam olduğunu, sunduğum tüm sonuç, doküman, bilgi ve belgeleri bizzat ve bu tez çalışması kapsamında elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara atıf yaptığımı ve bunları kaynaklar listesinde usulüne uygun olarak verdiğimi, tez çalışması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını, bu tezin herhangi bir bölümünü bu üniversite veya diğer bir üniversitede başka bir tez çalışması içinde sunmadığımı, bu tezin planlanmasından yazımına kadar bütün safhalarda bilimsel etik kurallarına uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul edeceğimi beyan ederim.

../../20..

İmzası

İlayda YILDIZ

ÖZET

DERİN ÖĞRENME TEKNİKLERİ KULLANILARAK KÜÇÜK VERİ SETLERİ İÇİN GÖRÜNTÜ ALTYAZISI ÜRETME

YILDIZ, İlayda

Doktora Tezi, Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Prof. Dr. Aybars UĞUR

Şubat 2025, 85 sayfa

Görüntüler için altyazı üretimi, temel olarak insan davranışıyla paralellik gösteren bir sistemin, derin öğrenme modelleri ve metotlarıyla sanal bir ortamda gerçekleştirilmesidir. Bu sistemlerde, girdiyi işlemeyi sağlayan bir kodlayıcı ve çıktıyı ifade etmeyi sağlayan bir kod çözücü mekanizma bulunmaktadır.

Yapay zekanın gelişimi ve derin öğrenme alanında yapılan çalışmaların önem kazanmasıyla uygun veri setlerine duyulan ihtiyaç artmaktadır. Bu tezde, derin öğrenme alanında küçük veri setleriyle çalışma problemi de ele alınarak az veriye sahip olunan durumlarda derin öğrenme algoritmaları kullanılarak geliştirilen modellerin başarısının hangi yollarla artırılabileceği üzerine çalışılmıştır.

Çalışmada, kodlayıcı olarak, derin öğrenme kullanılarak VGG16, ResNet, EfficientNetB2 gibi CNN modelleri ve VisionTransformer model; kod çözücü olarak ise LSTM, RNN, GPT, dikkat (attention) ve dönüştürücü (transformer) modeller gibi görüntülerden altyazı üretmek için değişik mimariler oluşturularak gerçekleştirimleri yapılmıştır. Önerilen farklı model kombinasyonları üzerinde deneysel çalışmalar yapılmış ve başarım değerleri elde edilerek performansları karşılaştırılmıştır. Görüntüler için altyazı üretimi kapsamında üretilen modellerle birlikte hem görüntü hem de metin veri tipleri için ayrı ayrı ve birlikte veri artırımı uygulanarak etkilerinin gözlemlenmesi tezin katkısını oluşturmaktadır.

Anahtar sözcükler: Görüntü altyazılama, küçük veri setleri, derin öğrenme, kodlayıcı-kodçözücü sistemler, veri artırımı, CNN, LSTM, RNN, dikkat modeli, dönüştürücü modeller, GPT.

ABSTRACT

IMAGE CAPTION GENERATION FOR SMALL DATASETS USING DEEP LEARNING TECHNIQUES

YILDIZ, İlayda

Ph.D. in Department of Computer Engineering

Supervisor: Prof. Dr. Aybars UĞUR

February 2025, 85 pages

Caption generation for images is basically the realisation of a system that parallels human behaviour in a virtual environment with deep learning models and methods. In these systems, there is an encoder mechanism to process the input and a decoder mechanism to express the output.

With the development of artificial intelligence and the increasing importance of studies in the field of deep learning, the need for appropriate data sets is increasing. In this thesis, the problem of working with small datasets in the field of deep learning is also addressed and the ways in which the success of the models developed using deep learning algorithms can be increased in situations with small datasets are used.

In this study, various architectures were implemented to generate captions from images using deep learning-based CNN models, including VGG16, ResNet, and EfficientNetB2, as well as the Vision Transformer model as the encoder. For the decoder, models such as LSTM, RNN, GPT, attention mechanisms, and transformers were utilized. Experimental studies were conducted on different combinations of these models, and their performance was evaluated through the use of success metrics. This thesis contributes by examining the effects of data augmentation, which was applied both separately and jointly to images and text, in conjunction with the proposed models for image captioning.

Keywords: Image caption generation, small dataset, deep learning, encoder-decoder systems, data augmentation, CNN, LSTM, RNN, attention model, transformer model, GPT.

ÖNSÖZ

Tıptan astronomiye, her türlü bilimsel çalışmada kullanılan görüntüler ve bu görüntülerin analiz edilmesi, anlamlı bilgilerin çıkarılması yapay zeka alanında büyük önem arzetmektedir. Günümüz sistemlerinde başarısı kanıtlanmış derin öğrenmenin çok miktarda veri ile çalıştığı ancak her alanda büyük boyutlarda veriye sahip olunamadığı bilinmektedir. Bu kapsamda bu çalışmayla görüntüler için altyazı üretimi alanında küçük veri setleri kullanılarak en hızlı ve en düşük maliyetli şekilde en efektif sonuçları üretebilmek adına çalışmalar gerçekleştirilmiştir.

İZMİR
12/02/2025

İlayda Yıldız

İÇİNDEKİLER

Sayfa

İÇİNDEKİLER (devam)

Sayfa

İÇİNDEKİLER (devam)

Sayfa

ŞEKİLLER DİZİNİ

Şekil **Sayfa**

ŞEKİLLER DİZİNİ (devam)

Şekil **Sayfa**

TABLOLAR DİZİNİ

Çizelge **Sayfa**

SİMGELER VE KISALTMALAR DİZİNİ

Simgeler	Açıklama
CNN	Evrişimli Sinir Ağları - Convolutional Neural Network
$ETrAtt$	EfficientNetB2-Attention-Transformer Model
GAN	Çekişmeli Üretici Ağ - Generative Adversarial Network
GPT	Üretken Ön Eğitimli Dönüştürücü
$ImgAug$	Görüntü Veri Artırımı
$ImgTxtAug$	Görüntü ve Metin Veri Artırımı
$LSTM$	Uzun ve Kısa Süreli Bellek - Long Short-Term Memory
$ResRNNAttGlo$	ResNet-Attention-RNN-GloVe Model
RNN	Tekrarlı Sinir Ağı - Recurrent Neural Network
$TxtAug$	Metin Veri Artırımı
ViT	VisionTransformer
$ViTGPT2Seq$	VisionTransformer-GPT2 Model

1 GİRİŞ

İnsan gözleriyle gördüğü bir olayı, hafızasından yardım alarak ve beyniyle işleyerek algılar, anlamlı bir yorumda bulunur. İnsanın doğasında bulunan bu temel davranış, günlük yaşamının her saniyesinde ihtiyaç duyduğu bir eylemdir. Bulunduğumuz çağın, en önemli çalışma alanı ise insan gücüne duyulan ihtiyacı azaltarak sistemlerin otomatizasyonunu sağlamaktır. Böylece hem insan hayatını kolaylaştırmak hem de daha güvenilir sistemlerin üretilmesi hedeflenmektedir. Görüntüler için altyazı üretimi, insan davranışına parallelik gösteren yapısıyla, organik olmayan sistemler üzerine uyarlanarak tam olarak bu ihtiyacı gidermektedir.

İfadeden de anlaşılacağı üzere görüntüler için altyazı üretimi konusunda hem görüntüler hem de metinsel verilerle çalışılmaktadır. Dolayısıyla yapay zeka alanında önem kazanmış bilgisayarlı görü ve doğal dil işleme alanlarını birleştirerek içermektedir. Bilgisayarlı görü, elektronik cihazlar yardımıyla elde edilen görüntülerin, bilgisayar sistemlerince anlaşılmasıdır. Doğal dil işlemenin odağı ise metinsel ifadelerdir. Metinlerin işlenip, bu metinlerden duygu ya da anlam çıkarımını sağlamaktadır.

Bu tür sistemleri, derin öğrenme ile ifade edecek olduğumuzda karşımıza kodlayıcı-kodçözücü sistemler çıkmaktadır. Bilgisayarlı görü kapsamında, kodlayıcı sistemler geliştirilip görüntülerin özniteliklerinin çıkarılması ve bilgisayar sisteminde ifade edilmesi sağlanır. Kodlayıcılar, bir girdinin bilgisayar sistemi üzerinde temsil edilmesini gerçekleştirir. (Doaa et al., 2023) Görüntülerin kodlanmasında kullanılan en popüler yöntemlerden biri CNN mimarisidir.

Doğal dil işlemeyi temel alan kod çözücü sistemler ise elde edilen bu bilginin anlaşılıp yorumlanmasını sağlayarak bir çıktı üretir. Kod çözücü sistemlerin mimarileri, önceden eğitilmiş yapıları kullanabilen kodlayıcı sistemlerden farklılık göstermektedir. Mimari içerisinde öz dikkat (self attention), çapraz dikkat (cross attention) ya da ileri beslemeli ağlar (feed forward network)

içerebilmektedir. (Shuming et al., 2021) Böylelikle üretilecek çıktının, girdideki hangi özelliklere odaklanılarak üretileceğine karar verilmektedir.

Tez çalışması kapsamında, görüntüler için altyazı üretimini sağlayan ResNet-Attention-RNN, ResNet-Attention-LSTM, ResNet-LSTM, VGG16-LSTM, EfficientNetB2-Attention- Transformer ve ViT-GPT2 yöntemleri ile farklı kodlayıcı - kod çözücü modeller gerçekleştirilmiştir. Böylece, görüntüden öznitelik çıkaran modeller ile özniteliklerden altyazı üreten modellerin kombinasyonları oluşturularak altyazı üretme başarımları elde edilerek sunulmuştur. Kodlayıcı sistemler kapsamında bu çalışmada, derin öğrenmede aktarmalı öğrenme (transfer learning) yöntemi kullanan VGG16, ResNet, EfficientNetB2 CNN modelleri ve görü dönüştürücü (VisionTransformer) model kullanılmıştır. Uzun-Kısa Süreli Bellek (Long-Short Term Memory), Tekrarlı Sinir Ağları (Recurrent Neural Network), Üretken Önceden Eğitilmiş Dönüştürücü Model (Generative Pre-trained Transformer) ve dönüştürücü (Vision Transformer) modeller, kod çözücü mimariler olarak kullanılmıştır.

Geleneksel makine öğrenmesi tekniklerinin ve derin öğrenme sistemlerinin, sergiledikleri başarı, eğitildikleri veri setinin içerdiği çeşitlilik ya da veri setinin boyutu gibi özelliklere bağlıdır. Özellikler derin öğrenme algoritmalarının çok miktarda veriyle çalıştığı bilinmektedir. Verinin az olduğu durumlarda algoritma yeterince eğitilememekte ve öğrenme işlemi başarılı bir şekilde gerçekleştirilememektedir. Bazı çalışma alanlarında veri setinin yeterli olmaması söz konusu olabilir. Böyle durumlarda veri artırımına ihtiyaç duyulmaktadır. Bu çalışmada odaklanılan bir diğer konu da bu gibi küçük veri setlerine sahip olunan durumlarda uygulanabilecek efektif yöntemlerin bulunmasıdır. Görüntüler için altyazı üretiminde hem görüntü hem de metinsel verilerle çalışıldığı bilinmektedir. Dolayısıyla veri artırımı bu alanda uygulandığında her iki veri tipi için de artırma işleminin yapılması gerekmektedir. Tahmin edilebileceği üzere görüntü ve metinsel veriler yapısal olarak farklılıklar gösterdiklerinden uygulanan artırma teknikleri de farklı olmaktadır. Çalışmada ayrıca, görüntüler için altyazı üreten bir modelin; sadece görüntülerin, sadece

metinlerin ya da her iki veri tipinin de artırıldığı durumlarda gösterdiği performanslar elde edilerek incelenmiştir.

Herkes tarafından kolaylıkla elde edilebilen veri tipi olan görüntülerle çalışılarak bu görüntülerden anlamlı bilgi çıkarımını sağlayan altyazı üretimi alanında çalışmalar gerçekleştirilmiştir. Bu tezde yer alan diğer bölümler şu şekilde özetlenebilmektedir: ikinci bölümde, görüntüler için altyazı üretiminin nasıl çalıştığı, neleri barındırdığı ile ilgili detaylı bilgiler yer almaktadır. Üçüncü bölümde görüntüler için altyazı üretiminde kullanılan ve bu çalışma kapsamında gerçekleştirilmiş derin öğrenme mimarileri ilgili teknik detaylar ve sisteme uyarlanmalarını içeren bilgiler bulunmaktadır. Dördüncü bölümde, tezin odağındaki, literatürde bilinen küçük veri setleriyle çalışma problemi, detayları ve getirilebilecek çözüm yöntemlerinden bahsedilmiştir. Beşinci bölümde literatürde yer alan önceki çalışmalara değinilmiştir. Altıncı bölüm, tez kapsamında gerçekleştirilen tüm model kombinasyonları ve elde edilen sonuçları detaylarıyla içermektedir. Yedinci bölümde deneysel çalışmalar, sekizinci ve son bölümde ise tez süresince elde edilen genel sonuçlar ortaya konulmakta ve tartışılmaktadır.

2 GÖRÜNTÜ ALTYAZI ÜRETİMİ

Görüntü altyazı üretimi, insan algısı doğrultusunda görsel zekanın daha fazla araştırılması için çok önemli bir görevdir. Basit bir ifadeyle, makinelerin, görüntülerin öznitelikleri üzerine öğrenip sonrasında o görüntüler için metinsel olarak tanımlayıcı ifade verme yeteneğidir. Görüntülerden öznitelik çıkarılmasında bilgisayarlı görü ve altyazı üretiminde ise doğal dil işleme kullanılmaktadır. (Veena et al., 2022) Bu iki alanı barındıran geniş kapsamlı bir araştırma konusudur.

Görüntü altyazı sistemlerinin çok çeşitli uygulama alanları mevcuttur. Örneğin, görsel yorumlamada zorluk yaşayan farklı engelli kişilere yardımcı olmaktadır. Bu sistemler tarafından oluşturulan altyazılar, metinden konuşmaya dönüştürme sistemlerine beslenebilmektedir. Yayıncılıkta, gönderilen çok büyük haber bültenleri ve görsel resimler bulunmaktadır ve görüntü altyazılama yardımı ile en anlamlı yorumları üretmekte ve en çok aranan bilgilere odaklanmayı sağlamaktadır. Ayrıca tıbbi teşhis araçları, hastadaki herhangi bir anormalliğe ilişkin fizyolojik parametreleri ve tıbbi görüntü yorumunu bir araya getirerek metin resim yazısı ekleme aracı olarak sağlık alanında kullanılabilmektedir. (Biradar et al., 2023) Görüntüler için altyazı üretimi yaşamımızın birçok alanına uyarlanabilmekte ve karşılaşabileceğimiz problemlere çözüm sunmaktadır.

Derin öğrenmeyi kullanan görüntü altyazısı üretimi, görme engelli bireylere yardımcı olmak, içerik indeksleme ve erişimini geliştirmek ve e-ticaret ve eğlence gibi alanlarda kullanıcı deneyimlerini geliştirmek de dahil olmak üzere çeşitli uygulamalarda fayda sağlamaktadır. (Chitteti & Madhavi, 2024) İşlem görüntüyü analiz etmekle başlamaktadır. Görüntü üzerindeki nesnelerin ve bunlar arasındaki ilişkilerin tespit edilmesi gerekmektedir. Bu işlemin otomatikleştirilmesi, insan gözlemcilerin gözden kaçırdığı nesnelere için de metinsel ifadeler eklemek için kullanışlı olmaktadır. Bu detayları yakalayabilmek için derin öğrenmeye başvurulmaktadır. Derin öğrenmeden önce, özniteliklerin

çıkarılması ve nesnelerin sınıflandırılması adına sınıflandırıcılar kullanılmıştır. Ancak, büyük veri kümelerinden özellik çıkarmanın karmaşıklığı nedeniyle, geleneksel yaklaşımlar, özellikleri otomatik olarak öğrenen derin öğrenme kapsamındaki yaklaşımlarla karşılaştırıldığında iyi bir seçim olmadıkları görülmektedir. (Naga et al., 2023)

Yapay zekanın zorlu ve anlamlı bir alanı olarak görüntü altyazılarının otomatik olarak oluşturulması giderek daha fazla ilgi görmektedir. Görüntü altyazı üretiminin amacı, görüntünün içeriğine anlamsal olarak uygun olan dilsel açıdan makul cümleler oluşturmaktır. Bu nedenle görüntü tanımlamanın iki temel yönü vardır: bilgisayarlı görü ve dil işleme. Bu teknikler, oluşturulan cümlelerin dilbilgisi ve anlamsal olarak doğru olmasını sağlamak için, ilgili yöntemlerin yarattığı sorunların üstesinden gelmek ve bunları uygun şekilde bütünleştirmek için kullanılmalıdır. (He & Lu, 2019)

2.1 Bilgisayarlı Görü

Bilgisayarla görme teknolojisi, görüntü işleme, otomasyon, bilgisayar bilimi ve optik dahil olmak üzere çok çeşitli araştırma alanlarını kapsayan çok disiplinli bir alandır. Kamu güvenliği, güvenlik, bilimsel araştırma, ulaşım, askeri, havacılık, meteoroloji, sanayi ve benzeri birçok alanda yaygın olarak kullanılmaktadır. Bilgisayarla görme teknolojisi olarak adlandırılan teknoloji, temel olarak insanın görsel işlevlerinin gerçekleştirilmesini kolaylaştırmak için bilgisayar teknolojisine dayalı yargılama ve ölçüm anlamına gelmektedir. Genel olarak bilgisayarlı görme teknolojisi, lens kontrol ekipmanı, edinim ekipmanı ve ilgili algoritmalar vb. dahil olmak üzere çeşitli öğelerle entegre edilmiş donanım ve yazılım olmak üzere iki parçadan oluşmaktadır. (Qu, 2023) Bu teknolojinin gerçek kişilerin işlem yapamadığı bazı işletim ortamlarında uygulanması umut vericidir.

Yapay zekanın önemli bir kolu olan bilgisayarlı görünün amacı görüntüleri anlamak, yorumlamak ve anlamlı bir bilgi çıkarmaktır. Derin öğrenme yöntem-

leriyle bu alanda iyi bir ilerleme kaydedilmiştir. Bilgisayarlı görü de kendi içinde farklı teknikler barındırmaktadır. Örneğin, görüntü üzerindeki nesneleri yerelleştirme ve tanıma görevlerini ele alan nesne tespiti işlemi en önemli uygulamalarından bir tanesidir. (Zu et al., 2024) Bu işlem için yaygın olarak derin öğrenmedeki CNN mimarisi kullanılmaktadır. Aktarmalı öğrenmeyi barındıran CNN modellerinin kullanılması, görüntüye ait bilgilerin çıkarımını kolaylaştırmakta ve doğru bilgilere odaklanabilmeyi, dolayısıyla sağlayarak gereksiz ayrıntılardan kaçınmayı sağlamaktadır.

Uygulanan bilgisayarlı görme algoritmaları iyi bir verimliliğe sahip olmasına rağmen, görüntü işleme konusunda hala sınırlamaları ve zorlukları vardır. Örneğin bilgisayarlı görü, özellikle de derin öğrenme, yüksek doğruluğa sahip olabilmeleri için milyonlarca verinin eğitilmesine ihtiyaç duymaktadır. (Charleen et al., 2021) Görüntü miktarının artması, çeşitliliğin artmasını sağlamakta ve öğrenmenin kalitesini arttırmaktadır. Böylelikle bilgisayarlı görü algoritmalarının performansı iyileştirilmektedir.

2.2 Doğal Dil İşleme

Doğal dil işleme (DDL), bilgisayarların insan dilini değerli bir şekilde anlamasını, yorumlamasını ve yanıt vermesini sağlayan bir yapay zeka dalıdır. (Lopez & Kalita, 2017) Bu konu, metinlerin ne ifade ettiğini anlayabilmeyi sağlayan duygu analizi, cümle benzerliği keşfi, metinden metine dönüşüm sağlayan makine tercümesi, bir ifadeden bir metin üretebilmeyi ya da bütün metni bir cümle ile ifade edebilmeyi sağlayan metin özetleme gibi birçok alanda kullanılmaktadır.

Doğal Dil İşleme (NLP) alanındaki önemli görevlerden biri, metin cümleleri arasındaki anlamsal benzerliğin belirlenmesidir. Geleneksel makine öğrenimi algoritmaları çok büyük miktarda eğitim verisi gerektirir, ancak bu zaman alıcı bir süreçtir. Önceden eğitilmiş modeller, sinir ağı topolojilerinin ve dil temsillerinin özelliklerini genel olarak öğrenmeye yönelik yöntemler kullandıklarından,

çeşitli aşağı yönlü uygulamalar için değiştirilebilir. Dönüştürücülerden Çift Yönlü Kodlayıcı Gösterimleri - BERT ve GPT, etkili sonuçlar üretmek için minimum ince ayar çabası kullanılmasını sağlayan Doğal Dil İşleme'deki popüler mimarilerdir. (Mayil & Jeyalakshmi, 2023) Bunların dışında yapay sinir ağları da ürettikleri yüksek başarıyla doğal dil işlemede önemli rol oynamaktadır.

Yalnızca metin analizi dikkate alınırsa, dilbilimciler tarafından kullanılan metodolojinin yüksek doğrulukta bir sonuç sağlaması zordur, çünkü dil belirsizdir ve daha fazla veri doğruluğu artırmaya yardımcı olsa da bilgisayarlar dili insanlar gibi anlayamazlar. Bu durumda, birçok sorun bir sinir ağı devreye girene kadar çözülmeyecektir. (Wang, 2024) Örneğin, tekrarlayan sinir ağları (RNN), zaman serisi verilerini işleyebilen bir tür sinir ağıdır. (Topbaş et al., 2021) Dolayısıyla cümleler ve metinler gibi sıralı verilerle çalışmada kullanılabilecek en güzel yöntemlerden bir tanesidir. Merkezi kelimenin hem sağındaki hem de solundaki kelimelerden herhangi bir zamanda yararlanmak, cümledeki yalnızca önceki kelimelere bakmaktan ziyade kelimelerin bağlamı hakkında daha yararlı bilgiler sağlayabilir. (Topbaş et al., 2021) Ancak RNN'lerin temelini oluşturan önceki ve sonraki kelimelere duyarlılık sağlamak bir süre sonra performanslarını azaltmaktadır. Bu sorun, uzun ve kısa süreli bellek kullanımını etkinleştiren LSTM'lerin ortaya atılmasını sağlamıştır. LSTM ile sağlanan yaklaşım, geleneksel unutma kapısını optimize ederek hesaplama masraflarını azaltmayı sağlamaktır. Başka bir yaklaşım ise standart prosedürün önceki adımından elde edilen çıktıya güvenmek yerine, göreve özgü bilgileri çıkarmak için LSTM'nin nihai çıktısı içindeki hem zamansal hem de özellik boyutlarında bir dikkat modeli sağlamaktır. (Upadhyay et al., 2024) Böylelikle sıralı verilerin bütününe değil parçalarına odaklanılarak anlam çıkarma işleminin başarısı artırılmaktadır.

3 DERİN ÖĞRENME

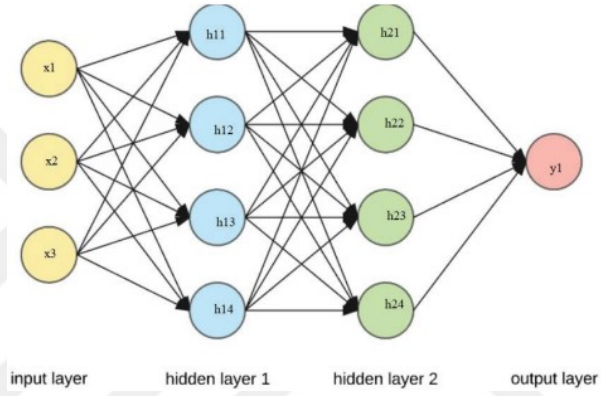
Görüntü altyazısı üretme, makine öğrenmesi tabanlı modeller kullanılarak resimde mevcut bilgilerin çıkarılması ve görsel görüntü özelliklerinin anlamlı hale dönüştürülmesidir. (Biradar et al., 2023) Özel matematiksel yapılar olan derin öğrenme modelleri, belirli veri kümeleri kullanılarak eğitime tabi tutulur ve tahminler oluşturmak için kullanılır. Çeşitli alanlarda, derin öğrenme teknikleri geçmişte cesaret verici sonuçlar ortaya koymuştur. (Niteesh & Pooja, 2024) Derin öğrenmenin her alana uyarlanabilirliği ve ürettiği yüksek başarılarla güvenilirliği göz önünde bulundurularak popülerliği artmıştır. Bilgisayarla görme, ses tanıma, makine çevirisi, doğal dil işleme gibi çok çeşitli endüstrilerde derin öğrenme uygulamalarıyla karşılaşmak mümkündür.

Birçok makine öğrenmesi tekniğinde olduğu gibi derin öğrenmenin temelinde de insan davranışı yer almaktadır. İnsan sinir hücresinden esinlenilerek geliştirilen algılayıcılar (perceptron) yapay sinir ağlarının temelini oluştururken yapay sinir ağları da derin öğrenmenin temelini oluşturmaktadır. Derin öğrenmedeki "derin" terimi, ağın çeşitli gizli katmanlardan yararlanmasını ifade etmektedir. Sonsuz sayıdaki boyut-sınırlı gizli katman mimarisini kullanan bu tasarım, daha yüksek düzeyde bilgi çıkarmak için ham girdi verilerinden yararlanmaktadır. Eğitim verilerinin karmaşıklığı, gizli katmanların sayısının belirlenmesi için kullanılmaktadır. Etkili bir şekilde iyi sonuçlar elde etmek için, daha karmaşık verilerin daha fazla gizli katmana ihtiyacı bulunmaktadır. (Patel et al., 2023) Bu katmanlar, öğrenmede gerekli bilgilerin çıkarılması hususunda önem arz etmektedir.

3.1 Yapay Sinir Ağları

Bilgisayar bilimi ile doğrusal olmayan dinamikler ve kaos teorisi gibi teorik temellerin birleştirilmesi, kendilerini yüksek karmaşıklıktaki sorunlara dinamik olarak adapte edebilen yapay sinir ağları (YSA) gibi zeki etmenlerin oluşturulmasına olanak tanımıştır. YSA'ları, birden fazla faktörün dinamik

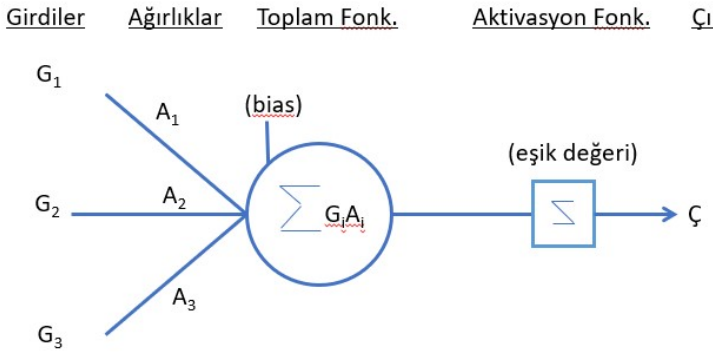
etkileşimini aynı anda yeniden üretebilmektedir, bu da karmaşıklığın incelenmesine olanak tanımaktadır. Ortalama eğilimler olarak değil, bireysel bazda da sonuçlar çıkarabilmektedirler. Bu araçlar klasik istatistiksel tekniklere göre gözle görülür avantajlar sunmaktadır. Yapay sinir ağı, insan sinir merkezinin dışarıdan gelen bilgileri kabul ederek iç yapısını değiştirebilen bir simülasyon yapısıdır. Daha spesifik olarak, bu, kademeli bir kişisel adaptasyon sürecidir. Örneğin bir bilgisayar, kediyi diğer hayvanlardan ziyade bir kedi görüntüsünden tanıyabilir. (Grossi & Buscema, 2007) Bu, aynı insanların öğrenme sürecinde olduğu gibi öncelikle benzer görüntülerle eğitilme ve sonrasında karşılaşılan yeni bir durumda anlamlı bir çıkarım yapabilmektir.



Şekil 3.1: Yapay Sinir Ağı (Patel et al., 2023)

YSA'nın temel elemanları, işleme elemanları olarak da adlandırılan düğümler ve bağlantılardır (Şekil 3.1). Her düğümün, diğer düğümlerden ve/veya ortamdan iletişim aldığı kendi girişi ve diğer düğümlerle veya çevreyle iletişim kurduğu kendi çıkışı vardır. (Grossi & Buscema, 2007) İnsan beynindeki sinir hücresinin bir taklidi olan bu yapı, gerçek bir sinir hücresinde yer alan akson ve dendridlerin diğer sinir hücreleriyle iletişim kurdukları gibi düğümlerde de birbirleriyle iletişim kurmaktadır. Şekil 3.2'de örnek bir yapay sinir hücresi temsil edilmiştir. Sinir hücresinin her bir girdisi farklı ağırlık değerleriyle işleme tabi tutulabileceği gibi eşit ağırlık değerlerine sahip olup, sisteme eşit oranda da etki edebilirler. Bir nörona giren her bir girdi değeri bir ağırlık değeri ile çarpılarak uyarlanır. (Adalıer & Uğur, 2008). Girdi ve ağırlıkların

işleme alınması, tanımlanan ve Formül 3.1'de ifade edilen toplam fonksiyonuyla sağlanmaktadır. Hesaplanan bu değerin yapay ağda dikkate alınıp alınmayacağı, sistemde tanımlanan bir eşik değeriyle kontrol edilmektedir. Eşik değeri, belirli bir değerin belirlenip alt ve üst değerler olarak anlaşılmasını sağlayan ikili basamak (binary step); sigmoid, tanh gibi sürekli olmayan (non-linear) ya da sürekli (linear) fonksiyonlar, sadece pozitif değerleri dikkate alan ReLU gibi fonksiyonlarla bir aktivasyon fonksiyonu kullanılabilmektedir. Belirlenen eşik değerine göre kullanılabilir bulunan değerler çıktı olarak üretilmektedir. Sistemden istenen çıktının elde edilememesi durumunda, çıktıya göre ağırlıkların güncellenmesi geri yayılım yöntemiyle söz konusu olabilmektedir.



Şekil 3.2: Yapay Sinir Hücresi

$$T = G_1A_1 + G_2A_2 + G_3A_3 + bias \quad (3.1)$$

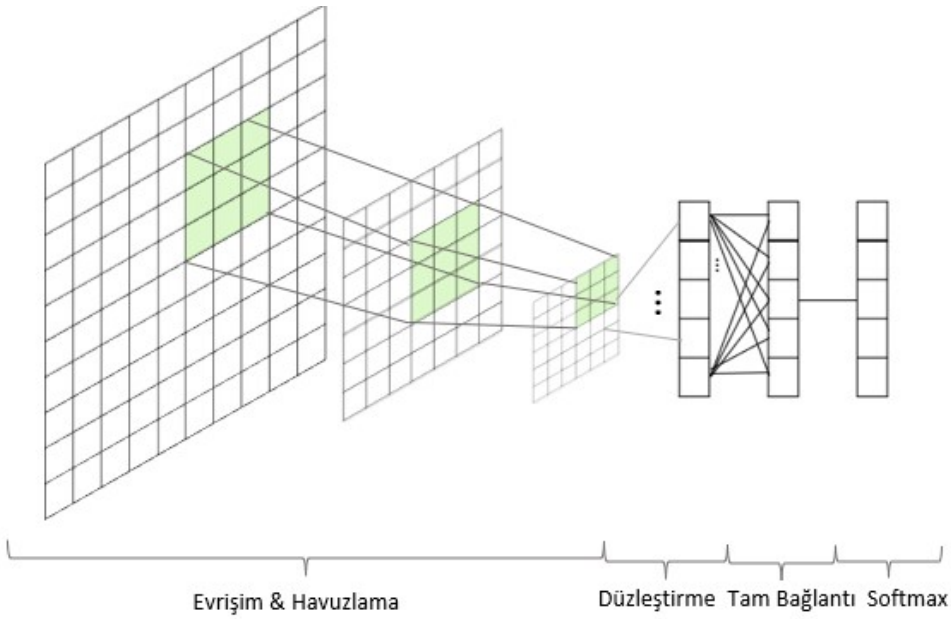
Yapay sinir ağları ile gerçekleştirilen karmaşık modellerden, evrişimli sinir ağı (CNN) ve tekrarlayan sinir ağı (RNN), doğal dil işleme ve görüntü altyazı üretimi alanında yaygın olarak kullanılmaktadır. Evrişimli sinir ağları, sınıflandırma görevleri için kullanılırken, tekrarlayan sinir ağları, doğal dil işlemede çeviri alanında daha çok kullanılmaktadır. (Wang, 2024) Tekrarlayan sinir ağları, yapıları itibariyle sıralı veriyle çalışmak için geliştirilmiş modellerdir. Bu modelde, diziden diziye çeviri problemini çözmek için yeni bir yapının geliştirilmesine ihtiyaç duyulmuştur. Böylelikle kodlayıcı ve kod çözücü sistemler tasarlanmıştır ve en son gelişmeler bu yapıya dayanmaktadır.

(Wang, 2024) Bir başka yapay sinir ağı olan LSTM sinir ağları, tekrarlayan sinir ağlarındaki sorunların üstesinden gelebilmek ve zaman serisi modellemede kullanılabilmek için en uygun seçimlerden birisidir. LSTM model mimarisi, yinelenen sinir ağlarında kaybolan değişimlerin üstesinden gelinmesi ve zaman serisi verilerindeki uzun vadeli bağımlılıkları öğrenmek için depolama birimleri ve kapıların kullanılması tahminleme başarısını artırmaktadır. (Lopez et al., 2023)

3.2 CNN

Evrişimli sinir ağları, görüntü verileriyle çalışmada klasikleşmiş bir derin öğrenme mimarisidir. Genellikle görüntü sınıflandırma amacıyla kullanılan bu mimari otonom sürüş, görüntü altyazı üretimi, hastalık teşhisi gibi farklı alanlarda da popüler olarak başvurulan bir yöntemdir.

Temelinde derin öğrenme bulunduğundan, CNN'ler de katmanlı bir yapıya sahiptir. CNN modeli, girdi verilerini alan ve çıktıyı üretmek için modelin hesaplama adımını takip eden, ardından modelin parametrelerini ayarlayarak çıktı hatasını ve gerçek değeri en aza indiren temel sinir ağı modeliyle aynı mantığı izlemektedir. Bunlar aynı zamanda ileri yayılım ve geriye yayılım olarak da bilinmektedir. (Wang, 2024) Şekil 3.3'te de gösterilen CNN mimarisinde, görüntü ilk olarak bilgisayar dilinde bir matrisle ifade edilmektedir. Bu matrise, çekirdek filtreler (kernel function) uygulanarak ham ve değerli bilginin çıkarılması işlemi CNN mimarisinin başlangıcında bulunan evrişim ve havuzlama katmanı ile sağlanmaktadır. Elde edilen bu yeni özet matris, mimaride takip eden düzleştirme (flatten) katmanı ile tek boyutlu bir vektörle ifade edilmektedir. Bu katmanı yapay sinir ağının temelini oluşturan tam bağlı (fully connected/dense) katman izler ve girdiler üzerinden çıktıların hesaplanması gerçekleştirilir. Son olarak softmax katmanıyla üretilen çıktılar gösterilmektedir.



Şekil 3.3: CNN Katmanları (Tabian et al., 2019)

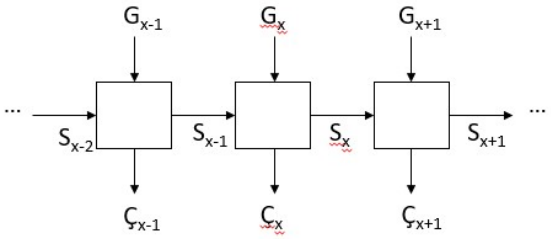
Son yıllarda derin öğrenmeye verilen önemin artması beraberinde bu alanda kullanılan mimarilerin de önemini artırmıştır. Görüntü kodlamada akla gelen ilk yöntemlerden biri olan CNN'e duyulan, gereklilik farklı alanlar için özelleşmiş, farklı CNN modellerinin geliştirilmesine neden olmuştur. VGG16, implentasyonunun/gerçekleştiriminin kolaylığı nedeniyle sınıflandırma çalışmalarında en yaygın kullanılan CNN modellerinden bir tanesidir. Karen Simonyan ve Andrew Zisserman tarafından geliştirilen bu yapı, bilinen yaygın görüntü veri setlerinden, içerisinde 14 milyon görüntü ve 1000 kategori bulunan, ImageNet veri setiyle eğitilmiştir. İsminde yer alan "16", barındırdığı katman sayısını ifade etmektedir. Katmanlı mimarisinde; evrişim ve ReLU, maks havuzlama, tam bağlı katman ve ReLU ve son olarak softmax katmanları yer almaktadır. Bir diğer popüler CNN modeli ResNet'tir. ResNet50, ResNet101, ResNet152 olarak adlandırılan ResNet modelleri VGG'de olduğu gibi adlarını içerdikleri katman sayısından almaktadır. Yapısında kullanılan artık (residual) katman bazı durumlarda ağırlıklı katmanların göz ardı edilmesini sağlamaktadır. Böylelikle veri setine aşırı uyum engellenmekte ve eğitim işleminin hızı artmaktadır. ImageNet veri setiyle eğitilmiştir ve kaybolan gradyan (vanishing gradient) problemini çözmek için geliştirilmiştir. Bir başka CNN modeli olan

InceptionV3'ün geliştirilmesinin altında yatan sebep, işlem gücüne duyulan ihtiyacın azaltılmasıdır. Daha az parametreyle çalışılarak daha hızlı eğitim işlemi gerçekleştirilmektedir. (Veena et al., 2022) Bilinen ve yaygın olarak kullanılan bu üç modelin dışında literatürde özelleşmiş Xception, DenseNet, EfficientNet gibi daha birçok model bulunmaktadır.

3.3 RNN

Makine öğrenmesi algoritmalarında eğitim süreci ve bu süreçte sağlanan veri o modelin başarısını artıran en önemli faktörlerden biridir. Veri, model tarafından anlaşılıp işlenir ancak her veri aynı yapısal özelliğe sahip değildir. Örneğin, bilgisayarlı görü sisteminde kullanılacak veri seti, görüntülerden oluşurken doğal dil işleme alanındaki veriler cümleler ya da metinlerdir ve bu iki veri türü yapıları gereği birbirlerinden farklıdır. Bir görüntünün anlaşılıp bilginin çıkarılması o görüntüye bağlıdır. Ancak bir cümleden bilgi çıkarımı yapılmak istendiğinde, o cümledeki kelimelerin tek başlarına ve bir araya gelerek ne ifade ettiklerinin, hangi kelimenin anlam çıkarımında diğerinden daha değerli olduğunun, kelimelerin sıralamalarının da göz önünde bulundurulması gerekmektedir.

Farklı alanlar için özelleşmiş farklı modeller, o alana özgü veri setleriyle eğitilirler. Metinsel veriler ile çalışıldığında akla gelen ilk yöntemlerden bir tanesi de tekrarlayan sinir ağlarıdır (RNN). Bu ağlar, Şekil 3.4'te de gösterilen tekrarlı yapıları sayesinde sıralı veriyle çalışma konusu için özelleşmiştir. Sıralı verideki elemanlar teker teker işleme alınır ancak sonraki elemena geçildiğinde önceki elemanların ne ifade ettiği unutulmaz. Elemanlar arası ilişkiye bakılarak ihtiyaç duyulduğunda önceki elemanlara da başvurularak anlam çıkarma işlemi gerçekleştirilir. Özetle, tekrarlayan sinir ağlarının, CNN gibi ileri beslemeli ağlardan farkı, önceki katmanda elde ettiği çıktıyı saklaması ve bu çıktıyı sonraki katmana girdi olarak sağlamasıdır.



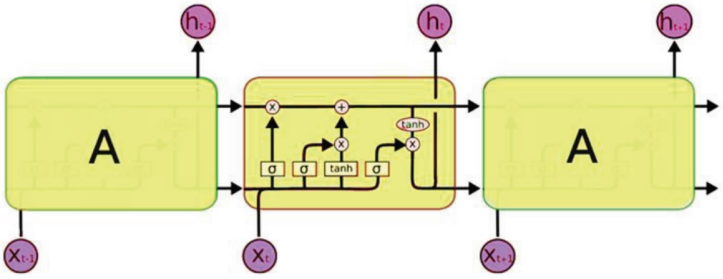
Şekil 3.4: RNN Mimarisi

Tekrarlayan sinir ağları, görüntü altyazı üretiminde popüler olarak kullanılan yapay sinir ağlarındandır. Farklı problemlere göre adapte edilen bu sistemler; görüntü işleme alanında kullanılan vektörden sıralı veriye, duygu analizinde kullanılan sıralı veriden vektöre, dil çevirisinde kullanılan sıralı veriden sıralı veriye dönüşüm sağlayan farklı yapılarla genel olarak üç bölümde ele alınabilmektedir. Görüntü altyazı üretiminde, tekrarlayan sinir ağının girdisi, eğitim setinde bulunan görüntüye ait kodlayıcıdan gelen öznitelikler olurken ve beklenen çıktı da kodlayıcıya verilen görüntünün tahmini altyazısıdır. Yani tekrarlayan sinir ağının hem girdisi bir vektör iken çıktısı ise metindir bir başka deyişle sıralı veridir.

3.4 LSTM

Yapay sinir ağları, genelleme ve karar verme yeteneğine sahip güçlü sınıflandırıcılardır. Güçlerini, Şekil 3.5'te de gösterilen gelişmiş katmanlı yapılarından almaktadırlar. Yapay sinir ağlarındaki katmanlı yapılar, değiştirilip özelleştirilerek farklılar gösterebilir. Uzun ve kısa süreli bellek (LSTM) yöntemi, tekrarlayan sinir ağları (RNN) gibi ardışık verilerle çalışma konusunda özelleşmiş bir yapıdır ve RNN'leri daha da iyileştirerek daha geniş aralıktaki verilere ulaşabilme yeteneğine sahiptir. RNN'lerdeki en önemli sorunlardan bir tanesi uzun vadeli bağımlılıkları dikkate alamamalarıdır. Bu problem RNN'lerin çıkarım yapma konusundaki başarısını olumsuz yönde etkileyen bir faktördür.

LSTM'ler mimari açıdan diğer sinir ağlarına göre biraz daha karmaşık bir yapıdadır. Hatırlama ve saklama prensibinden yola çıkılarak geliştirilmiştir.



Şekil 3.5: LSTM Mimarisi (Rohitharun et al., 2022)

Genel olarak bir hücre ve bu hücrenin kapıları olarak düşünüldüğünde Şekil 3.5'te olduğu gibi üç bölümde ele alınabilir. İlk olarak, hücrenin girdisini x(i) ve önceki hücreden aktarılan değeri h(t-1) dikkate alan, sigmoid fonksiyonunu kullanan (Formül 3.2) ve hangi bilginin silineceğine karar veren unutma kapısıdır.

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f \qquad (3.2)$$

Hücrenin sonraki aşamasında, hücrede hangi bilginin girdi olarak kullanılacağına karar veren girdi kapısı yer almaktadır. Bu kapıda, Formül 3.3'te de gösterildiği gibi sigmoid ve tanh fonksiyonları birlikte kullanılmaktadır. Hücrenin mevcut durumuna etki edecek olan girdi belirlenmektedir.

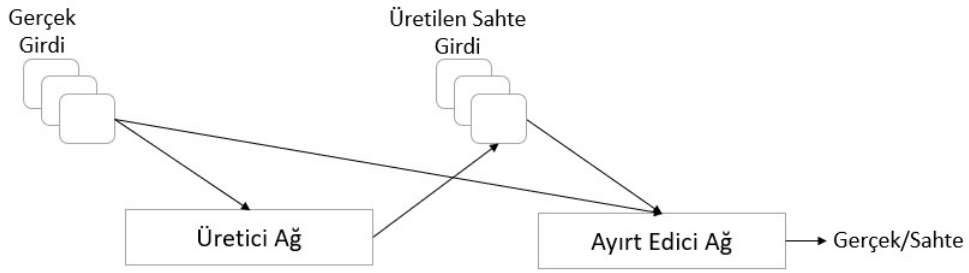
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t]) + b_i$$
$$\tilde{c}_t = tan\, h(W_c \cdot [h_{t-1}, x_t]) + b_c \qquad (3.3)$$

Son olarak, sonraki hücrede, bu hücredeki hangi bilginin kullanılması gerektiğine karar veren çıktı kapısı bulunmaktadır. Mevcut durumun hangi kısmının çıktı olduğuna karar verilmektedir. Formül 3.4 ile gösterilmiştir.

$$Ot = \sigma(W_0 \cdot [h_{t-1}, x_t] + b_0)$$
$$h_t = Ot \cdot tanh(Ct) \qquad (3.4)$$

3.5 GAN ve Difüzyon Modelleri

Çekişmeli üretici ağlar, derin öğrenme mimarisine sahip yapay sinir ağlarının, üretim ve ayırt etme üzerine yoğunlaşmış özel bir çeşididir. GAN, minmax oyununu oynayabilmek için geliştirilmiş ağlardır. Üretici ağ amaç fonksiyonunu (objective function) maksimize etmeye çalışırken, ayırt edici ağ bu değeri minimize etmeye çalışmaktadır. Bu ağlar yapısal olarak Şekil 3.6 ile gösterildiği gibi iki bölümden oluşmaktadır. İlk olarak üreteken ağ (generative) bölümü, ağa girdi olarak verilen verilerle eğitilerek öğrenir ve benzeri çıktılar üretmektedir. Çekişmeli üretici ağın diğer kısmı ise ağa başlangıçta gerçek girdi olarak verilen verilerle, ağın kendisi tarafından üretilen verileri ayırt etmeyi sağlayan ayırt edici (discriminative) ağdır. Her iki ağda çalışması sırasında kendi içerisinde ağırlıklarını güncellemekte ve kendisini optimize etmektedir.



Şekil 3.6: Üretici Çekişmeli Ağ (GAN)

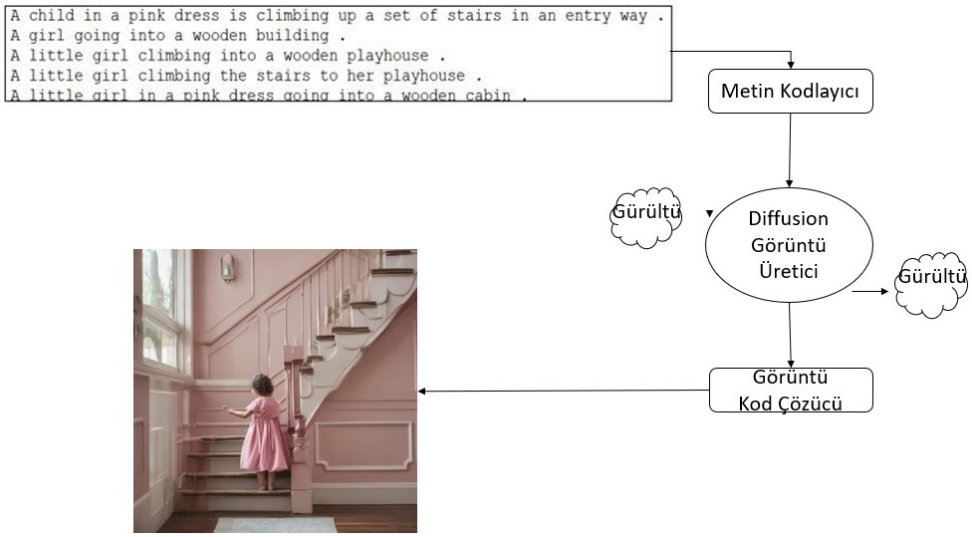
Güçlü donanım ve veri setine ihtiyaç duyan çekişmeli üretici ağlar, sahte veri üretimi sağlamaktadır. Sahte veri üretimi, sisteme verilen gerçek girdilere, rastgele gürültü ekleme işlemi ile sağlanmaktadır. GAN'ların performanslarını iyileştirmek amacıyla geliştirilen bazı modeller mevcuttur. Örneğin koşullu (conditional) GAN (C-GAN), üretici ve ayırt edici ağın sınıf etiketlerine bağlı olarak çalışmasıyla gerçekleştirilmiştir. Bu ek bilgi, eklenecek olan gürültü ile birlikte ele alınır ve böylece üretici ağ ekleyeceği gürültüyü buna göre ayarlamaktadır. Geliştirilmiş bir diğer model, görüntü verileriyle çalışmak

için özelleşmiş, üretici ve ayırt edici ağı evrişimli ağdan oluşturulmuş DC-GAN'dır. Adımlı (strided) ve kesirli adımlı (fractionally strided) evrişimli ağ, uzaysal yukarı örnekleme (up sampling) ve aşağı örnekleme (down sampling) operatörlerini öğrenmek için kullanılır. (Shahriar, 2022) Bu da görüntüyü, olduğundan daha düşük boyutlu bir uzayda temsil edebilmeyi sağlamaktadır. Ayrıca sıralı veriyle çalışmayı kolaylaştıran tekrarlayıcı GAN'lar (RGAN) da mevcuttur. VanillaGAN (Zeng et al., 2019), AttnGAN (Zhang et al., 2021), StackGAN (Karpathy & Fei-Fei, 2019) literatürde metinden görüntü üretimi için geliştirilmiş diğer GAN çeşitleridir. Literatürdeki CycleGAN (Zu et al., 2017), StyleGAN (Karras et al., 2019), Pix2PixGAN (Isola et al., 2017) gibi mimariler, belirli bir koşula dayalı olarak çalışmakta ve üretilen görüntülerin istenilen özelliklere sahip olmasını sağlamaktadır. Genel olarak bu sistemler koşullu GAN (CGAN) mimarileri olarak adlandırılmaktadır.

Difüzyon modeller GAN'ların bir alternatifi olarak düşünülebilecek, metinlerden görüntü üretimini destekleyen benzer bir yöntemdir. Difüzyon modellerin bir türü olan kararlı difüzyon, geleneksel GAN modellerine kıyasla daha çeşitli ve daha yüksek kaliteli görüntüler üretebilmekte ve görüntü sentezine hakim olan GAN ailesi modellerini geride bırakabilmektedir (Yidan et al., 2014). Bu modeller gereksiz detaylardan uzaklaşıp, kayıp değerleri azaltmayı amaçlamaktadır. Difüzyon modelleri, verilere kademeli olarak gürültü ekleme ve ardından orijinal verileri yeniden oluşturmak için bu gürültüyü tersine çevirmeyi öğrenme süreciyle çalışmaktadır. (Şekil 3.7) Bu yaklaşım, difüzyon üretici modellerini özellikle karmaşık veri dağılımları için çok uygun hale getirmektedir. Böylelikle daha yüksek kalitede görüntüler elde edilmektedir.

3.6 Dikkat Modeli (Attention Model)

Kodlayıcı - kod çözücü mimariyle gerçekleştirilmiş görüntü altyazı üretimi modelinde, kod çözücü bölüm, altyazı üretme işlemini kelimeleri tek tek inceleyip baz alarak gerçekleştirmektedir. Bu durum, alyazı üretimi esnasında, sadece üretilecek kelimeye odaklanılmasına ve üretilecek cümledeki kelimelerin

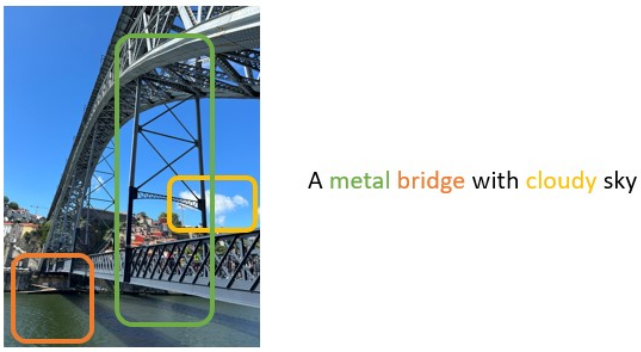


Şekil 3.7: Difüzyon Model

birbirleriyle bağlantılarının göz ardı edilmesine neden olabilmektedir. Probleme çözüm olması amacıyla tekrarlayan sinir ağları kullanılabilir ancak bu ağların dahi uzun vadedeki bağlılıkları koruyamadığı bilinen bir açıktır. Başka bir problem ise cümledeki kelimelerin önem derecelerinin birbirilerinden farklı olmasıdır. Cümle üretilirken, oluşturulan bazı kelimeler anlam ve tahminleme bakımından yetersiz bilgi sağlamaktadır. Dolayısıyla altyazı üretimi sırasında bir kelime üretilirken, sadece bir önceki kelime değil, önceden üretilen serinin tümü, genel olarak dikkate alınmalıdır.

Dikkat modeli bahsedilen problemlerden yola çıkılarak, bütündeki bağlantıyı kaybetmeden parçaya odaklanıp iyi bir tahminde bulunmayı hedeflemektedir. İnsan beyninden yola çıkılarak geliştirilmiş bir girdi işleme tekniğidir. Karmaşık görevleri, küçük parçalara bölerek ve sıralı olarak işleme alan önemli bir tekniktir. Her bir bölüme ayrı ayrı dikkat edilmekte ve her biri ayrı ayrı ele alınmaktadır. Böylelikle karmaşık sorunları çözmedeki işlem maliyeti, küçük problemlere indirgenip ele alınarak düşürülmekte ve daha detaylı inceleme yapıldığından başarı artırılmaktadır.

Görüntü verilerine dikkat modeli uzamsal (spatial) dikkat yöntemiyle (Xu & Saenko, 2016), metinsel verilere ise anlamsal (semantic) dikkat modeli (Cheng



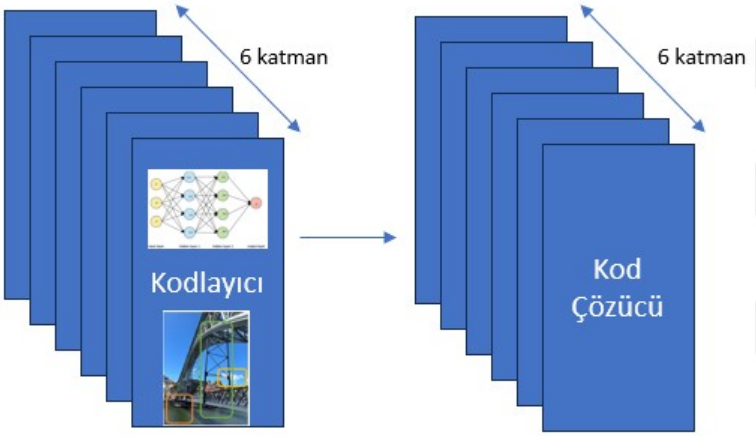
Şekil 3.8: Dikkat Modeli (Attention Model)

et al., 2020) ile uygulanmaktadır. Şekil 3.8'de verilen örnek görüntü üzerinde dikkat edilen kısımlar için üretilmesi beklenen kelimeler gösterilmiştir. Dikkat modeli ayrıca, sert (hard) ve yumuşak (soft) dikkat (Xu et al., 2015) olarak adlandırılan yöntemler ile de gerçekleştirilebilmektedir. Sert dikkat modelinde, girdi olarak verilen görüntü üzerindeki belirli bölgeler seçilir, altyazı üretiminde bu bölgelere odaklanılır ve bu bölgeler dikkate alınır. Yumuşak dikkat modelinde ise bölgeler seçilir ve bölgelerin önem derecelerine göre bir ağırlık değeri verilir. Bu yöntemin bir artısı, yapay sinir ağlarındaki geri yayılım yöntemiyle ağırlıkların güncellenip, görüntüde seçilen bölgelerin altyazı üretimine etkisi artırılıp azaltılabilmektedir.

3.7 Dönüştürücü Model (Transformer Model)

Dönüştürücü model; LSTM, RNN gibi sıralı verilerle çalışma üzerine geliştirilmiş sinir ağlarıdır. İleri beslemeli olarak gerçekleştirilmiş bu ağlar, konuşma tanıma, metinden konuşmaya dönüştürme, makine çevirisi, cümle sınıflandırma, metin oluşturma gibi doğal dil işleme alanında yaygın olarak kullanılmaktadır. Dikkat modelini de bünyesinde barındıran bu modeller, sıralı verileri paralel olarak işleyebilme yetenekleri sayesinde işlem süresini kısaltmaktadır. Ayrıca dönüştürücü model, kodlayıcı - kod çözücü sistem mimarisinde yaygın olarak kullanılan yöntemlerden biridir.

Dönüştürücülerin yapısında standart olarak altı katmanlı kodlayıcı ve kod-çözücü bulunduğu bilinmektedir. Kodlayıcı bölümü içerisinde, öz dikkat modeli (self-attention) ve ileri beslemeli bir ağ bulundurmaktadır. (Şekil 3.9) Bu yapıyı daha da iyileştirerek girdiyi, hem baştan sona hem de sondan başa analiz edebilmeyi sağlayan BERT (Bidirectional Encoder Representations from Transformers) modeli GoogleAI tarafından geliştirilmiştir. Rakip olarak OpenAI, GPT (Generative Pre-Trained Transformers) modelini ortaya atarak, tek yönlü girdi işleme sağlayan bir sistem sunmuştur. Bu sistem, sürekli olarak cümledeki sonraki kelimeyi tahmin etme üzerine geliştirildiğinden metin oluşturma konusunda yüksek bir performans sergilemektedir.



Şekil 3.9: Dönüştürücü (Transformer) Model

Doğal dil işleme alanında yaygın olarak kullanılan dönüştürücü modellerin, bilgisayarlı görü alanına da uyarlanabilen tipleri mevcuttur. ViT (vision transformer) model (Kwok, 2023), BERT modelini temel alan, 2021 yılında ortaya atılmış bilgisayarlı görü alanında kullanılan bir modeldir. Görüntüyü yamalara (patch) ayırıp dikkat modelini kullanarak bölümler arası ilişkiyi, yerel ve genel olarak tutarak öznitelik çıkarımına yardımcı olmaktadır.

Ayrıca dikkat mekanizması kapsamında, öz (self) dikkat ve çapraz (cross) dikkat mekanizmaları tanımlanabilmektedir. Bu iki yöntem, dönüştürücü modellerde bir sıralı girdideki farklı bölümlere odaklanabilmeyi sağlamaktadır.

Öz dikkat mekanizması, hem kodlayıcı hem de kod çözücü bölümde kullanılabilen, modelin sıralı girdi verisindeki her bir bölümüne bakarak diğer bütün bölümler içerisindeki önemini anlayabilmeyi sağlamaktadır. Böylelikle uzun süreli bağımlılıkları yakalayabilmeyi sağlamaktadır. Çapraz dikkat mekanizması ise özellikle kod çözücü bölümde kullanılan bir yöntemdir. Çıktı üretilirken, girdideki farklı bölümleri kontrol edebilmeyi sağlamaktadır. Böylelikle çıktı üretilirken kodlayıcıdan gelen verilerdeki her bir bölüm dikkate alınmaktadır.

3.7.1 Bağlamsal Olmayan Kelime Gömme (Non-Contextual Embedding)

Makineler, gerçek insan dilini, metinsel ifadeleri hatta sözcükleri anlayamazlar. Bu tür yapıları, makine dilinde temsil edebilmek amacıyla vektörler kullanılmaktadır. Kelimeleri, vektörlere dönüştürme işlemi, kelime gömme (word embedding) olarak adlandırılmaktadır. Dönüştürme işlemi, girdinin içeriğine bağımlı ya da bağımsız olarak gerçekleştirilebilmektedir.

Word2Vec, GloVe (Global Vectors for Word Representation) yöntemleri metni, metnin içeriğinden bağımsız bir şekilde kodlamayı sağlamaktadır. İleri beslemeli sinir ağı tabanlı bu modeller, kelime kodlamalarını kullandıkları veri tabanlarında bulmaya çalışmaktadırlar. Temel olarak kelimelerin birlikte görülme sıklıkları yakalanmaya çalışılmaktadır. Sistemde kelimelerin anlam bakımından birbirlerine yakınlığına dikkat edilmektedir. GloVe'da bir kelimenin, global olarak, tüm metinde nasıl ve ne kadar yer aldığına, Word2Vec'te ise yerel bağlamda sadece komşu kelimelerle ilişkisine bakılmaktadır.

3.7.2 Bağlamsal Kelime Gömme (Contextual Embedding)

İçerik bağımlı kelime kodlama, kodlamanın kelime kelime değil bütün metne bakılarak yapılmasıdır. Kodlanan değer, o metnin bütününe özgü bir anlam barındırmaktadır. Her bir kelimenin, metnin içerisinde ifade ettiği anlama bakılmaktadır. Bu sistemlerde, kelimelerin birbirleri arasındaki ilişkinin

çıkarımı ve bu ilişkinin kullanımı söz konusu olmaktadır. Yapay sinir ağlarını kullanan bu sistemler kodlayıcı - kod çözücü sistem mimarisinde yaygın olarak kullanılan yöntemlerden biridir.

Kodlama işleminin başarısının, içeriğe bağımlı kodlayıcı modeller karşılaştırıldığında farklı olduğu gözlemlenmektedir. (Ethayarajh, 2019) Bilinen içeriğe bağımlı kodlayıcı modellerden biri olan ELMo, aynı metin içerisindeki kelimeleri kodlarken, kelime kodlarının birbirlerine daha çok benzediği bilinmektedir. Daha üst katmanlarda yaptığı kodlama işlemi sayesinde, kelime kodlamaları arasındaki BERT bu benzerliği düşürmektedir. GPT-2'de ise rastgele seçilmiş farklı iki metindeki kelimeler kadar birbirlerine az benzemektedir.

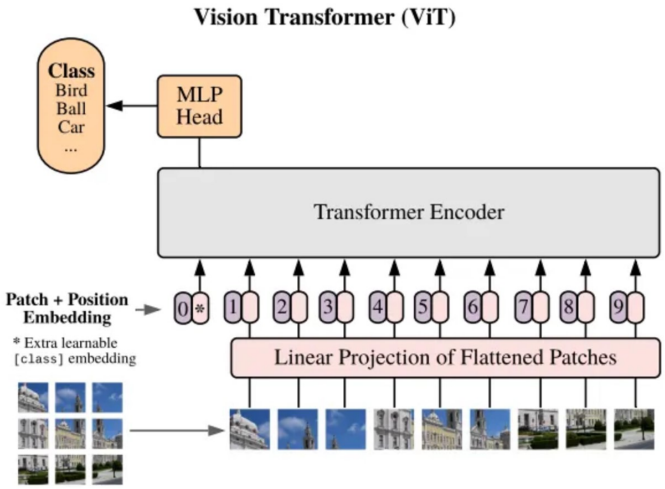
İçeriğe bağımlılık bulunduğundan kelime vektörleri yerine ifadenin tümüne özgü bir vektör tanımlanmaktadır. Bu da kodlanan vektörün, barındırdığı bilgi açısından daha zengin olmasını sağlamaktadır. İçeriğe bağımlı kodlama yapan sistemler, metindeki kelimeler arasındaki ilişkiyi görebilen dönüştürücü (transformer) model mimarisini kullanmaktadırlar.

BERT bilinen en popüler içerik bağımlı gömme yöntemlerinden birisidir. Verilerdeki eksik kelimeleri görmek için dil modellemeyi kullanacak şekilde eğitilmiştir. (Alammary, 2022) BERT, kelimeleri tahmin edebilen çift yönlü bir model olarak tasarlanmıştır. Büyük miktarda metin verisi kullanılarak önceden eğitilmiştir ve bu yüzden kelime tahminlemedeki performansı yüksektir. Gözetimsiz (unsupervised) yöntemle geliştirilmiş olması, internet üzerindeki milyarlarca etiketsiz metin içeriğiyle eğitilebilmesini kolaylaştırmaktadır. Temelinde, öznitelik çıkarımı ve ince ayar yapmak olan iki temel prensip yer almaktadır. (Devlin et al., 2018) Öznitelik çıkarımı esnasında önceden eğitilmiş BERT modeli herhangi bir model parametresi değiştirilmeden kullanılır. Çıkarılan bu öznitelikler, genel model tarafından kullanıma hazır olmaktadır. Ancak ince ayar kısmında mevcut BERT modeline parametreleri iyileştirilmiş fazladan bir katman eklenmektedir. BERT, doğal dil işleme alanındaki görevler üzerinde başarıyla çalışan bir yöntemdir.

3.8 ViT (Vision Transformer Model)

Dönüştürücü modeller, büyük ölçekte ve herhangi bir alana bağlı olmayan, genel veri setleri ile önceden eğitilmiş bir model sunmaları nedeniyle öznitelik çıkarımındaki başarılarını arttırdığından, gelişmiş performansa sahip ve çok yönlü olmaları sebebiyle güncel çalışmalarda yaygın olarak tercih edilen modellerdir. Aktarmalı öğrenmeyi bünyelerinde barındırmaları, hem doğal dil işleme hem de bilgisayarlı görü sistemlerinde kullanılmalarını kolaylaştırmaktadır.

BERT modeline benzer şekilde tasarlanmış görü dönüşüm modeli, 2020 yılında ortaya çıkmıştır. (Dosovitskiy et al., 2020) Şekil 3.10'da da görülebileceği üzere, metinsel parçalar yerine görüntü eşit boyutlu parçalara bölünerek sıralı bir girdi oluşturulmaktadır. Görüntünün parçaları, görü dönüşüm modelinin kodlayıcısında girdi olarak kullanılabilecek şekilde sıralanıp kodlanarak dönüştürülmektedirler.



Şekil 3.10: ViT Modeli (Vision Transformer Model)(Fields & Kennington, 2023)

Görü Dönüşüm Modeli (Vision Transformer Model), dönüştürücü model mimarisini bünyesinde barındırarak, bilgisayarlı görü sistemlerinde gözle görülebilir iyileştirmeler sağlamıştır. Bu sistemin amacı, görüntüyü küçük parçalara bölüp işleme alarak görüntü analizinde yeni bir yöntem ortaya atmaktır. Mimarisinde öz dikkat mekanizmasını (self-attention) barındırmaktadır. Böylelikle,

görüntüdeki yerel ve global ilişkilerin tümünü yakalayabilmeyi sağlamaktadır.

3.9 GPT (Generative Pre-trained Transformer)

Üretken ön eğitimli dönüştürücü (GPT) model, kodlayıcı-kod çözücü sistem mimarisinde, kod çözücü olarak kullanılabilen yaygın bir yöntemdir. Aktarmalı öğrenmenin avantajını kullanarak geliştirilmiş bir dil modelidir. Genel amaçlı bir veri setiyle önceden eğitilmiş bu model için parametre optimizasyonu da gerçekleştirildiğinden doğrudan kullanıma hazırdır. Bünyesinde dönüştürücü tabanlı bir kod çözücü barındırmaktadır. Girdiyi, BERT'ten farklı olarak tek yönlü işlemektedir. GPT-2 modeli ise GPT'nin gelişmiş bir versiyonudur. İşlediği girdi boyutu (batch size), sahip olduğu kod çözücü blok sayısı, eğitildiği veri seti bakımından farklılıklar göstermektedir.

Öğreticisiz öğrenme yöntemi ile önceden eğitilmiş bu modellerin, üretme ve çıkarım konusunda daha başarılı olduğu bilinmektedir. Soru cevaplama, dil çevirisi, metin özetleme gibi doğal dil işlemedeki hemen hemen her alana uyarlanabilmektedir. Şekil 3.11'de de gösterilen katmanlı mimarilerinde; girdileri nümerik vektörlerle kodlamayı sağlayan bir girdi kodlama katmanı, öz dikkat (self-attention) mekanizmasını kullanan ve öznitelik çıkarımını sağlayan dönüştürücü kodlayıcı katman, girdinin sırasını da göz önünde bulundurarak işleme almayı sağlayan pozisyon kodlayıcı katman, girdileri indirgeyip çıktıyı normalleştiren bir normalleştirme katmanı, girdiden çıktıya dönüşümü sağlayan doğrusal katman ve son olarak çıktıları tutan bir softmax katmanı yer almaktadır. (Hirway et al., 2023)

Şekil 3.11: GPT Modeli

4 VERİ ARTIRIMI (DATA AUGMENTATION)

Veri artırımı, modeli eğitmek için kullanılan verinin benzer özellikteki verilerin üretimi yoluyla çoğaltılmasıdır. Varolan verinin, üretilen sentetik veriyle birlikte modeller üzerinde kullanıldığında, genel olarak modellerin performansını arttırdığı gözlemlenmektedir.

Özellikle derin öğrenme modellerinin performanslarının, eğitildikleri veri setiyle doğru orantılı olduğu bilinmektedir. Geleneksel makine öğrenmesi yötemlerinde olduğu gibi derin öğrenmede de eğitim ve test veri setlerine ihtiyaç duyulmaktadır. Modelin eğitimi, eğitim verisi ile gerçekleştirilmektedir. Eğitim işlemi sırasında ağ, ilgili ağ parametrelerini güncelleyerek kendisini iyileştirmekte ve en iyi hale getirmektedir. Eğitim işlemi bittikten sonra modelin performansını ölçümlemek için test veri seti kullanılmaktadır.

Modelin performansı eğitildiği veri setine bağlıdır. Veri setindeki çeşitlilik ne kadar fazla olursa öğrendiği bilgi ve parametreler de o derece fazla ve çeşitli olmaktadır. Ancak veri seti yetersiz olursa, model yeterince öğrenememekte (underfitting) ve bununla doğru orantılı olarak performansı düşmektedir. Oluşabilecek bir başka problem ise modelin veri setine aşırı uyum (overfitting) sağlamasıdır. Veri çoğaltma uygulanarak, veriler artırılıp çeşitlendirilebilmekte ve bu problemlerin üstesinden gelinebilmektedir.

Veri artırma bir düzenlileştirme (regularization) stratejisidir. (Shorten, 2021)(Srivastava et al., 2014)(Kukačka et al., 2017) Seyreltme (dropout) veya ağırlık cezaları gibi diğer düzenlileştirme teknikleri de mevcuttur. Bu teknikler, ağın ara aktivasyonlarına gürültü ekleyerek veya işlevsel forma kısıtlamalar getirilerek düzenlileştirmenin uygulanmasını sağlamaktadır.

Bir görüntüyü çoğaltarak yeni bir görüntü oluşturmak için görüntünün noktalarını farklı bir konuma eşleştiren veya görüntü yoğunluk değerlerini

değiştiren bir yöntem izlenebilmektedir. (Chlap et al., 2021) Üretilen veri, mevcut veri kümesine eklenerek veri havuzu genişletilmektedir. Literatürde bir çok veri artırma tekniği mevcuttur. Bu teknikler; ölçekleme, döndürme, çevirme (translation), perspektif dönüşümleri gibi verinin koordinat dönüşümlerine dayalı özelliklerini değiştirirek uygulanan geometrik dönüşümler, girdi görüntülerinin bazı kesitlerinin alınıp boyutlarının artırılıp sisteme tekrar verilmesi ya da bu kesitin görüntüden kaldırılarak sisteme tekrar verilmesi, yoğunluk değerleri değişimi ve filtrelerin uygulanması, gürültü eklenmesi, iki girdi verinin birbirlerine eklenmesi ya da çıkarılması yoluyla yeni veri elde edilmesi ve sistemde kullanılması, difüzyon model kullanılarak sahte görüntülerin üretilmesi olabilmektedir.

4.1 Küçük Veri Setleri (Small Datasets)

Derin öğrenme alanında başarı sağlayabilmek için güçlü miktarda veriye ihtiyaç duyulduğu bilinmektedir. Ancak her alanda çok veriye sahip olmak mümkün olmamakta ve küçük veri setleriyle çalışmak zorunda kalınabilmektedir. Bu kapsamda küçük veri setleri ile derin öğrenmede başarı sağlayabilmek adına yapılan çalışmalar giderek önem kazanmaktadır.

Aktarmalı öğrenme ile eğitilen modeller, hali hazırda genel bir veri seti ile önceden eğitildiklerinden, daha küçük veya belirli veri kümeleri üzerinde eğitime devam etmek için kullanılabilir. Ancak yine de veri setinin yetersiz olması yetersiz öğrenme, aşırı uyum, elde edilen doğruluk değerinin düşük olması, gürültü ve sonuç olarak düşük performanslı sistemlere neden olmaktadır. Veri artırımı bu noktada genel bir çözüm yöntemi olarak karşımıza çıkmaktadır. Veri setinin boyutunu ve çeşitliliğini artıran bu yöntem sistemler açısından kullanışlı olmaktadır.

Küçük veri setleriyle çalışabilmek için alternatif yollar mevcuttur. (Fong et al., 2020) (Lateh et al., 2017) (Shaikhina et al., 2017) (Andonie, 2010) (Slifker & Shapiro, 1980) Bunlardan bir tanesi veri setindeki eğitim verilerini

çoğaltarak veri setini genişletmektir. İkinci bir yol ise toplu olarak tahmin sonuçları oluşturmak için bir topluluk öğrenme yöntemini kullanmaktır ve en düşük hataya sahip birkaç algoritmadan birinin tahmin sonucu kullanılmak üzere seçilmektedir. Adayların geri kalanından alınan sonuçlar ise göz ardı edilmektedir. Üçüncü bir yaklaşım, tek bir tahmin algoritmasına odaklanmaktır ve bu yaklaşım genellikle ayarlanacak birkaç parametre ile gelir. Ortaya çıkan modelin doğruluk düzeyi, parametre ayarına çok duyarlı olma eğilimindedir. Bu tür bir algoritma için varsayılan parametre değerleri genellikle maksimum performansı sağlamaz, doğruluk seviyesini artırmak için parametre değerlerinde ince ayar yapılması gerekir.

4.2 Görüntü Veri Artırımı

Yapay zekanın kazandığı önem dolayısıyla bu alanda yapılan çalışmalar popülerlik kazanmış ve artmıştır. Bu alanda kullanılabilecek en yaygın veri tiplerinden biri de görüntülerdir. Ancak her çalışma alanında, istenen boyutta görüntü verisine sahip olunamayabilmektedir.

Görüntüleri artırmak için kullanılan bir çok farklı teknik mevcuttur. (Shorten, 2019) Bunlar yatay kaydırma, renk alanı geliştirmeleri ve rastgele kırpma gibi basit dönüşümler olabileceği gibi geometrik dönüşümler, renk uzayı dönüşümleri, çekirdek filtreleri, görüntüleri karıştırma, rastgele silme, özellik alanı büyütme, çekişmeli eğitim, GAN tabanlı büyütme, nöral stil aktarımı ve meta-öğrenme şemaları gibi görüntüler üzerinde daha detaylı çalışmalar gerektiren yöntemler de olabilmektedir. Geometrik dönüşümler kapsamındaki işlemlerin bazıları; kaydırma, renk uzayı dönüşümleri, kırpma, döndürme, tercüme, gürültü enjeksiyonu gibi gerçekleştirilmesi kolay ve etkili tekniklerdendir.

Basit görüntü işleme teknikleri, görüntü artırımı kapsamında çok yardımcı olmaktadır. Görüntüler yatay eksende çevirme (flipping), rastgele ya da kurallı olarak (örneğin sadece merkezden bir kısım seçilip) kırpılarak (cropping),

görüntü belirli bir eksende, 1° ile 359° arasında döndürülerek (rotation) yeni görüntülerin üretilmesi işlemi gerçekleştirilebilmektedir. Ayrıca görüntünün matris temisili kullanılarak; R,G,B renk tonları değiştirilerek, görüntünün parlaklığı artırılıp azaltılarak, histogram eşitlemesi yapılarak basit matris işlemleri ile de yeni görüntülerin üretilebilmesi söz konusudur.

Bilinen ve yaygın olarak kullanılan başka bir yöntem olan gürültü enjeksiyonu, varolan görüntülerin değiştirilip tekrar kullanılabilmesi için zekice geliştirilmiş bir tekniktir. Genellikle bir Gauss dağılımından alınan rastgele değerler matrisinin, görüntü matrisine enjekte edilmesinden oluşmaktadır. Görüntülere gürültü/parazit eklemek, CNN'lerin daha sağlam özellikleri öğrenmesine yardımcı olabilmektedir. Bir başka etkili yöntem olan tercüme yöntemiyle; görüntüleri sola, sağa, yukarı veya aşağı kaydırmak, verilerde konumsal yanlılığı önlemek için çok faydalı bir dönüşüm olabilmektedir. Örneğin, yüz tanıma veri kümelerinde yaygın olan bir veri kümesindeki tüm görüntüler ortalanmışsa, bu, modelin mükemmel şekilde ortalanmış görüntüler üzerinde de test edilmesini gerektirmektedir. Orijinal görüntü bir yönde çevrilirken, kalan boşluk 0'lar veya 255'ler gibi sabit bir değerle veya rastgele Gauss gürültüsü ile doldurulabilmektedir. Bu dolgu, büyütme sonrası görüntünün uzamsal boyutlarını koruma altına almaktadır.

4.3 Metinsel Veri Artırımı

Metinsel veriler, karakteristiksel özellikleri dolayısıyla görüntülerden farklıdır ve bu verilere görüntülere uygulanan veri artırma yöntemlerini uygulamak mümkün olmamaktadır. Metin içerisindeki kelimelerin lokasyonları önemlidir ve dolayısıyla herhangi bir modifikasyon uygulandığında cümle anlam değişikliğine uğrayabileceği için risklidir.

Bazı yaklaşımlar, cümlenin morfolojik özelliklerini değişikliğe uğratmaya dayanırken bazıları dil ve dil modellerine odaklanmaktadır. Başka bir tanımlamayla metinsel veri çoğaltma sembolik (symbolic) ve sinirsel (neural) çoğaltmalar

olarak iki başlık altında incelenebilmektedir. (Shorten, 2021) Bu iki yöntem arasındaki temel fark sembolik çoğaltmada sembolik kuralların kullanılması ve sinirsel çoğaltmada yardımcı sinir ağlarının veya diğer istatistiksel model türlerinin kullanılmasıdır. Sembolik çoğaltmanın önemli bir faydası, insan tasarımcı için yorumlanabilir olmasıdır. Sembolik büyütmeler ayrıca, artırılmış örnekler oluşturmak için sözcükleri veya tümceleri değiştirmek gibi kısa dönüşümlerle daha iyi çalışmaktadır. Literatürde yer alan yaklaşımlardan biri olan EDA (easy data augmentation) tekniğinde (Wei, 2019), metindeki kelimelere odaklanılmaktadır. Metin içerisinden rastgele kelime çıkarılarak ya da metin içerisindeki mevcut kelimeleri eş anlamlılarıyla yer değiştirerek, metne rastgele kelime eklenerek, metindeki kelimelerin yerleri değiştirilerek metinsel veri artırımı uygulanmaktadır. Bu işlemler rastgele yapılabileceği gibi kural tabanlı olarak da yapılabilmektedir.

Sinirsel çoğaltma kapsamında uygulanabilecek yöntemlerden biri metne çeviri uygulamaktır. Bu yöntemle, metin önce bir başka dile, sonrasında tekrar orijinal dile çevrilerek aynı anlam ifade eden yeni metin elde edilmektedir. Her metnin kendisine özgü bir stili bulunmaktadır. Metindeki bu özellik kullanılarak, aynı stile sahip başka bir metin üretimi, stil aktarımı yöntemiyle çoğaltma ile gerçekleştirilebilmektedir. Bilinen bir başka yöntem ise GPT, GPT-2, LSTM ya da GAN tabanlı kodlayıcı-kod çözücü sistemler kullanılarak önceden eğitilmiş modeller vasıtasıyla metinler üretmektir. BERT, GloVe, Word2Vec gibi kodlayıcılar kullanılarak metinde yer alan bir kelimenin eş anlamlısının, ilgili kodlayıcının veri tabanından bulunup yer değiştirilmesiyle de veri artırımı gerçekleştirilebilmektedir. (Csanyi & Orosz, 2021) Bahsedilen tüm bu yöntemleri, görüntü altyazı üretimi kapsamındaki metinsel verilere uygulayabilmek mümkündür.

4.4 Difüzyon Model ile Veri Artırımı

Çekişmeli üretici ağlar, 2014 yılında geliştirilmiş, sahte veri üretimi üzerine özelleşmiş ağlardır. (Goodfellow et al., 2014) Girdi olarak verilen veriyi en

iyi şekilde taklit ederek sahte veri üretimi gerçekleştirilmektedir. GAN'larla üretilen verilerin kalitesinin yüksek olması, kullanımlarını yaygınlaştırmıştır.

Çalışma prensipleri bakımından GAN'lara yüksek oranda benzerlik gösteren bir başka yöntem difüzyon modelidir. Difüzyon model metinsel ifadelerin görüntülere dönüştürülmesini desteklemesi sebebiyle tercih edilebilmektedir. Gereksiz detayları göz ardı ederek GAN'lara kıyasla daha kaliteli görüntüler üretilebilmektedir.

Difüzyon model, kapsamlı açıklamalı eğitim verileri olmadan, derin ilişkileri öğrenmenin bir yolunu bulmaktadırlar. İçerdikleri metin kodlayıcı sayesinde girdi olarak verilen metin kodlanır. Sonrasında difüzyon katmanı eklenen gürültü ile yeni bir görüntünün sentezlenmesini sağlamaktadır. Son aşamada ise görüntü kod çözücüsü, çıktı görüntünün üretilmesini sağlamaktadır. Veri çoğaltma alanında da yaygın olarak kullanılan bu ağların ürettikleri veriler son derece başarılı olmakta ve hatta bazen gerçek verilerden ayırt edilmeleri zor dahi olabilmektedir.

Geleneksel makine öğrenmesi yöntemlerinde de kullanılan, görüntünün döndürülerek, kesilerek, kaydırılarak, renk değerleriyle oynanarak fiziksel özelliklerinin değiştirilip, veri setinin genişletilmesi akla gelen ilk yöntemlerdendir. Ancak genişletilmiş veri setindeki bu veriler halen bir takım benzerlikler barındırmaktadır. Difüzyon model ise tamamen yepyeni bir görüntünün üretilmesini sağlayarak veri setini genişletmektedir. Bu kapsamda kullanılan tekniklerden biri görüntü stili aktarımıdır. Görüntüye ait stilin sinir ağları vasıtasıyla anlaşılıp kullanılarak yeni görüntülerin elde edilmesi işlemidir.

5 ÖNCEKİ ÇALIŞMALAR

Mucizevi bir varlık olan insan, hakkında her gün yeni bir bilginin keşfi sayesinde bilimsel araştırmaların merkezinde bir konumdadır. Keşfedilen her yeni bilgi hayranlık uyandırmakta ve sanal ortamlara uyarlanmaya çalışılmaktadır. Sayısız insan davranışından sadece biri olan öğrenme, makine sistemlerine entegre edilerek en iyileştirilmeye çalışılmaktadır. Bu kapsamda derin öğrenme, insan beynini taklit eden yapısıyla en merak uyandıran popüler alanlardan biridir. (Sudhakar et al., 2022) tarafından yapılan çalışmada, insanoğlu çok miktarda bilgiyi bir anda işleyebildiği ve bu bilgilerin büyük olasılıkla resimler, videolar ve yazılı formattaki herhangi bir şey olduğundan bahsedilmektedir. Her görüntü, insanların onu deşifre edip işlediği büyük miktarda bilgiye sahiptir ve bir görüntüyü tanımlamak için dil kullanılmaktadır. Ancak görüntüler için altyazı üretmek günümüz dünyasındaki makineler için oldukça zahmetli bir iştir ve bir makine kullanarak altyazı oluşturmak; doğal dil işlemeye, farklı nesneleri ayırt etmeye ve bunları ilişkilendirmeye ilişkin temel bir anlayışı içermektedir. Çalışmada derin öğrenme kullanılarak görüntüler için altyazı üretmek amaçlanmıştır. Görüntüyü algılayan ResNet50 CNN modeli ve altyazı üretimi için RNN birlikte kullanılmıştır. Flickr8K veri seti ile denenen bu sistem ile ResNet50 modelinin ortalama doğruluk değeri %79 olarak hesaplanmıştır ve VGG16 modelinin ürettiği %29 doğruluk değerinden daha başarılı bir sonuç ürettiği görülmüştür.

CNN-RNN modelini kullanan bir başka çalışma (Chang et al., 2016) tarafından geliştirilen Im2Txt modeldir. Bu sistem verilen bir görüntünün, eşleniğinin bulunması ve bu eşleşen görüntüye ait altyazının bu yeni görüntü için kullanılmasıdır. Dolayısıyla üretme (generative) tabanlı değil bulup getirme (retrieval) tabanlı çalışmaktadır. Çalışmada makine çevirisine benzer şekilde, resim yazısı oluşturmaya yönelik bir diziden diziye yinelenen sinir ağları (RNN) modeli kullanılmaktadır. Tüm görüntünün evrişimli sinir ağları (CNN) modeli ile temsil edildiği mevcut çalışmaların çoğundan farklı olarak, giriş

görüntüsünü, RNN modelinin kaynak dizisi olarak hizmet verecek şekilde tespit edilen nesnelerin bir dizisi olarak temsil etmek önerilmiştir. Sistem ayrıca kullanıcıların bir görüntüdeki belirgin nesneleri tespit etmesine ve benzer görüntüleri ve karşılık gelen açıklamaları bir veritabanından almasına olanak tanımaktadır.

Bulup getirme tabanlı geliştirilmiş çalışmalardan bir diğeri, RNN ve LSTM tekniklerini birleştirerek, (Vinyals et al., 2014) tarafından ortaya atılan altyazı üretim modelidir. Bu model, altyazı üretimi ve görüntüleri tanımadaki son gelişmeleri kullandığından, görüntüler için altyazı üretiminde referans alınan modellerden biridir. Çalışmada, görüntülerden altyazı üretmek için sinirsel ve olasılıksal bir çerçeve birleştirilerek kullanılmıştır. Kodlayıcı olarak RNN'yi derin evrişimli sinir ağı (CNN) ile değiştirmek önerilmiştir. Böylece değişken uzunluklu bir girdiyi sabit boyutlu bir vektöre kodlayan tekrarlayan bir sinir ağından yararlanılmaktadır. Kod çözme bölümünde, bir giriş cümlesi verildiğinde doğru çeviri olasılığını en üst düzeye çıkarmak amaçlanmıştır. Modelde ayrıca LSTM tabanlı bir cümle üreticisi kullanılmıştır. Girdi, unutma ve çıktı kapıları yardımıyla RNN'lerdeki kaybolan ya da patlayan gradyan problemine çözüm getiren bu mimari üretilecek altyazıdaki her bir kelimeyi, metinde üretilmiş önceki kelimelerin tümünü de dikkate alarak üretmektedir. Çalışmada BLEU-1 skor değeri ile Flickr8K veri seti üzerinde ölçümlenen değer %58 olmuştur.

Görüntü altyazı üretiminde kullanılabilecek yöntemlerden bir diğeri şablona uyum (template base) yöntemidir. Bu yöntemde, cümle oluşturma kalıpları bir şablon olarak kullanılmaktadır. Görüntüden bilgi çıkarımı şablona uymak adına nesne ve fiillerin çıkarımı olarak gerçekleştirilmektedir. Sonrasında çıkarılan bu bilgiler cümle oluşturma kuralları çerçevesinde bağlaçlarla birleştirilmektedir. (Li et al., 2011) tarafından geliştirilen bu çalışmada, bilgisayarlı görü tabanlı girdiler verilmiştir ve web ölçeğinde n-gramlar kullanılmıştır.

İnternetteki görsel ve multimedya verilerinin giderek artması, multimedya analizi ve bilgisayarlı görme alanında görsel içerik anlayışının gerekliliğini doğurmuştur. Görsel analiz ve içeriğin anlaşılması, bir görüntüden daha fazla kullanılabilecek metinsel açıklamalar oluşturmayı amaçlayan (Tiwari & Bhatnagar, 2021) tarafından yapılan bir başka çalışmada CNN-RNN mimarisi kullanılmıştır. Ancak RNN, görselde göze çarpan şeyleri belirginleştirme ve dolayısıyla görsele daha iyi altyazılar üretebilmeyi sağlayan dikkat (attention) yöntemiyle iyileştirilmiştir. Dikkat mekanizmasının kullanılmadığı sistemlerde, üretilen altyazıdaki bir sonraki kelimenin gizli durumunu ve oluşturulmasını sağlamak için görüntünün tamamı dikkate alınmakta ve bu nedenle görüntüdeki farklı nesnelerin anlamlarını ve özünü yakalamak için görüntünün farklı bölümlerine odaklanılamamaktadır. Bu problemin önüne geçebilmek için geliştirilmiş olan sistemde CNN, görüntüdeki her bölüm için farklı gizli durumlar üretmektedir ve RNN daha sonra dikkati, cümledeki bir sonraki kelimenin oluşturulmasıyla ilgili olarak, görüntünün bölümlerine odaklanmak için kullanmaktadır. Modelin başarısını ölçümlemek için BLEU skor değeri kullanılmıştır ve seçilen bazı örnek görüntüler için elde edilen skor değerlerinin %69 ile %86 arasında olduğu görülmektedir.

Dikkat (attention) mekanizmasını kullanan ve (Kulkarni et al., 2023) tarafından geliştirilen bir başka çalışmada; görüntüler, görüntü bilgilerinin özellik vektörlerine dönüştürüldüğü kodlayıcıya beslenmektedir. Kodlayıcının çıktısı, özellikleri cümlelere çeviren kod çözücüye iletilmektedir. Bu yöntem bilinen temel görüntü altyazı üretim tekniğidir. Ancak bu klasik altyazı üretim yönteminin bir sorunu bulunmaktadır. Çünkü görüntünün uzaysal özelliklerinin dikkate alınması mümkün olamamaktadır. Sonuç olarak, görüntünün hassas veya önemli özellikleri dikkate alınmadan, görüntünün tümü bir sahne olarak dikkate alınarak bir altyazı oluşturulmaktadır. ImageNet veri tabanıyla önceden eğitilerek, ağırlıkları güncellenmiş bir CNN modeli olan InceptionV3, sistemde bir kodlayıcı görevi görmektedir. Görüntülerden öznitelikleri çıkarmakta ve kod çözücü bölüme iletmektedir. Burada kod çözücü olarak

kullanılan GRU bir RNN çeşididir ve amacı cümlenin kodlamasını çözmektir. Modelde ek olarak Bahdanau dikkat modeli (Bahdanau et al., 2014), altyazıları üretirken görüntülerin önemli yönlerine odaklanmasını sağlayarak kod çözücünün yeteneğini geliştirmek için kullanılmıştır. Böylece oluşturulan altyazıda görselin hassas kısımlarının göz ardı edilmemesine dikkat edilmektedir. Sistemde, Flickr8K veri seti kullanılarak elde edilen doğruluk değeri dikkat mekanizması kullanılmadığında %79 iken, dikkat mekanizması da dahil edildiğinde %90 olarak ölçümlenmiştir.

Görüntü altyazı üretimi modelleri ile ilgili (Chandra et al., 2023) tarafından yapılan genel bir çalışmada kodlayıcı kod çözücü sistemler karşılaştırılmıştır. Sistemlerde girdi görüntüler ön işleme tabi tutularak CNN-RNN modeline verilmiştir ve altyazı üretim işlemi gerçekleştirilmiştir. MobileNet, ResNet, Inception, DenseNet, EfficientNet ya da VGG16 gibi CNN modelleri anlamlı özniteliklerin çıkarılmasını sağlamaktadır. Sonrasında kodlanmış özellikler sabit uzunlukta bir vektör halinde düzleştirilmekte veya yeniden şekillendirilmektedir. Kod çözme için LSTM gibi bir RNN modeli kullanılmaktadır. Kod çözme işlemi, gizli durumun başlatılması ve bir başlangıç sözcük belirtecinin üretilmesiyle başlamaktadır. RNN, girdi olarak sözcük belirtecini ve gizli durumu almaktadır, bir çıktı sözcüğü tahmini ve güncellenmiş gizli durum üretmektedir. BLEU skor değeri, Flickr8K veri seti ile VGG16-LSTM için %51, InceptionResNet-LSTM için %48, InceptionV3-LSTM için %54, MobileNet-LSTM için %55, EfficientNet-LSTM için %54 olarak elde edilmiştir.

Görüntü altyazı üretimi için mevcut bir çok yöntem ve alternatif yollar mevcuttur. Mevcut görüntü altyazı üretim modellerinin çoğu esas olarak tüm görüntü özelliklerini temsil eden global dikkati, nesne özelliklerini temsil eden yerel dikkati veya bunların bir kombinasyonunu kullanmaktadır. Görüntünün çeşitli nesne bölgeleri arasındaki ilişki bilgisini entegre edecek az sayıda model bulunmaktadır. Ancak bu ilişki bilgisi aynı zamanda altyazı oluşturmak için de oldukça öğretici ve önemlidir. Bu bilgiler ışığında (Chen et al., 2022)

tarafından yapılan çalışmada öz dikkat (self-attention) mekanizması GAN'lar ile birleştirilerek SC-GAN modeli ortaya atılmıştır. Global ve yerel dikkat mekanizması birlikte kullanılarak farklı nesne bölgeleri arasındaki iç görsel ve anlamsal ilişkiyi keşfedebilen Pyramid Attention modeli oluşturulmuştur. Ayrıca, maruz kalma yanlılığı (exposure bias) problemini hafifletmek ve eğitim sürecini daha verimli hale getirebilmek için, GAN sıralı veri üretiminde (sequence generation) kullanılmıştır.

Dönüştürücü model ve LSTM yakından ilişkilidir. Her ikisi de öğe bazında korelasyonlu ilişki için dikkat mekanizmasını kullanmaktadır, ancak dönüştürücü tabanlı sistemlerin belirgin bir avantajı, dizileri paralel olarak işleyebilme yeteneğidir. Giriş dizilerini sırayla işleyen geleneksel tekrarlayan sinir ağlarının (RNN'ler) aksine, dönüştürücülerin herhangi bir dizi kısıtlaması yoktur, bu da onların dizideki tüm öğeleri aynı anda işleyebilmesine olanak tanımaktadır. Hesaplamanın bu şekilde paralelleştirilmesi, sistemin önemli derecede hızını artırmasına yardımcı olabilmektedir. RNN'ler tarafından, hesaplama açısından mümkün olmayan daha uzun dizilerin işlenmesini mümkün kılmaktadır. BERT, GPT gibi dönüştürücü modeller, sadece doğal dil işlemede değil aynı zamanda bilgisayarlı görü sistemlerinde de yeri olan modellerdir. Görü dönüştürücü (ViT) modeli ise, BERT'in temelleri temel alınarak tasarlanmış bir sistemdir ve ağın mekansal bilgileri ve görüntü yamaları arasındaki korelasyonları yakalamasını sağlayan dikkat mekanizması nedeniyle popülerlik kazanmıştır. Bahsedilen bilgilere dayanarak, kodlayıcı olarak ViT, kod çözücü olarak BERT ve GPT'yi ayrı ayrı kullanan bir kodlayıcı kod çözücü sistem mimarisi (Kwok, 2023) tarafından geliştirilmiştir. Flickr8K veri seti kullanılarak denenen sistemde BERT için yedi devir sonunda, Precision %43, Recall %42 ve F-Measure %41 olarak ölçülmüştür. Aynı skor değerleri GPT kod çözücüsü kullanıldığında sırasıyla %74, %59, %13 olarak gözlemlemiştir.

Veri artırma teknikleri, derin öğrenme (DL) algoritmalarının performansını artırmak amacıyla eğitim verilerinin çeşitliliğini ve kalitesini yükselt-

mek için veri işlemede önemli bir tekniktir. (Shorten, 2019) çalışmalarında, görüntü işlemede veri artırmanın, model doğruluğunu artırmak için etkili bir yaklaşım haline geldiğinden bahsedilmiştir. Veri artırma, veri kümelerini döndürme, kaydırma, kırpma, ölçekleme, geometrik bozulma vb. yoluyla sentetik olarak dönüştürerek manipüle etmektedir. Veri artırmanın amacı, modelin görüntü temsilindeki daha büyük çeşitlilikten ders çıkarabilmesi için veri kümesinin çeşitliliğini artırmaktır. (Alin et al., 2023) tarafından yapılan bir çalışmada aynı model üzerinde veri artırımı olmadan ölçümlenen precision-recall değeri %94 iken görüntü verileri döndürme, kaydırma yöntemleri uygulanarak artırıldığında %95, renk tonu değişimiyle %96, mozaik uygulanmasıyla %99 olarak ölçümlenmiştir. Metin verilerinin arttırılması ile ilgili (Muftie & Haris, 2023) tarafından yapılan bir başka çalışmada, metinsel verilerin artırılması metin sınıflandırma ve duygu analizine yönelik modellerin veya algoritmaların performansını geliştirebildiğinden bahsedilmiştir. Kural ya da model tabanlı artırmanın mümkün olabileceğine değinilmiştir. Eğitim verilerinin sınırlı olduğu durumlarda, daha büyük veriye ihtiyaç duyulan doğal dil işleme yöntemlerinde, metinsel veri artırımının önemli bir odak noktası olduğu bilinmektedir. Çalışmada; CNN kullanılarak ölçümlenen doğruluk değeri, metinsel veri artırımı olmadığında %86 iken, rastgele kelime eklenerek veri artırımı sonrasında bu değer %87, çalışmada gerçeklenen model ile %88 olmuştur. Bi-LSTM modeli ile ölçümlenen doğruluk değeri ise metinsel veri artırımı olmadığında %83 iken, rastgele ekleme ve çalışmada gerçeklenen model ile metinsel veri artırımı sonrasında bu değer %86 olarak elde edilmiştir.

Difüzyon model özellikle metinsel verilerden görüntü üretimini desteklemesi sebebiyle veri artırımı alanında da kullanılabilmektedir. Kararlı difüzyon modelini kullanan (Yidan et al., 2014) tarafından gerçekleştirilen çalışmada, akciğer film görüntüleri için görüntü veri artırımında kullanılmıştır. Metin kodlayıcı, difüzyon ve görüntü kod çözücüyü içeren ve üç ana bölümden oluşan kararlı difüzyon modeli metinden görüntü sentezleme amacıyla kullanılmıştır. Difüzyon model ile veri artırımını kullanan bir başka çalışmada, Flickr30K

veri seti üzerinde görüntü ve metin artırımı gerçekleştirilmiştir. ((Medina & Alejandro, 2024)) Görüntülerin artırımında kullanılan bu yöntemin model performansını artırmadığı ancak metin verilerinin artırıldığı durumda, model performansının BLEU skor değerine göre %0.7 arttığı görülmüştür. Benzer bir çalışmada ((Xiao et al., 2023)), MS COCO veri seti kullanılmış ve BLEU, METEOR, ROUGE ve CIDEr metriklerine göre çoklu model veri artırımının model performansına pozitif yönde etki ettiği görülmüştür.

6 KÜÇÜK VERİ SETLERİ İÇİN ÖNERİLEN GÖRÜNTÜ ALTYAZILAMA MODELLERİ

Bu bölümde, doktora tezi kapsamında gerçekleştirilen çalışmalardan ve önerilen yöntemlerden bahsedilmektedir. Çalışmamız öncelikli olarak küçük veri setleri ile ilgili olduğundan büyük boyutlu veri setleri yerine, bilinen veri setlerinden Flickr1K, Flickr8K ve Flickr30K tercih edilmiştir. İlk olarak farklı kodlayıcı ve kod çözücü model bileşimleri önerilerek, bu veri setleri üzerindeki performans değerleri elde edilmiştir. Bu sonuçlara Deneysel Çalışmalar bölümünden ulaşılabilir. Bu şekilde belirlenen en iyi model kullanılarak, seçtiğimiz en küçük veri seti olan Flickr1K üzerinde hem görüntüler hem de altyazılar üzerinde farklı veri artırım stratejileri uygulanarak, küçük veri setleri için derin öğrenme tabanlı başarılı bir altyazılama yaklaşımı sunulmuştur.

6.1 Veri Seti

Çalışmaya, kullanılacak veri setlerinin netleştirilmesi üzerine araştırmalar yapılarak başlanmıştır. Görüntüler için altyazı üretimi konusunda bulunan güncel derleme makalelerden elde edilen veriler ışığında, çalışmaya Flickr1K (Young et al., 2014) & (Garg, 2020), Flickr8K (Rashtchian et al., 2010) ve Flickr30K (Young et al., 2014) veri setleri ile çalışılmaya karar verilmiştir. Veri setlerinin içeriği Tablo 6.1'de özetlenmiştir.

Tablo 6.1: Çalışmada kullanılmak üzere seçilmiş veri setlerine ait genel bilgiler

	Boyut	Görüntü/Metin Sayısı
Flickr1K	140MB	1000/(x5)
Flickr8K	1GB	8091/(x5)
Flickr30K	4GB	31783/(x5)

Flickr veri setlerinde yer alan her bir görüntüye ait beş adet tanımlayıcı metin yer almaktadır. Her bir metin, metnin başlangıcını ve bitişini belirten "startseq" ve "endseq" anahtar kelimeleri sayesinde birbirinden ayırt

edilmiştir. Örnek bir görüntü ve bu görüntüye ait metin bilgisi Şekil 6.1'de yer almaktadır.

```
# after preprocess of text
mapping['1000092795']
```

```
[16]: ['startseq two young guys with shaggy hair look at their hands while hanging out in the yard endseq',
       'startseq two young white males are outside near many bushes endseq',
       'startseq two men in green shirts are standing in yard endseq',
       'startseq man in blue shirt standing in garden endseq',
       'startseq two friends enjoy time spent together endseq']
```

Şekil 6.1: Veri setlerinde yer alan bir görüntüye ait metin bilgisi

Flickr veri setleri birbirlerinin alt setleri olarak aynı verileri içermektedir. En küçük küme olan Flickr1K içerisindeki veriler, Flickr8K ve Flickr30K'da da yer alan ilk 1000 görüntü ve 5000 metinden oluşmaktadır. Aynı şekilde Flickr8K, Flickr30K' nın bir alt kümesidir. Şekil 6.2 ve Şekil 6.3'te, Flickr1K ve Flickr8K'dan alınmış aynı isimdeki görüntü ve bu görüntüye ait metin bilgileri gösterilmiştir.

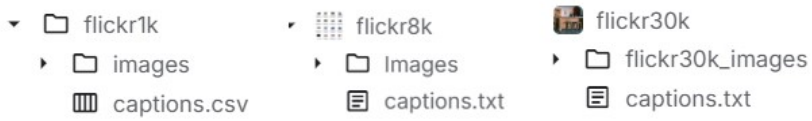
```
[42]: generate_caption("1002674143_1b742ab4b8.jpg")
```

```
---------------------Actual---------------------
startseq little girl covered in paint sits in front of painted rainbow with her hands in bowl endseq
startseq little girl is sitting in front of large painted rainbow endseq
startseq small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it endseq
startseq there is girl with pigtails sitting in front of rainbow painting endseq
startseq young girl with pigtails painting outside in the grass endseq
--------------------Predicted--------------------
startseq little girl in pigtails painting with painted painted endseq
```

Şekil 6.2: Flickr1K'dan alınmış örnek görüntü ve görüntüye ait metin bilgisi

Çalışma kapsamında gerçekleştirilen sistemlerde kullanılan ve Şekil 6.4'te de gösterilen bu veri setlerinin yapısı; görüntülerin tümünü içeren bir klasör ve metinleri içeren bir dosyadan (.csv, .txt) oluşmaktadır.

Şekil 6.3: Flickr8K'dan alınmış örnek görüntü ve görüntüye ait metin bilgisi



Şekil 6.4: Flickr Veri Setlerinin Dosya Yapısı

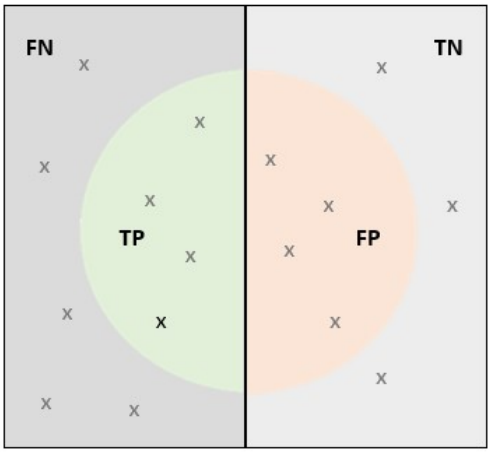
Flickr1K, Flickr8K ve Flickr30K veri setleri kendiliğinden eğitim, test ve validasyon için bölümlenmiştir. Her biri dosya bazında ayrılmış birbirinden bağımsız verileri içermektedir. Tablo 6.2'de her bir veri seti için eğitim, test ve validasyon için bölümlendikten sonra toplam görüntü (G) ve metin (M) bilgileri verilmiştir.

Tablo 6.2: Veri setlerinin eğitim, test ve validasyon bölümlenme bilgileri

	Eğitim	Test	Validasyon
Flickr1K	600 G/3000 M	300 G/1500 M	100 G/500 M
Flickr8K	4855 G/24275 M	2427 G/12135 M	809 G/4045 M
Flickr30K	19070 G/95350 M	9535 G/47675 M	3178 G/15890 M

6.2 Ölçümleme Metrikleri

Çalışılan alan görüntüler için altyazı üretimi olunca, bilinen ve sınıflandırmada başarı ölçüm metrikleri olarak kullanılan standart doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F-skor (f-score) gibi değerleri yalnız başına izlemek, sistemin performansını ölçümlemek için yeterli olamamaktadır.



Şekil 6.5: Karmaşıklık Matrisi

Şekil 6.5'te örnek bir veri kümesi verilmiştir. Bilinen en temel örneklerden biri ile açıklayacak olursak, verilen bir görüntünün kedi ya da köpek olduğunu tahminleyen bir sınıflandırma algoritmasını düşünelim. TP durumu, verilen görüntünün ve tahminlenen çıktının kedi olduğu durumdur. TN ise, verilen görüntü bir köpektir ve çıktı değeri köpek olarak tahminlenmiştir. FP durumunda ise verilen bir köpek görüntüsü, kedi olarak tahminlenmiştir. FN ise, verilen görüntü kedidir ve köpek olarak tahmin edilmiştir. Bu ayrıştırma sayesinde sınıflandırma algoritmasının ne kadar doğru çalıştığı hesaplanabilmektedir. Matematiksel olarak bunun hesaplayabilmek için Şekil 6.6'da gösterilen accuracy, precision, recall, f1-measure değerleri kullanılmaktadır.

Accuracy, modelin performansını ölçümlemek için kullanılmaktadır. Precision, yapılan doğru pozitif tahminlerin sayısını ölçen bir ölçüm iken Recall, yapılabilecek tüm olumlu tahminlerden yapılan doğru olumlu tahminlerin sayısı olmaktadır. Ancak ne precision ne de recall tam olarak doğru bilgi vermektedir. Çünkü bazı durumlarda precision değeri çok yüksek iken recall

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Şekil 6.6: Accuracy, Precision Recall, F-Measure Değer Hesaplamaları

değeri çok düşük ya da tam tersi olabilmektedir. Bu yüzden alternatif olarak her ikisini de dikkate alan bir ölçümleme metriği olarak F-measure değeri karşımıza çıkmaktadır.

BLEU, METEOR, ROUGE gibi ölçümleme metrikleri, bu hesaplamaları temel alarak oluşturulmuştur.

6.2.1 BLEU

BLEU metinsel ifadelerin başarılarının ölçümlemesinde kullanılan en popüler yöntemlerden biridir. Üretilen precision değerine odaklanarak çalışmaktadır. n-gram yöntemini kullanarak üretilen metindeki n'li setlere bakılarak başarı ölçümlenebilmektedir. (Papineni et al., 2002) tarafından gerçekleştirilen çalışmada yer alan bilgilere göre BLEU, tahminlenen ve referans alınan metne bakılarak tokenlar arasındaki örtüşmenin gücüne göre bir skor değeri üretir ancak anlam bakımından bir karşılaştırma yapmamaktadır. Daha kısa n' setleri daha yüksek başarı üretmektedir ve bunun önüne geçebilmek adına brevity penalty değeri ile ceza puanlaması yapılarak bir denge sağlanmaya çalışılmaktadır.

6.2.2 ROUGE

Metin başarı ölçümlemede kullanılan bir diğer yöntem ise recall değerine odaklanarak çalışan ROUGE metriğidir. BLEU gibi n'li setler baz alınarak da çalışabilmektedir ayrıca ROUGE metriğinde en uzun ortak alt küme grubu da dikkate alınabilmektedir. Her metrik gibi referans metin ve üretilen metni karşılaştırmak esastır. (Lin, 2004)'nın çalışmasında yer alan bilgiye göre, büyük küçük harfe duyarlıdır.

6.2.3 METEOR

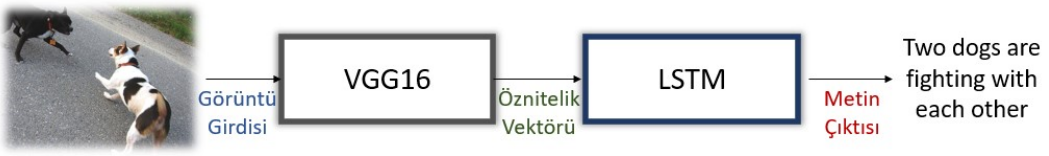
METEOR, unigram eşleştirme konseptine dayanan, makine çevirisi değerlendirmesi için otomatik bir ölçümdür. Unigram precision, unigram recall'u kullanarak, eşleşen kelimelerin ne kadar iyi sıralandığını bulabilmek için tasarlanmış bir yöntemdir. (Banerjee & Lavie, 2005) F-measure değerini kullanan ve dolayısıyla precision ve recall değerlerine dayanarak çalışan bu metrik, kelimelerin hizalanmasına bakılarak çalışmaktadır. Referans alınan ve üretilen metindeki kelimeler arasında bire bir eşleme yapılmaya çalışılmaktadır.

6.3 Önerilen Modeller

Her alanda çok sayıda görüntü, yediden yetmişe herkesin sahip olduğu telefonlar ve çok çeşitli diğer cihazlarla kolaylıkla elde edilebilmektedir. Sahip olunan bu verinin işlenmesi ve anlamlı bilgilerin çıkarılması önem arzetmektedir. Görüntü altyazı üreten sistemlerin yaygınlaşmasını sağlayan bu neden, bu alanda efektif sonuçlar alabilmek adına bir çok model mimarisiyle gerçekleştirilmeye çalışılmıştır. Bu bölümde, bu çalışma kapsamında gerçekleştirilmiş kodlayıcı kod çözücü modellerden bahsedilmiştir. Güncel teknolojilere ve üretilen performanslara dayanılarak farklı bileşimler oluşturulmuş ve en iyi model önerileri elde edilmeye çalışılmıştır.

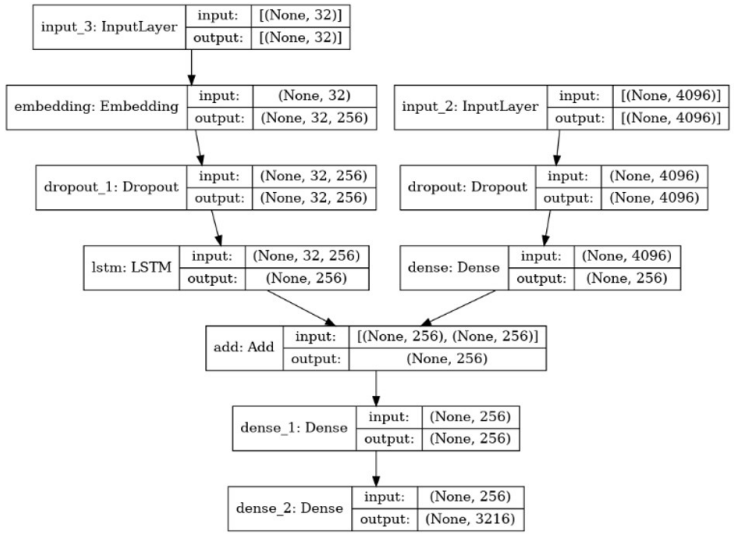
6.3.1 VGG16-LSTM

İlk olarak CNN'in en temel modellerinden biri olan VGG16 ve metinsel verilerle çalışmada akla ilk gelen yöntem olan LSTM ağları birleştirilerek bir model önerisi yapılmıştır.



Şekil 6.7: VGG16-LSTM Kodlayıcı Kod Çözücü Mimarisi

Şekil 6.7'de gösterilen modelde VGG16 kodlayıcı ve LSTM kod çözücü olarak kullanılmıştır. İlk olarak eğitim veri setindeki görüntülerden öznitelik çıkarım işlemi gerçekleştirilmiştir. features_flickr1k.pkl isimli bir dosyada tutulmuştur. Daha sonra eğitim veri setindeki altyazıları içeren dosya okunup metin temizleme işlemleri gerçekleştirilmiştir. Bu aşamada başlangıç bitiş anahtar kelimeleri tanınarak işleme dahil edilmemiş, noktalama işaretleri dikkate alınmayarak kullanılmıştır. Sonrasında sistem görüntü metin eşlemesini gerçekleştirmiş ve bir veri tablosunda bu bilgiyi saklamıştır.



Şekil 6.8: VGG16-LSTM Kodlayıcı Kod Çözücü Mimari Katmanları

Şekil 6.8'de kodlayıcı kod çözücü modeli oluşturan katmanlar gösterilmiştir. Sol taraf metinsel ifadelerin işlenmesini ele alan bölümdür. Metinlerin sistemde kullanılmasını sağlayan bir girdi katmanı, metinsel ifadelerin kodlanmasını sağlayan embedding katmanı, düzenlileştirme stratejisi ile aşırı uyumu önlemek adına, verilerin bir kısmının dikkate alınmamasını sağlayan dropout katmanı ve kod çözme işlemini gerçekleştirecek olan LSTM katmanı yer almaktadır. Bu katman sayesinde her adımda bilgilerden hangisinin silineceğine, mevcut duruma ne ölçüde etki edeceğine karar verilmektedir. Sağ tarafta ise görüntülerin sistemde temsil edilip kullanılmasını sağlayan bir girdi katmanı, aşırı uyumu önlemek amacıyla bir dropout katmanı ve tam bağlı bir katman yer almaktadır.

LSTM ve bu tam bağlı katman bir araya getirilerek birleştirilmekte ve kodlayıcı kod çözücü sistem mimarisi oluşturulmaktadır. Bu katmanı tam bağlı iki katman takip etmektedir.

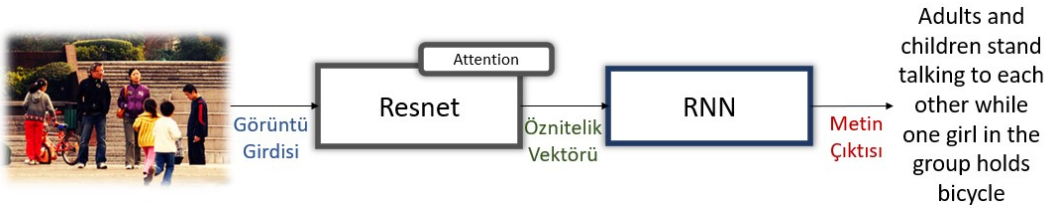
Önerilen bu model Flickr1K, Flickr8K ve Flickr30K veri setleriyle ayrı ayrı eğitilip test edilmiştir. Sonuçlar takip eden bölümde yer almaktadır. Modelin performansı, test veri seti kullanılarak gözlemlenmiş, tahminlenen değerleri için elde edilen başarılar ölçümlenmiştir. Performans ölçüm metrikleri olarak BLEU, ROUGE ve METEOR kullanılmıştır. Önerdiğimiz bu ve bundan sonraki modellerden elde edilen başarım değerleri Deneysel Çalışmalar bölümünde sunulmuştur.

6.3.2 ResNet-Attention-RNN

Transfer öğrenme yönteminde popüler olarak kullanılan önceden eğitilmiş ve başarısı kanıtlanmış başka bir CNN modeli ise ResNet'tir. Geri yayılımdan destek alınmaktadır. Çok fazla gizli katman oluştuğunda işleyişin zorlaştığı bilinen bir problemdir. Bunu çözmek adına residual bloklar bulunmaktadır. Böylece bazı bağlantılar atlanır ve katmanların kaybolması ya da aşırı katmana sahip olunması problemlerinin önüne geçilebilmektedir. Bu nedenle sistemde ResNet152 modeli kullanılarak bir kodlayıcı oluşturulmuştur. Imagenet ile önceden eğitilmiş bu model, görüntülerden özniteliklerin çıkarılmasında yardımcı olmuştur.

Şekil 6.9'da gösterilen modelin, kod çözücü bölümü temelde RNN mimarisini kullanmaktadır ancak dikkat modeli ve GRU kullanılarak genişletilmiştir. Kod çözücü bölümde, kapıları sayesinde RNN'i iyileştiren GRU katmanı, tam bağlı bir katman, dropout katmanı, bir başka tam bağlı bir katman yer almaktadır. Bu katmanlara ek olarak özniteliklerin seçilip, ağırlık değerlerinin belirlendiği dikkat modeli katmanları da yer almaktadır. Dikkat modeli Softmax katmanındaki ağırlık değerlerinden hesaplanan skor değerleri, çıkarılan özniteliklerin değerleriyle çarpılarak, dikkat modelinin ağırlık değerleri hesaplanmaktadır

ve bu değerler yeni ağırlık değerleri olarak kullanılmaktadır. Hesaplanan bu ağırlıklar GRU katmanına iletilmektedir ve çıktıya etki etmektedir.



Şekil 6.9: ResNet-Attention-RNN Kodlayıcı Kod Çözücü Mimarisi

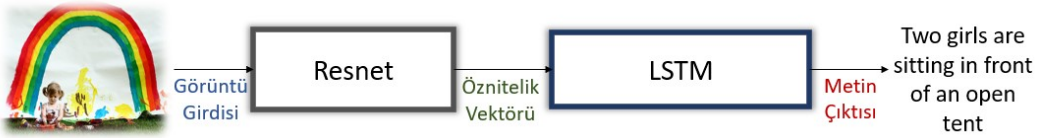
Geliştirilmiş bu mimari Flickr1K, Flickr8K ve Flickr30K kullanılarak denenmiş ve sonuçlar elde edilmiştir. Sistemde ilk olarak görüntü metin eşleşmelerini tutan dosya okunmuştur. Metinler üzerinde; nümerik değerler, anlamsız karakterler, noktalama işaretleri gibi gereksiz bilgilerden temizleme işlemleri gerçekleştirilir ve metinler kelime bazında sistem tarafından algılanır. Kodlayıcı mimari görüntülerden özniteliklerin çıkarımı işlemini gerçekleştirir ve bu değerler ihtiyaç duyulduğunda başvurmak üzere flickr_features isimli bir dosyada tutulur. Metinler için kodlama (embedding) işlemi gerçekleştirilir. Modelde içerik bağımsız kodlayıcılarda GloVe yöntemi kullanılmıştır. Bu kodlama yöntemiyle kelimelerin sadece yerel bağlamına değil, kelime vektörü elde edebilmek için global kelime birlikteliği de göz önünde bulundurulmaktadır.

Model üzerinde eğitim işlemi gerçekleştirilmiş ve her adımda kayıp değerleri hesaplanmıştır. Test veri setiyle test edilmiş ve modelin ürettiği başarı elde edilmiştir. Üretilen çıktılar BLEU, METEOR ve ROUGE metrikleriyle ölçümlenmiştir.

6.3.3 ResNet-LSTM

ResNet'in görüntülerden öznitelik çıkarımındaki başarısı göz önünde bulundurularak, çalışma kapsamında, kodlayıcı mimarisinde bir kez daha kullanılmıştır. Ancak bu kez, kod çözücü görevinde popüler olarak kullanılan

başarısı kanıtlanmış bir başka derin sinir ağı olan LSTM mimarisi gerçekleştirilmiştir. Şekil 6.10' da kodlayıcı kod çözücü mimari gösterilmiştir.

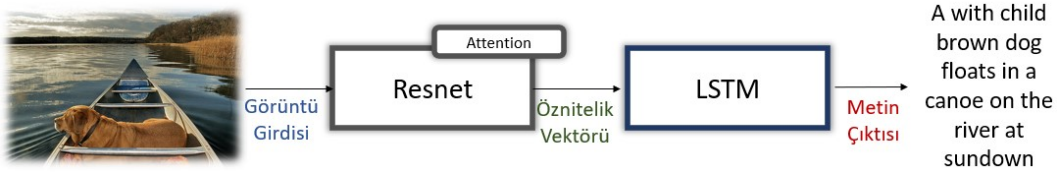


Şekil 6.10: ResNet-LSTM Kodlayıcı Kod Çözücü Mimarisi

Öncekilerde olduğu gibi öncelikle model üzerinde eğitim işlemi, görüntülerden özniteliklerin çıkarılması ve kelime kodlamaları yapılarak gerçekleştirilmiştir.

6.3.4 ResNet-Attention-LSTM

Tez çalışması kapsamında edinilen tecrübeler sonucu, önceki iki alt bölümde bahsedilen ResNet-Attention-RNN ve ResNet-LSTM modellerinin performansları gözlemlenerek dikkat mekanizmasının model performansına etkisini izleyebilmek amacıyla ResNet-Attention-LSTM modeli gerçekleştirilmiştir. ResNet-LSTM modelinden farklı olarak kodlayıcı bölüme dikkat mekanizması eklenmiştir. Şekil 6.11' de kodlayıcı kod çözücü mimari gösterilmiştir.



Şekil 6.11: ResNet-Attention-LSTM Kodlayıcı Kod Çözücü Mimarisi

Diğer modellerde olduğu gibi, model eğitim ve test aşamalarından geçerek; BLEU, METEOR ve ROUGE metrikleriyle ölçümlenen sonuçlar "Deneysel Çalışmalar" bölümünde verilmiştir.

6.3.5 EfficientNetB2-Attention-Transformer

Bu modelin gerçekleştirilme amacı, dönüştürücü modellerin görüntü altyazı üretimindeki performansının gözlemlenmesi ve LSTM, RNN gibi sinir ağlarıyla sonuçlarının karşılaştırılmasıdır. Bilindiği gibi dönüştürücü modeller, sıralı verilerdeki ilişkileri izleyerek, bağlamı ve dolayısıyla anlamı öğrenen bir sinir ağıdır. Ayrıca bu durum içeriğe bağlı metinsel kodlama yapabilmeyi de beraberinde getirmektedir. Şekil 6.12'de verilen modelde ayrıca farklı bir CNN modeli olan EfficientNetB2 kullanılmıştır. Bu model; Google AI tarafından geliştirilmiş, RNN'lerdeki bellek problemlerinin üstesinden gelmek amacıyla tasarlanmıştır.

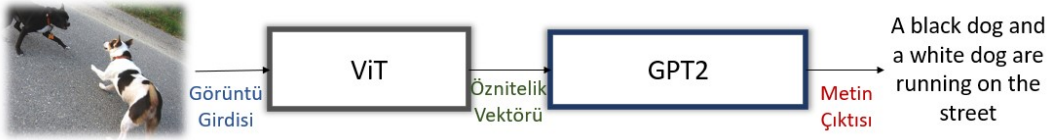
Şekil 6.12: EfficientNetB2 - Attention - Transformer Kodlayıcı Kod Çözücü Mimarisi

Sistemde ilk olarak altyazıları içeren dosya okunup, altyazı verilerindeki gürültü giderme işlemi gerçekleştirilmiştir. Sistemde görüntüler de okunup, görüntü metin eşlemesi yapılmış ve bir veri seti değişkeninde tutulmuştur. EfficientNetB2'nin standart görüntü boyutu ile çalışıyor olmasından dolayı görüntüler, 260 x 260 boyutunda olarak şekilde yeniden boyutlandırılarak bir ön işleme tabi tutulmuştur. Ayrıca gerçekleştirilen bu modelin kodlayıcı kısmında çok başlı dikkat (multi-head attention) yöntemi kullanılmıştır. Sonrasında kod çözücü model de bu yönteme uyarlanarak metinsel veriler için, metin içerisinde bulundukları pozisyonlar da dikkate alınarak, kelime vektörleri oluşturulmuştur. Çok başlı dikkat (multi head attention) yöntemi, paralel olarak çalışan birden fazla dikkat modeli kullanabilme olanağı tanımaktadır. Sistemde öz ve çapraz dikkat mekanizması böylelikle birlikte kullanılmıştır. Öz dikkat modeli ile üretilen çıktıdaki her bir kelime için, üretilen tüm metin içerisinde önemi dikkate alınırken, çapraz dikkat mekanizması ile

üretilen metin, kodlayıcıdan gelen verideki her bir bölümü dikkate alabilmeyi sağlamıştır. Modeldeki kodlayıcı kısım, kodlama, dropout, normalleştirme ve tam bağlı katmanlardan oluşmaktadır.

6.3.6 ViT-GPT2

Geliştirilen bu modelde amaç, dönüştürücü modellerin, görüntüler için altyazı üreten sistemlerde kodlayıcı olarak görev aldığında performansın ne ölçüde değişeceğini gözlemlemektir. Görü dönüşüm modeli, görüntüler için altyazı üretimi konusunda özelleşmiş, özetleme veya çeviri gibi diziden diziye görevlere yönelik eğitim modelleri için uyarlanmış bir dönüştürücü model mimarisine sahiptir. CNN'deki piksel işlemeden farkı olarak görüntüyü parçalara bölerek işlemektedir. Öz dikkat mekanizmasını içerisinde barındırdığından, bu parçalar arasındaki bağlantılar keşfedilmektedir. Böylelikle sadece yerel değil, görüntünün tümüne ait bir bilgi çıkarılmaktadır. Şekil 6.13'teki kodlayıcı kısım bu şekilde özetlenebilir.



Şekil 6.13: Vision Transformer - GPT2 Kodlayıcı Kod Çözücü Mimarisi

Modelin kod çözücü bölümünde, normal bir dönüştürücü mimari kullanmak yerine sıralı verilerle çalışma üzerinde özelleşmiş, önceden eğitilmiş üretici dönüştürücü model olan GPT teknolojisi kullanılmıştır. Bu bölümdeki girdi ve çıktı verileri, belirli bir uzunluktaki sürekli metinsel ifadelerdir. Dönüştürücü modellerin temelinde olan içerik bağımlı olarak kelime kodlama işlemi gerçekleştirilmektedir.

6.4 Önerilen Veri Artırma Modelleri

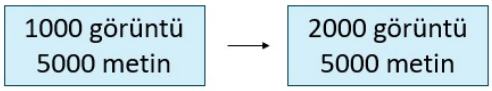
Çalışmada önerilen modellerde veri artırımının uygulandığı durumlarda, hem görüntülerin hem de metinlerin çoğaltılması göz önünde bulundurulmuştur. Yani çalışma kapsamında veri artırma işlemi, hem görüntlere hem de bu görüntülere ait metinlere uygulanmıştır. Bu işlemler ayrı ayrı ve birlikte alternatif yollarla denenmiştir. Sonraki bölümlerde veri artırma işlemlerimiz detaylandırılmıştır.

6.4.1 Görüntü Veri Artırımının Uygulanması

Bu yöntemle, görüntü altyazı üretiminde, görüntü veri artırımının sisteme etkisi gözlemlenmiştir. Flickr1K veri setinde bulunan her bir görüntü için artırma işlemi gerçekleştirilmiştir. Her görüntü, rastgele döndürme işlemine tabi tutularak artırılmıştır. Şekil 6.14'te bu yöntemle artırılmış bazı örnek görüntüler yer almaktadır.

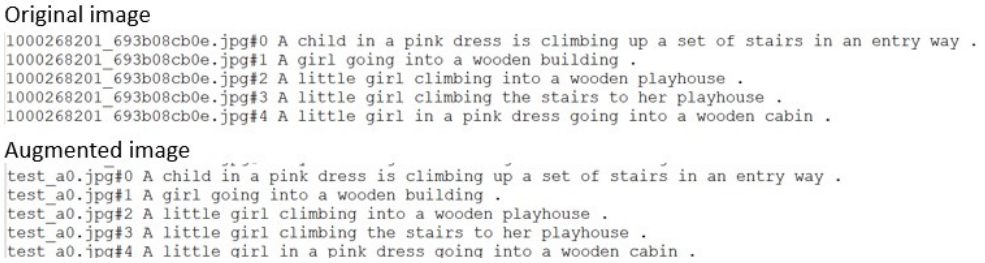
Şekil 6.14: Görüntülere Uygulanan Rastgele Döndürme Yöntemi

Şekil 6.15'te, Flickr1K veri setine artırma uygulanmadan önce bu veri setinde yer alan toplam görüntü ve her bir görüntüye ait metin sayısı soldaki kutuda yer almaktadır. Sağda ise görüntü veri artırımı gerçekleştirildikten sonra kullanılan toplam görüntü ve metin sayısı verilmiştir. Görülebileceği üzere, yeni veri setindeki metin sayısı sabit tutulurken, görüntü sayısı iki katına çıkarılmıştır.



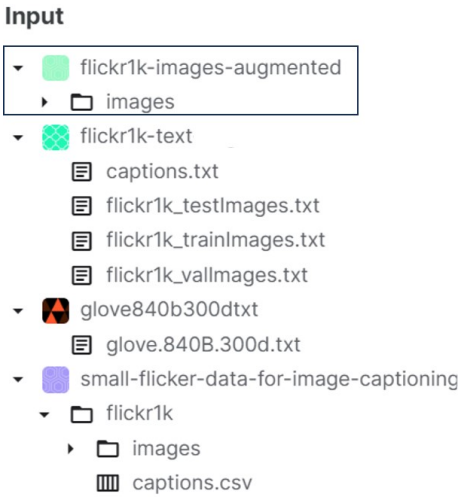
Şekil 6.15: Flickr1K Veri Setindeki Görüntüler İçin Veri Artırımı Gerçekleştirilmeden Önce Ve Gerçekleştirildikten Sonraki Veri Setlerinin Boyutları

Görüntüler artırıldıktan sonra, veri seti sadece görüntüler için genişletilerek, artırılmış görüntü için hali hazırda veri setinde bulunan uygun metinsel ifadelerle eşleştirilme işlemi gerçekleştirilmiştir. Şekil 6.16'da bir örnek yer almaktadır. Veri setindeki görüntüler ve ilgili altyazılar "GörüntüAdı#Altyazı" deseni ile tanımlanmıştır. Artırılmış görüntü, "test_a0.jpg" olarak isimlendirilmiştir ve orijinal görüntü ile aynı altyazılarla eşleştirilmiştir.



Şekil 6.16: Orijinal Ve Artırılmış Görüntünün Birlikte Kullanımı

Son olarak Şekil 6.17'de, modelin eğitimi ve testi için kullanılan veri seti gösterilmiştir. Model, orijinal veri seti (small-flickr-data-for-image-captioning) yerine artık artırılmış görüntüleri içeren (flickr1k-images-augmented) ve Şekil 6.17'da gösterilen şekilde artırılmış görüntülere ait altyazıları da içerisinde barındıran, güncellenmiş (flickr1k-text) klasörler ile çalışmaktadır.



Şekil 6.17: Flickr1K Veri Setindeki Görüntüler İçin Veri Artırımı Gerçekleştirildikten Sonra Model Üzerinde Kullanılan Veri Seti

Veri artırma işlemi, veri setindeki tüm görüntüler üzerinde gerçekleştirildikten sonra veri seti, eğitim ve test aşamalarında kullanılmak üzere, orijinal veri setindeki gibi bölümlenmiştir ve Tablo 6.3'te gösterilmektedir. Artırılmış veriler sadece eğitim aşamasında kullanılacak şekilde güncelleme işlemi gerçekleştirilmiştir. Model performansını veri artırımı bazında karşılaştırabilmek adına test verileri sabit tutulmuştur, orijinal halinden herhangi bir değişikliğe uğratılmamıştır. Test veri seti artırılmış verileri içermemektedir.

Tablo 6.3: Görüntü Artırımından Sonra Veri Setinin Eğitim ve Test Aşamaları İçin Bölümlenmesi

	Veri Art. Önce	Veri Art. Sonra
Toplam	1000	2000
Eğitim	600	1600 (600 orj. + 1000 aug.)
Test	300	300 (orj.)

Modelin, artırılmış veri seti ile ürettiği sonuçlar "Deneysel Çalışmalar" bölümünde yer almaktadır.

6.4.2 Metinsel Veri Artırımının Uygulanması

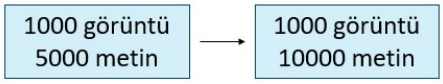
Çalışma başka bir açıdan ele alınarak, sadece metinlere artırma uygulandığında, görüntü altyazı üretim modellerine nasıl bir etkide bulunduğu gözlemlenmek istenmiştir. Flickr1K veri seti kullanılarak, içerdiği her bir metin için eş anlamlısıyla yer değiştirme yöntemi kullanılarak artırma işlemi gerçekleştirilmiştir. Şekil 6.18'de örnek bir metin ve artırılmış örneği verilmiştir.

7]:

	index	filename	caption
0	0	1000268201_693b08cb0e.jpg	a child in a pink dress is climbing up a set o...
1	1	1000268201_693b08cb0e.jpg	a girl going into a wooden building .
2	2	1000268201_693b08cb0e.jpg	a little girl climbing into a wooden playhouse .
3	3	1000268201_693b08cb0e.jpg	a little girl climbing the stairs to her playh...
4	4	1000268201_693b08cb0e.jpg	a little girl in a pink dress going into a woo...
5	5	1000268201_693b08cb0e.jpg	a child in a pink dress comprise climb upwardl...
6	6	1000268201_693b08cb0e.jpg	a fille going into a wooden construction.
7	7	1000268201_693b08cb0e.jpg	a picayune girl climb into a wooden playhouse.
8	8	1000268201_693b08cb0e.jpg	a lilliputian girl rise the stair to her playh...
9	9	1000268201_693b08cb0e.jpg	a petty girl in a garden pink dress going into...
10	0	1001773457_577c3a7d70.jpg	a black dog and a spotted dog are fighting

Şekil 6.18: Metinlere Uygulanan Eş Anlamlısıyla Yer Değiştirme İşlemi

Bu yöntemde bu kez görüntü sayısı sabit tutulup, sadece altyazılar üzerinde artırma işlemi gerçekleştirilmiştir. Şekil 6.19'da sol kutucukta veri artırımından önceki, sağ kutucukta ise sonraki toplam veri boyutu özetlenmiştir.



Şekil 6.19: Flickr1K Veri Setindeki Metinler İçin Veri Artırımı Gerçekleştirilmeden Önce Ve Gerçekleştirildikten Sonraki Veri Setlerinin Boyutları

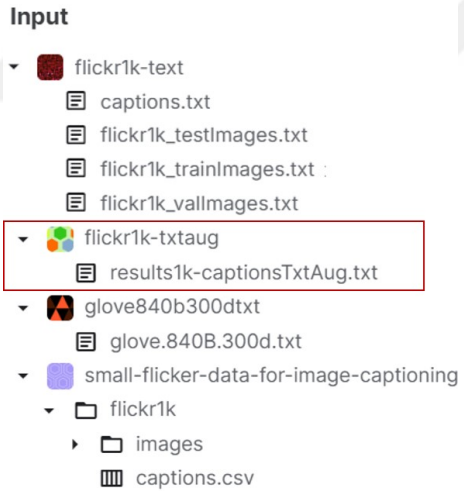
Metinler için veri artırımı gerçekleştirildikten sonra referans alınan veri setine entegrasyonu gerçekleştirilmiştir. Şekil 6.20'de, bir görüntü için verilen

orijinal ve çerçeve ile işaretli artırılmış metin eşleşmesi verilmiştir. Görülebileceği üzere, üretilen yeni metinler, aynı görüntü ile eşleştirilerek, sadece metinsel veri artırımı gerçekleştirilmiştir.

```
1000268201_693b08cb0e.jpg#0 A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1 A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2 A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3 A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4 A little girl in a pink dress going into a wooden cabin .
1000268201_693b08cb0e.jpg#5 A child in a pink dress comprise climb upwardly a readiness of stairs in an entry wayz.
1000268201_693b08cb0e.jpg#6 A fille going into a wooden construction.
1000268201_693b08cb0e.jpg#7 A picayune girl climb into a wooden playhouse.
1000268201_693b08cb0e.jpg#8 A lilliputian girl rise the stair to her playhouse.
1000268201_693b08cb0e.jpg#9 A petty girl in a garden pink dress going into a wooden cabin.
```

Şekil 6.20: Orijinal Ve Artırılmış Metinlerin Birlikte Kullanımı

Şekil 6.21'de, sadece metinsel veri artırımını içeren ve model için kullanılan veri seti gösterilmiştir. Metin artırma işlemi gerçekleştirildikten sonra Şekil 6.20'de gösterildiği gibi görüntü yeni metin eşlemeleri yapılıp yeni bir klasörde (flickr1k-txtaug) tutulmuştur. Görüntüler, orijinal veri seti klasöründen kullanılmaktadır.



Şekil 6.21: Flickr1K Veri Setindeki Metinler İçin Veri Artırımı Gerçekleştirildikten Sonra Model Üzerinde Kullanılan Veri Seti

Artırılmış metinleri içeren yeni veri seti, eğitim ve test aşamlarında kullanılmak üzere bölümlenmiştir. Tablo 6.4'te bölümlenmiş veri setinin boyutları gösterilmiştir. Artırılmış metinsel veriler sadece eğitim aşamasında

kullanılmış, test verisi değiştirilmeden korunarak model performansı elde edilerek karşılaştırılmıştır.

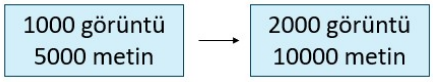
Tablo 6.4: Metin Artırımından Sonra Veri Setinin Eğitim ve Test Aşamaları İçin Bölümlenmesi

	Veri Art. Önce Veri Boyutu	Veri Art. Sonra Veri Boyutu
Toplam	5000	10000
Eğitim	3000	8000 (3000 orj. + 5000 aug.)
Test	1500	1500 (orj.)

6.4.3 Görüntü ve Metinsel Veri Artırımının Birlikte Uygulanması

Bu model önerisinde, önceki iki bölümde açıklanımış görüntü ve metinsel veriler için artırma yöntemleri birleştirilmiş ve model üzerinde birlikte kullanılmıştır. Görüntü altyazı üreten sistemlerde sadece görüntülerin değil görüntülerle birlikte metinlerin de artırıldığı durumda modelin performansı gözlemlenmek istenmiştir.

Veri setindeki her iki veri tipi için de veri artırımı uygulanmadan önce ve sonraki veri setinin boyutları Şekil 6.22' de gösterilmiştir.



Şekil 6.22: Flickr1K Veri Setindeki Görüntüler Ve Metinler İçin Veri Artırımı Gerçekleştirilmeden Önce Ve Gerçekleştirildikten Sonraki Veri Setlerinin Boyutları

Orijinal veri seti artırılmış her iki veri tipi için de genişletilmiştir. Şekil 6.23' te gösterilen birinci kutucukta orijinal veri setindeki veri aynen korunarak kullanılmıştır. İkinci kutucukta ise artırılmış görüntü, artırılmış metin verisiyle eşleştirilerek veri setine eklenmiştir. Örnekte gösterilen, "test_a0.jpg" olarak isimlendirilmiş görüntü artırılmış bir görüntüdür ve artırılmış metinlerle uygun tanımlama deseni ("GörüntüAdı#Altyazı") kullanılarak eşleştirilmiştir.

```
1000268201_693b08cb0e.jpg#0 A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1 A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2 A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3 A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4 A little girl in a pink dress going into a wooden cabin .
```

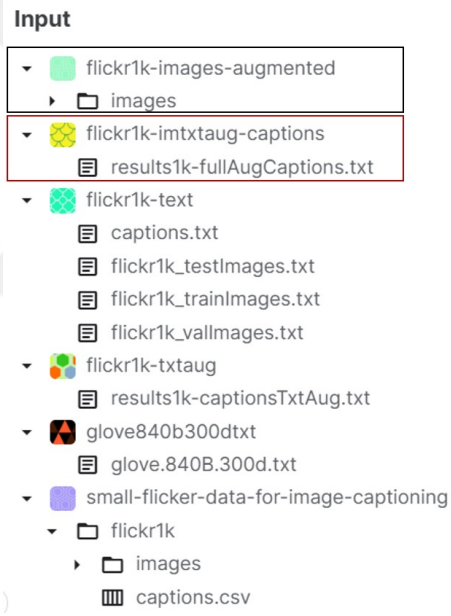
1

```
test_a0.jpg#0 A child in a pink dress comprise climb upwardly a readiness of stairs in an entry wayz.
test_a0.jpg#1 A fille going into a wooden construction.
test_a0.jpg#2 A picayune girl climb into a wooden playhouse.
test_a0.jpg#3 A lilliputian girl rise the stair to her playhouse.
test_a0.jpg#4 A petty girl in a garden pink dress going into a wooden cabin.
```

2

Şekil 6.23: Orijinal Ve Artırılmış Metinlerin Birlikte Kullanımı

Şekil 6.24'te model üzerinde kullanılan, oluşturulmuş yeni veri seti gösterilmektedir. Yeni sistemde, artırılmış görüntüleri de barındıran "flickr1k-images-augmented" klasörüne başvurumaktadır. Ayrıca Şekil 6.23'te bahsedilen eşleştirmeyi barındıran yeni bir dosya oluşturulmuştur ve "flickr1k-imtxtaug-captions" klasöründe yer almaktadır.



Şekil 6.24: Flickr1K Veri Setindeki Görüntüler Ve Metinler İçin Veri Artırımı Gerçekleştirildikten Sonra Model Üzerinde Kullanılan Veri Seti

Artırılmış görüntü ve metinleri içeren genişletilmiş veri seti, eğitim ve test için bölümlenmiştir. Artırılmış veriler sadece eğitim aşamasında kullanılmış, test verisi ise değiştirilmeden kullanılmıştır.

Tablo 6.5: Görüntü ve Metin Artırımından Sonra Veri Setinin Eğitim ve Test Aşamaları İçin Bölümlenmesi

	Veri Art. Önce Gör./Met.	Veri Art. Sonra Gör./Met.
Toplam	1000/5000	2000/10000
Eğitim	600/3000	1600/8000 (600/3000 orj. + 1000/5000 aug.)
Test	300/1500	300/1500 (orj.)

Bu veri seti ile elde edilen sonuçlar "Deneysel Çalışmalar" bölümünde gösterilmiş ve diğer yöntemlerle karşılaştırılmıştır.

6.4.4 Difüzyon Model ile Görüntü Veri Artırımının Uygulanması

Çalışmada farklı veri artırma yöntemlerinin görüntü altyazı üretimi modelinin performansına etkisi gözlemlenmek istenmiştir. Bu sebeple bilinen difüzyon model mimarisi kullanılarak görüntü verileri için artırma işlemi gerçekleştirilmiştir.

Önerilen bu modelde girdiler altyazılardır. Altyazılar üzerinden görüntü üretme işlemi gerçekleştirilmiştir. Şekil 6.25'te Flickr1K veri setinden alınmış örnek bir görüntü ve bu görüntüye ait altyazı yer almaktadır.

Şekil 6.25: Flickr1K Veri Setinden Örnek Bir Görüntü ve İlgili Altyazı

Model bu altyazı bilgisi ile beslenerek benzer bir görüntünün üretilmesi işlemi gerçekleştirilmiştir. Şekil 6.26'da bahsedilen örnek metin için üretilen görüntü gösterilmektedir.

Şekil 6.26: Flickr1K Veri Setindeki Örnek Bir Altyazıdan Üretilmiş Görüntü

Veri setindeki her bir görüntüye ait altyazılar kullanılarak ilgili yeni görüntülerin üretimi işlemi gerçekleştirilmiş ve bu metinlerle eşleştirilmiştir. Şekil 6.27'de genişletilmiş veri setinden, orijinal ve artırılmış görüntüler için birer metin ile eşleştirme görüntüsü yer almaktadır.

```
1000268201_693b08cb0e.jpg#0 A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1 A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2 A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3 A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4 A little girl in a pink dress going into a wooden cabin .
```

```
test_a0.jpg#0 A child in a pink dress is climbing up a set of stairs in an entry way .
test_a0.jpg#1 A girl going into a wooden building .
test_a0.jpg#2 A little girl climbing into a wooden playhouse .
test_a0.jpg#3 A little girl climbing the stairs to her playhouse .
test_a0.jpg#4 A little girl in a pink dress going into a wooden cabin .
```

Şekil 6.27: Flickr1K Veri Setindeki Bir Görüntüden Üretilen Bir Görüntünün Altyazılarının Orjinalinden Alınması

Bu yöntem de sadece görüntü veri artırımı yöntemindeki veri boyutunun artırıldığı kadar artırılmıştır. Tablo 6.6'da genişletilmiş veri setinin boyutu verilmiştir.

6.4.5 Difüzyon Model ile Görüntü ve Metinsel Veri Artırımının Birlikte Uygulanması

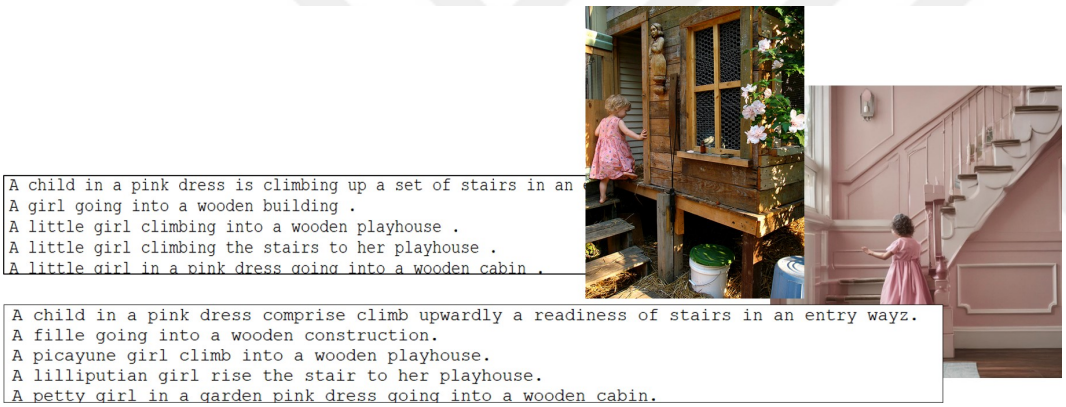
Bu model önerisinde, kararlı difüzyon ile artırılan görüntüler, önceki bölümlerde bahsedilen artırılmış metinlerle eşleştirilmiştir. Böylece, görüntü altyazı üretim

Tablo 6.6: Görüntü Artırımından Sonra Veri Setinin Eğitim ve Test Aşamaları İçin Bölümlenmesi

	Veri Art. Önce	Veri Art. Sonra
Toplam	1000	2000
Eğitim	600	1600 (600 orj. + 1000 aug.)
Test	300	300 (orj.)

modelinin performansının, kararlı difüzyon modeli ile görüntü artırımının etkisinin, altyazıların da arttığı durumda ne şekilde değiştiği gözlemlenmiştir.

Önceki bölümlerden yola çıkılarak tahmin edilebileceği üzere, üretilen yeni görüntülerle, yeni metin verileri eşleştirilerek orijinal veri seti genişletilmiştir. Şekil 6.28'de bir örnek gösterilmiştir.



Şekil 6.28: Orijinal Ve Artırılmış Veri Setlerinden Birer Görüntü ve İlgili Altyazıları

Bu yöntem de görüntü ve metin veri artırımı yöntemindeki veri boyutunun artırıldığı kadar artırılmıştır. Tablo 6.7'de genişletilmiş veri setinin boyutu verilmiştir.

Tablo 6.7: Görüntü ve Metin Artırımından Sonra Veri Setinin Eğitim ve Test Aşamaları İçin Bölümlenmesi

	Veri Art. Önce Gör./Met.	Veri Art. Sonra Gör./Met.
Toplam	1000/5000	2000/10000
Eğitim	600/3000	1600/8000 (600/3000 orj. + 1000/5000 aug.)
Test	300/1500	300/1500 (orj.)

Bu veri seti ile elde edilen sonuçlar "Deneysel Çalışmalar" bölümünde gösterilmiş ve diğer yöntemlerle karşılaştırılmıştır.

7 DENEYSEL ÇALIŞMALAR

Bu bölümde, tez çalışması kapsamında önerilen modellerle elde edilen sonuçlar gösterilmektedir. İzleyen alt bölümlerde, kullanılan gerçekleştirme ortamı ve kütüphaneler, yapılan deneylerle ilgili teknik detaylar, önerilen görüntü altyazı üretimi kodlayıcı - kod çözücü modeller için elde edilen deneysel sonuçlar ve önerilen veri artırma yöntemlerinin, kodlayıcı - kod çözücü model üzerinde etkisini ölçümleyen deneyler ve sonuçları yer almaktadır.

7.1 Ortam ve Kütüphaneler

İşlemler derin öğrenme ve yapay zekâ denince akla ilk gelen python programlama dili ile gerçekleştirilmiştir. Makine öğrenmesi kapsamında derin öğrenme yöntemlerini kullanabilmek için Tensorflow'un Keras API'sı; görüntü, metin verileri ile çalışmada, derin öğrenmeden aktarmalı öğrenme modellerine, bir ağda bulunabilecek katmanlara sahip olup tanımlayabilmek için kullanılmıştır. Bunun dışında numpy, tüm matematiksel işlemlerin gerçekleştirilmesinde yardımcı olmuştur. Sklearn, eğitim süreci işlemlerinde; NLTK, doğal dil işleme operasyonlarında; Pandas, veriler ile ilgili süreçlerde; torch, derin öğrenmede kullanılan sinir ağlarının tanımlanması, dönüştürücü modellerin içeri aktarılmasında; transformers, dönüştürücü modellerin ve ilgili eğitim yöntemlerinin içeri aktarılmasında kullanılmıştır.

Deneylerin tümü 20 epok ve batch boyutu 64 olarak gerçekleştirilmiş ve üçer kez tekrarlanmıştır. Adam optimize etme (optimizer) yöntemi kullanılarak ağ üzerindeki ağırlıkların (w) iyileştirilmesi için, öğrenme oranı gerçek zamanlı olarak her adımda güncellenmiştir.

Tüm gerçekleştirimler Kaggle platformunda yapılmıştır. Öncesinde Google Colab denenmiş ancak RAM yetersizliği dolayısıyla gerçekleştirmeler Kaggle platformuna taşınmıştır. Bu platform GPU desteği sunmaktadır ve bu sayede işlem süreleri kısaltılmıştır. Örneğin Tablo 7.1'de, öznitelik çıkarım süreleri

gösterilmiştir. Eğitim sürecinde, GPU kullanılmadığı durumda kodun çalışması mümkün olamamaktadır.

Tablo 7.1: Flickr Veri Setlerinin GPU Olmadan ve GPU İle Öznitelik Çıkarım Süreleri

	GPU Kullanılmadan	GPU Kullanarak
Flickr1K	0:17:17s	0:00:23s
Flickr8K	2:19:48s	0:01:53s
Flickr30K	6:30:48s	0:04:49s

Modellerin çalışma süreleri veri setleri bazında yakınlık göstermektedir. GPU kullanılarak; Flickr1K veri seti ile önerilen modellerin ortalama çalışma süresi yaklaşık 4 saat, Flickr8K ile yaklaşık 7 saat, Flickr30K ile ise yaklaşık 11 saat olmaktadır.

7.2 Önerilen Modeller için Performans Karşılaştırmaları

Çalışmada, görüntüler için altyazı üreten bir sistem için beş farklı kodlayıcı - kod çözücü model önerisinde bulunulmuş, önceki bölümlerde detaylandırılmıştır. Bu bölümde her bir model için elde edilen sonuçlar gösterilmekte ve birbirleriyle karşılaştırılmaktadır. Geliştirilen her bir kodlayıcı - kod çözücü model BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR ve ROUGE metrikleri kullanılarak skor değerleri hesaplanmıştır.

İlk olarak VGG16 ve LSTM mimarileri kullanılarak geliştirilen görüntü altyazı üretim modeli için gözlemlenen sonuçlar Tablo 7.2'de verilmiştir. Önerilen model; Flickr1K, Flickr8K ve Flickr30K veri setleri ile denenmiş, her biri için elde edilen skor değerleri gösterilmiştir.

Tablo 7.2: VGG16-LSTM Modelinin Flickr Veri Setleri Üzerindeki Başarım Değerleri

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1(r)
Flickr1K	50.5	30.0	18.7	10.3	37.1	67.7
Flickr8K	53.2	30.8	18.9	10.7	37.0	69.8
Flickr30K	53.9	29.1	17.1	9.3	33.3	67.4

Önerilmiş bir başka kodlayıcı - kod çözücü model, ResNet-Attention-RNN mimarileri kullanılarak geliştirilen görüntü altyazı üretim modelidir. Tablo 7.3'te bu model üzerinde; Flickr1K, Flickr8K ve Flickr30K veri setleri ile elde edilen skor değerleri gösterilmiştir.

Tablo 7.3: Resnet-Attention-RNN Modelinin Flickr Veri Setleri Üzerindeki Başarım Değerleri

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1(r)
Flickr1K	89.2	87.5	84.1	81.6	85.6	69.9
Flickr8K	67.7	59.6	52.6	45.0	63.3	65.8
Flickr30K	22.5	14.6	10.0	6.8	27.2	55.3

Bu tez çalışması kapsamında, başka güncel teknolojileri kullanarak görüntü altyazı üretim sistemlerine katkısı gözlemlenmek için önerilen diğer dört kodlayıcı - kod çözücü model ise EfficientNetB2-Att-Transformer, ViT-GPT2, ResNet-LSTM ve ResNet-Attention-LSTM modelleridir. Literatürdeki çalışmalarla uyumlu olabilmek için seçilen Flickr8K veri seti ile tüm modeller eğitilip, test edilmiştir. Bahsedilen bu diğer üç modelin, ResNet-Attention-RNN ve VGG16-LSTM modellerinin performanslarına bakılarak gerisinde kaldıkları görülmüştür. Tablo 7.4'te tüm modellerin Flickr8K ve ilgili metrikler ile ürettikleri skor değerleri karşılaştırılmıştır.

Tablo 7.4: Çalışmada Önerilen Modellerin Flickr8K Veri Seti Üzerinde Karşılaştırılması

8K	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1(r)
VGG16-LSTM(Tez)	53.2	30.8	18.9	10.7	37.0	**69.8**
ResNet-Attention-RNN(Tez)	**67.8**	**59.1**	**52.6**	**45.2**	**63.0**	65.1
ResNet-LSTM(Tez)	49.6	39.7	24.9	6.0	-	64.4
ResNet-Attention-LSTM(Tez)	65.2	43.7	27.4	16.9	-	-
EfficientNetB2-Att-Transformer(Tez)	60.2	42.5	29.8	20.1	35.3	50.5
ViT-GPT2(Tez)	-	-	-	-	22.6	64.2

Flickr8K veri seti kullanılarak tüm modeller üzerinde elde edilen tüm skor değerleri, literatürde hali hazırda bulunan görüntü altyazı üretimi için geliştirilmiş, Flickr8K veri seti ile çalışılmış olan modeller ile karşılaştırılmıştır. Tablo 7.5'te geleneksel makine öğrenmesi yöntemlerini, popüler RNN, LSTM ağlarını ya da dönüştürücü modelleri kullanarak geliştirilen modeller için geniş çaplı bir araştırma yapılmış ve bu tez çalışması kapsamında geliştirilen model mimarilerine benzer çalışmalar seçilmiştir. Tablodaki sonuçlardan yola çıkılarak, literatürdeki çalışmalara kıyasla ResNet-Attention-RNN modelinin performansının METEOR ve ROUGE metriklerine göre gözle görülür, BLEU metriğine göre ise az bir farkla da olsa daha iyi olduğu görülmüştür.

Tablo 7.5: Çalışmada Önerilen Modellerin Literatürdeki Çalışmalarla Flickr8K Veri Seti Üzerinde Karşılaştırılması

8K	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1(r)
mRNN (Karpathy et al., 2014)	57.9	38.3	24.5	16.0	16.7	-
LogBilinear (Kiros et al., 2014)	65.6	42.4	27.7	17.7	17.3	-
DBDAF (Wang & Gu, 2022)	67.5	47.2	33.2	22.4	21.1	42.4
RC (Ji et al., 2022)	67.7	47.4	33.2	22.7	20.9	42.4
Hard-Attention (Xu et al., 2015)	67.0	45.7	31.4	21.3	20.3	-
gLSTM (Xu et al., 2015)	63.5	44.8	30.7	20.6	20.3	-
phiLSTM (Tan & Chan, 2016)	63.6	43.6	27.6	16.6	-	-
SS-Ensemble (Jin et al., 2015)	63.9	45.9	31.9	21.7	-	-
Visual enhanced gLSTM (Zhang et al., 2021)	65.0	46.3	31.7	21.5	20.5	51.8
ViT-GPT2 (Kwok, 2023)	-	-	-	-	-	59.0
ViT-BERT (Kwok, 2023)	-	-	-	-	-	42.0
VGG16-LSTM(Tez)	53.2	30.8	18.9	10.7	37.0	**69.8**
ResNet-Attention-RNN(Tez)	**67.8**	**59.1**	**52.6**	**45.2**	**63.0**	65.1
ResNet-LSTM(Tez)	49.6	39.7	24.9	6.0	-	64.4
ResNet-Attention-LSTM(Tez)	65.2	43.7	27.4	16.9	-	-
EfficientNetB2-Att-Transformer(Tez)	60.2	42.5	29.8	20.1	35.3	50.5
ViT-GPT2(Tez)	-	-	-	-	22.6	64.2

7.3 Önerilen Veri Artırım İşlemlerinin Model Performansına Etkisi

Tez çalışması kapsamında, küçük veri setlerinde görüntü altyazı üretimi için bir çözüm olması amacıyla veri setindeki eğitim verileri çoğaltılarak veri seti genişletilmiştir. Veri artırımı için görüntü çoğaltmada; rastgele görüntü döndürme, difüzyon model ve metin çoğaltmada metindeki eş anlamlı/benzer kelimelerle yer değiştirme yöntemleri uygulanmıştır. Bu yöntemler farklı kombinasyonlarla bir araya getirilip modelin performansına etkileri

gözlemlenmiştir. Çalışma kapsamında en yüksek performans elde eden ResNet-Attention-RNN modeli seçilmiş ve bu veri artırım metotları ile birleştirilerek yeni model önerileri yapılmıştır.

"Önerilen Veri Artırma Modelleri" bölümünde detaylandırıldığı gibi çalışma kapsamında; sadece görüntü artırımının, sadece altyazı artırımının ya da her ikisinin birlikte artırımının kodlayıcı - kod çözücü modele etkisi gözlemlenmiştir. Tablo 7.6'da Flickr1K için sırasıyla, herhangi bir veri artırımı olmadan, metinlere eş anlamlısıyla yer değiştirme yöntemi uygulanarak sadece metinsel veri artırımı, görüntülere rastgele döndürme yöntemi uygulanarak sadece görüntü veri artırımı, bahsedilen bu iki yöntemin birlikte uygulanarak hem görüntü hem de metinsel veri artırımı, metinlerden görüntü üretimini sağlayan kararlı difüzyon modeli ile sadece görüntü veri artırımı, bu yöntem ve bahsedilen metinsel veri artırımı birleştirilerek hem görüntü hem de metinsel veri artırımı ile elde edilen sonuçlar gösterilmektedir.

Tablo 7.6: Çalışmada Önerilen Flickr1K Üzerinde Veri Artırımının ResNet-Attention-RNN Modeli Sonuçlarına Etkisi

1K	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1(r)
ResNet - Attention - RNN	89.2	87.5	84.1	81.6	85.6	69.9
ResNet - Attention - RNN Txt Aug	**90.6**	**89.3**	**87.2**	**84.2**	**89.7**	67.6
ResNet - Attention - RNN Img Aug	87.1	85.4	83.0	79.9	88.6	**71.5**
ResNet - Attention - RNN Img(DifMod.) Aug	85.5	83.4	80.7	77.2	85.6	70.0
ResNet - Attention - RNN Img Txt Aug	88.9	86.9	84.1	80.8	86.7	70.1
ResNet - Attention - RNN Img(DifMod.) Txt Aug	86.6	84.2	81.4	77.9	86.9	71.0

Elde edilen sonuçlar, geliştirilen modelin, sadece metinsel verilere veri artırımı uygulanan veri setiyle eğitilip test edildiği takdirde normal veri setine göre BLEU metriğiyle %1.4 ve METEOR metriğiyle %4.1 ile daha iyi sonuçlar ürettiğini göstermektedir. Veri artırımı sadece görüntülere uygulandığında ise

sistemin başarısı ROUGE metriğine göre sistem performansını %1.6 artırarak en iyi değerine ulaşmaktadır.

Tablo 7.6' dan yola çıkılarak, ResNet-Attention-RNN modeli üzerinde uygulanan, veri artırma yöntemlerinden elde edilen sonuçlar karşılaştırıldığında, BLEU ve METEOR metriklerine göre en iyi sonucun sadece metinlere uygulanan veri artırımı ile elde edildiği görülmüştür. Bu bilgi ışığında, veri setinde bir görüntüye ait açıklayıcı daha çok metne sahip olunması, tanımlayıcı kelime sayısını artırıp çeşitlendirdiğinden, sistem performansını pozitif yönde etkilediği sonucuna da ulaşılmıştır. Gelecekte geliştirilebilecek derin öğrenme mimarileri kombinasyonları ile de daha başarılı görüntü altyazılama gerçekleştirimi için örnekler içermesi yönü ile önemlidir.

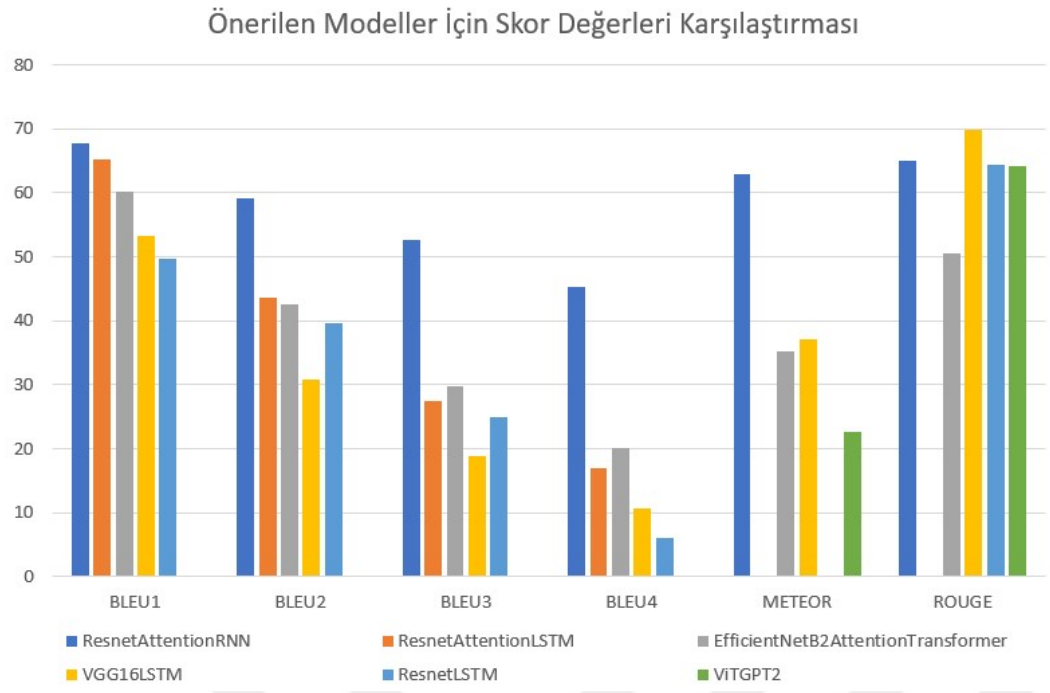
8 SONUÇ

Tezin önceki bölümlerinde değinilen çalışmaların tümü, görüntü altyazı sistemlerinde ilerlemeler sağlanması için modellerin geliştirilmesinin yolunu açmaktadır.

Tablo 7.4' te görüldüğü üzere, bu tez çalışması kapsamında önerilen kodlayıcı - kod çözücü modellerin, Flickr8K veri seti üzerinde, ROUGE skor metriği ile ürettikleri sonuçlar bazında performansı en yüksek modelin VGG16-LSTM modeli olduğu, bu modeli sırasıyla ResNet-Attention-RNN, ResNet-LSTM, VisionTransformer-GPT2 ve EfficientNetB2-Attention-Transformer kodlayıcı kod çözücü modellerinin izlediği görülmüştür. Ancak BLEU ve METEOR skor metrikleri göz önünde bulundurulduğunda en yüksek skor değeri ResNet-Attention-RNN modeli ile elde edilmiştir. Sonrasında bu modeli ResNet-Attention-LSTM, EfficientNetB2-Attention-Transformer, VGG16-LSTM modelleri takip etmektedir. BLEU metriğine göre ResNet-LSTM, METEOR metriğine göre ise ViT-GPT2 modeli çalışma kapsamında elde edilen skor değerleri bakımından en sonda yer almaktadır.

Önerilen modeller arasında performansı en yüksek olan ResNet-Attention-RNN modeli üzerinde, küçük veri seti olarak kabul edilen Flickr1K veri setine, veri çoğaltma işlemleri, görüntü ve metinsel veriler için ayrı ayrı ve birlikte uygulanarak aynı model üzerinde denenmiştir. Şekil 8.1' de önerilen tüm modellerin performansları grafik üzerinde karşılaştırılarak gösterilmiştir.

Gözlemlenen bir diğer sonuç, önerilen veri artırımı yöntemlerinin sistemin performansını olumlu yönde etkilediği olmuştur. Sistemin başarısının, görüntü ve metinsel veri çoğaltma işlemleri tek başlarına ya da birlikte uygulanmış veri setleriyle elde edilen sonuçların hem Flickr8K hem de Flickr30K veri setlerinin kullanıldığı durumlara göre tüm ölçümleme yöntemleri ile daha yüksek olduğu tespit edilmiştir. BLEU metriğine göre sistemde, sadece metinsel veri artırımı dışında gerçekleştirilen veri artımı yöntemlerinin, sistem

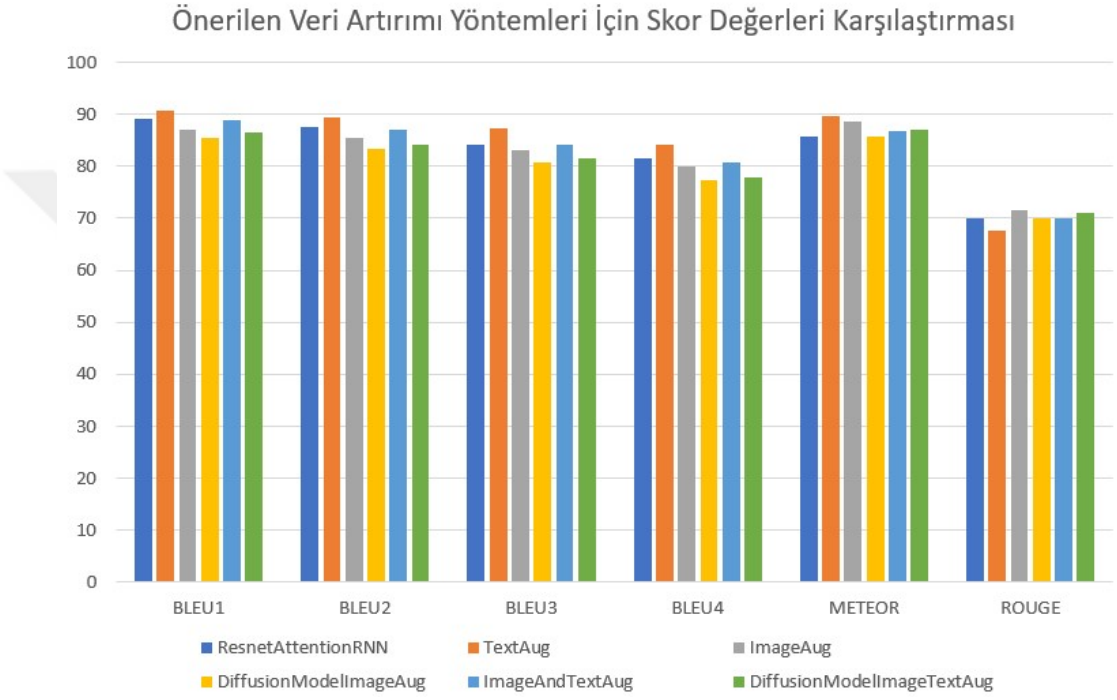


Şekil 8.1: Önerilen Modellerin Skor Değerleri Karşılaştırması

performansını pozitif yönde etkilemediği gözlemlenmiştir. Ancak METEOR ve ROUGE metriklerine göre önerilen veri artırımı yöntemleri, sistem performansına katkı sağlamıştır.

Tez çalışması kapsamında ayrıca veri artırımında farklı bir metot geliştirmek amacıyla, difüzyon model kullanılarak, bu mimarinin etkisi gözlemlenmiştir. Bu kapsamda metinlerden görüntüler üretmeyi sağlayan bir difüzyon model kullanılarak yine Flickr1K eğitim veri seti üzerinden sentetik veri üretimi gerçekleştirilmiştir. Sonuçlar incelendiğinde, görüntü rastgele döndürme (random rotation) yöntemiyle gerçekleştirilen görüntü veri artırımının, difüzyon model ile gerçekleştirilen görüntü veri artırımına göre tüm ölçümleme metrikleri dahilinde daha yüksek sonuç elde edildiği gözlemlenmiştir. Kararlı (stable) difüzyon modeli ile yapılan görüntü veri artırımı, tüm metriklerde görüntü işleme ile yapılan veri artırımına kıyasla daha düşük sonuçlar vermiştir. Bunun nedeninin, kararlı difüzyonun Flickr veri setini de içeren çeşitli veri kümelerini kapsayan LAION-5B veri kümesi üzerinde eğitilmiş

olması olduğu düşünülmektedir. Çalışma kapsamında önerilen bir diğer yöntem olan görüntü ve metinsel verilerin aynı anda artırıldığı durumlar da incelenmiştir. Bu kapsamda Tablo 7.6'dan çıkarılan bir diğer sonuç ise sistemde; difüzyon model ile görüntü artırımı ve metinsel veri artırımı birlikte uygulandığında, rastgele döndürme yöntemi ile görüntü veri artırımı ve metinsel veri artırımı uygulanması yöntemine göre ROUGE metriğine göre %0.9 ve METEOR metriğine göre %0.2 daha iyi olmaktadır. Şekil 8.2' de elde edilen sonuçlar grafik üzerinde karşılaştırılmıştır.



Şekil 8.2: Önerilen Veri Artırma Yöntemlerinin Skor Değerleri Karşılaştırması

Hem görüntü hem de görüntülere ait metinler için artırım yapıldığında elde edilen sonuçlar incelendiğinde, difüzyon model ile gerçekleştirilmiş olan görüntü veri artırımı sonucunda modelin başarısında ROUGE metriğine göre %1.1, METEOR metriğine göre %1.3 değer artışı olduğu saptanmıştır. Rastgele döndürme ve eş anlamlısıyla yer değiştirmeyi birlikte kullanan klasik veri artırımı yönteminin ise model performansını ROUGE metriğine göre %0.2, METEOR metriğine göre ise %1.1 artırmıştır.

Çalışmadaki sonuçların ilgili model ve yöntemlerin varsayılan değerleri ile elde edildiği, bu nedenle parametre değerlerinin değiştirilmesi veya eniyilenmesi durumunda sonuçların, önerilen modellerin sıralamalarının değişebileceği göz önünde bulundurulmalıdır. Ayrıca örnek olarak farklı çok sayıda olan görüntü işleme veri artırım / metinsel altyazı artırım yöntemlerinin veya bileşimlerinin de deneylere eklenmesi durumunda da daha yüksek başarı oranlarına ulaşılabileceği görülmektedir. Ek olarak, tez kapsamında önerilen GPT, dönüştürücü ve görü dönüşüm modelleri ve diğer derin öğrenme yöntemleri ile farklı kodlayıcı - kod çözücü model kombinasyonları denenerek deneysel çalışmalar genişletilebilir. Dönüştürücü modellerdeki özellikle son aylardaki gelişmeler, diğer alanlarda olduğu gibi görüntü altyazı üretiminde de gelecek vaad ettiğini göstermektedir. Tez çalışmasının odak noktası olmadığı için bu deneylerin bir kısmı kapsam dışında bırakılmıştır.

Kaynaklar

Adalıer O., Uğur A. (2008). YAPAY ZEKA YÖNTEMLERİ İLE YAZILIM PROJELERİNDE MALİYET KESTİRİMİ, 2008 EGE ÜNİVERSİTESİ Bilgisayar Mühendisliği Anabilim Dalı Doktora Tezi, İzmir/TÜRKİYE.

Alammary, A.S. BERT Models for Arabic Text Classification: A Systematic Review. Appl. Sci. 2022, 12, 5720. https://doi.org/10.3390/app12115720

Alin A. Y., Kusrini K., Yuana K. A. (2023). Data Augmentation Method on Drone Object Detection with YOLOv5 Algorithm. Eighth International Conference on Informatics and Computing (ICIC), Manado, Indonesia, 2023, pp. 1-6, doi: 10.1109/ICIC60109.2023.10382123.

Andonie R., (2010). Extreme data mining: Inference from small datasets, Int. J. Comput. Commun. Control, vol. 5, no. 3, pp. 280–291, 2010

Bahdanau D., Cho K., Bengio Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. doi: https://doi.org/10.48550/arXiv.1409.0473

Banerjee, S., Lavie A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65–72. Association for Computational Linguistics. https://www.aclweb.org/anthology/W05-0909

Biradar V. G., Agarwal M. G, S., Singh S. K. and Bharadwaj R. U. (2023). Leveraging Deep Learning Model for Image Caption Generation for Scenes Description. International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT), Bengaluru, India, 2023, pp. 1-5, doi: 10.1109/EASCT59475.2023.10393602.

Chandra K.S., Poda R., Vinod A., Arun R. A. (2023). Pixels to Phrases: Bridging the Gap with Computationally Effective Deep Learning models in Image Captioning. 12th International Conference on

Advanced Computing (ICoAC), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICoAC59537.2023.10249809.

Chang L., Changhu W., Fuchun S., Yong R. (2016). Image2Text: A Multimodal Image Captioner. In Proceedings of the 24th ACM international conference on Multimedia (MM '16). Association for Computing Machinery, New York, NY, USA, 746–748. https://doi.org/10.1145/2964284.2973831

Charleen C., Angelica H., Purnama and F. Purnomo, (2021). Impact of Computer Vision With Deep Learning Approach in Medical Imaging Diagnosis, 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), Jakarta, Indonesia, 2021, pp. 37-41, doi: 10.1109/ICCSAI53272.2021.9609708.

Chen T., Li Z., Wu J., Ma H., Su B. (2022). Improving image captioning with Pyramid Attention and SC-GAN. Image and Vision Computing, Volume 117, 104340, ISSN 0262-8856, doi: https://doi.org/10.1016/j.imavis.2021.104340.

Cheng, L., Wei, W., Mao, X., Liu, Y., Miao, C., 2020. Stack-vs: stacked visual-semantic attention for image caption generato. IEEE Access 8, 154953–154965.

Chitteti C., Madhavi K.R. (2024). Taylor African vulture optimization algorithm with hybrid deep convolution neural network for image captioning system, Multimed Tools Appl 83, 66393–66411 (2024). Springer. https://doi.org/10.1007/s11042-023-18080-0

Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. (2021). A review of medical image data augmentation techniques for deep learning applications. J Med Imaging Radiat Oncol. Aug;65(5):545-563. doi: 10.1111/1754-9485.13261.

Csanyi G.M., Orosz T., (2021). Comparison of data augmentation methods

for legal documentclassification. Acta Technica Jaurinensis, 15(1), 15–21. https://doi.org/10.14513/actatechjaur.00628

Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. (2018). Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv 2018, arXiv:1810.04805.

Doaa, B. E., Magda M.M., Adel A. E. (2023). Bi-directional Image–Text Matching Deep Learning-Based Approaches: Concepts, Methodologies, Benchmarks and Challenges, International Journal of Computational Intelligence Systems. Springer. doi: https://doi.org/10.1007/s44196-023-00260-3

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Ethayarajh K., (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. arXiv:1909.00512

Fields C., Kennington C. (2023). Vision Language Transformers: A Survey.

Fong S. J., Li G., Dey N., Crespo R. G., Herrera-Viedma E., (2020). Finding an Accurate Early Forecasting Model from Small Dataset: A Case of 2019-nCoV Novel Coronavirus Outbreak, International Journal of Interactive Multimedia and Artificial Intelligence 6.1 (2020): 132-40, https://doi.org/10.9781/ijimai.2020.02.002

Garg (2020). Flickr1K (Small flicker data for image captioning), Son erişim tarihi: 13.08.2024

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

Grossi E., Buscema M. (2007).Introduction to artificial neural networks. European journal of gastroenterology and hepatology, 19(12), 1046–1054

He, S.; Lu, Y. (2019) A Modularized Architecture of Multi-Branch Convolutional Neural Network for Image Captioning. Electronics 2019, 8, 1417. https://doi.org/10.3390/electronics8121417

Hirway C. , Fallon E., Connolly P., Flanagan K., Yadav D. (2023). A Comparative Study of Intent Classification Performance in Truncated Consumer Communication using GPT-Neo and GPT-2. International Conference on Emerging Techniques in Computational Intelligence (ICETCI), Hyderabad, India, 2023, pp. 97-104, doi: 10.1109/ICETCI58599.2023.10331337.

Isola P., Zhu J.Y., Zhou T., Efros A.A. (2017). Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017

Ji J., Wang M., Zhang X., Lei M., Qu L. (2022) Relation constraint self-attention for image captioning, Neurocomputing, Volume 501, 2022, Pages 778-789, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2022.06.062

Jin, J., Fu, K., Cui, R., Sha, F. and Zhang, C. (2015). Aligning where to see and what to tell: image caption with region-based attention and scene factorization. arXiv preprint arXiv:1506.06272

Rashtchian C., Young P., Hodosh M., Hockenmaier J. (2010). Flickr8K, Collecting Image Annotations Using Amazon's Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Son erişim tarihi: 13.08.2024

Karras T., Laine S., Aila T. (2019). A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

Karpathy A., Fei-Fei L., (2019). Deep Visual-Semantic Alignments for Generating on Image Descriptions" , IEEE Transactions on Pat-

tern Analysis Mechanism and Machine Intelligence, pp. 664-676,doi: 10.1109/TAMI.2019.2598339,2019

Karpathy A., Socher R., Le Q.V., Manning C.D., Ng A.Y. (2014). Grounded Compositional Semantics for Finding and Describing Images with Sentences. Transactions of the Association for Computational Linguistics 2014; 2 207–218. doi:10.1162/tacl_a_00177

Kiros R., Salakhutdinov R. and Zemel R. (2014). Multimodal neural language models. In Eric P. Xing & Tony Jebara (Eds.), Proceedings of the 31st international conference on machine learning, volume 32 of Proceedings of machine learning research (pp. 595–603). Bejing, China, 22–24 Jun 2014.

Kukačka J, Golkov V, Cremers D., (2017). Regularization for deep learning: a taxonomy. arXiv:1710.10686.

Kulkarni U., Tomar K., Kalmat M., Bandi R., Jadhav P., Meena S. (2023). Attention based Image Caption Generation (ABICG) using Encoder-Decoder Architecture. 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023, pp. 1564-1572, doi: 10.1109/ICSSIT55814.2023.10061040.

Kwok W. M. C. (2023). Image Captioning by ViT/BERT, ViT/GPT.

Lateh M. A., Muda A. K., Yusof Z. I. M., Muda N. A. and Azmi M. S., (2017). Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review, Journal of Physics: Conference Series, Volume 892, Conf. Ser. 892 012016.

Li S., Kulkarni G., Berg T.L., Berg A.C., Choi Y.. (2011). Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL '11). Association for Computational Linguistics, USA, 220–228.

Lin C.Y., (2004). ROUGE: A Package for Automatic Evaluation of Summaries. Book: Text Summarization Branches Out, 74–

81. Association for Computational Linguistics, Barcelona, Spain. https://www.aclweb.org/anthology/W04-1013

Liu A. ,Li J. and Ye H. (2023). A Prediction Model Combining Convolutional Neural Network and LSTM Neural Network. 2nd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS), Bristol, United Kingdom, 2023, pp. 318-321, doi: 10.1109/AIARS59518.2023.00071.

Lopez M.M., Kalita J. (2017). Deep learning applied to nlp. arXiv preprint arXiv:1703.03091.

Mayil V. V. and Jeyalakshmi T. R. (2023). Pretrained Sentence Embedding and Semantic Sentence Similarity Language Model for Text Classification in NLP, 2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP), VIJAYAWADA, India, 2023, pp. 1-5, doi: 10.1109/AISP57993.2023.10134937.

Medina P., Alejandro D. (2024). Synthetic data generation and data augmentation techniques for image captioning with Stable Diffusion and large language models. UNIVERSITAT POLITÈCNICA DE VALÈNCIA.

Muftie F., Haris M. (2023). IndoBERT Based Data Augmentation for Indonesian Text Classification. International Conference on Information Technology Research and Innovation (ICITRI), Jakarta, Indonesia, 2023, pp. 128-132, doi: 10.1109/ICITRI59340.2023.10250061.

Naga Srinivasu P, Krishna TB, Ahmed S, Almusallam N, Khaled Alarfaj F, Allheeib N (2023) Variational autoencoders-basedself-learning model for tumor identifcation and impact analysis from 2-D MRI images. J Healthc Eng 2023:1–17. https://doi.org/10.1155/2023/1566123

Niteesh K. R. and Pooja T. S. (2024). Application of Deep Learning in Detection of various Hepatic Disease Classification Using H and E Stained Liver Tissue Biopsy. 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare

and Internet of Things (AIMLA), Namakkal, India, 2024, pp. 1-6, doi: 10.1109/AIMLA59606.2024.10531423.

Patel S., Chandra S. K. and Jain A.(2023). DeepFake Videos Detection and Classification Using Resnext and LSTM Neural Network. 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2023, pp. 1-5, doi: 10.1109/SMARTGENCON60755.2023.10442131.

Papineni K., Roukos S., Ward T., Zhu W.J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 20th International Conference on Computational Linguistics, 501–507, Geneva, Switzerland. https://www.aclweb.org/anthology/C04-1072

Rohitharun S., Reddy L. U. K., Sujana S., 2022, "Image Captioning Using CNN and RNN", 2022 2nd Asian Conference on Innovation in Technology (ASIANCON) Pune, India. Aug 26-28, 78-1-6654-6851-0/22/31.00 ©2022 IEEE — DOI: 10.1109/ASIANCON55314.2022.9909146

Shahriar S. (2022). GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. Displays Volume 73, 2022, 102237, ISSN 0141-9382, 0141-9382/© 2022 Elsevier B.V. https://doi.org/10.1016/j.displa.2022.102237.

Shaikhina T., Khovanova N. A., (2017). Handling limited datasets with neural networks in medical applications: A small-data approach. Artificial Intelligence in Medicine, vol. 75, pp. 51-63, 2017

Shorten, C., Khoshgoftaar, T.M., Furht, B., (2021). Text Data Augmentation for Deep Learning. J Big Data 8, 101. https://doi.org/10.1186/s40537-021-00492-0

Shorten C., Khoshgoftaar T.M., (2019). A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60. https://doi.org/10.1186/s40537-019-0197-0

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, Furu Wei. (2021). DeltaLM: Encoder-Decoder Pre-training for Language Generation and Translation by Augmenting Pretrained Multilingual Encoders. doi: https://doi.org/10.48550/arXiv.2106.13736

Slifker J. F. and Shapiro S. S., (1980). The Johnson system: selection and parameter estimation. Technometrics, vol. 22, no. 2, pp. 239–246.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R., 2014, Dropout: a simple way to prevent neural networks from overftting. J Mach Learn Res. 2014;15(1):1929–58

Sudhakar J., Iyer V. V., Sharmila S. T. (2022). Image Caption Generation using Deep Neural Networks. 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, 2022, pp. 1-3, doi: 10.1109/ICONAT53423.2022.9726074.

Tabian I., Fu H., Khodaei Z. S. (2019). A Convolutional Neural Network for Impact Detection and Characterization of Complex Composite Structures. Sensors, 19(22), 4933. doi:10.3390/s19224933

Tan Y. H., Chan C. S. (2016). phi-lstm: A phrase-based hierarchical lstm model for image captioning. In Proceedings of Asian conference on computer vision, pages 101–117. Springer.

Tiwari V. and Bhatnagar C. (2021). Automatic Caption Generation via Attention Based Deep Neural Network Model. 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-6, doi: 10.1109/ICRITO51393.2021.9596255.

Upadhyay D., Tiwari A. and Niathani H. (2024). Sequential Pattern Analysis of Emotion in Speech With LSTM, 2024 2nd International Conference on Device Intelligence, Computing and Communica-

tion Technologies (DICCT), Dehradun, India, 2024, pp. 120-125, doi: 10.1109/DICCT61038.2024.10532789.

Veena S., Ashwin K.S., Gupta P. (2022). Comparison of various CNN encoders for image captioning, Journal of Physics: Conference Series, Volume 2335, International (Virtual) Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication and Computational Intelligence 2022 21/04/2022 - 24/04/2022 Online, Phys.: Conf. Ser. 2335 012029. doi: 10.1088/1742-6596/2335/1/012029

Vinyals O., Toshev A., Bengio S., Erhan D. (2014). Show and Tell: A Neural Image Caption Generator. CVPR, page 3156-3164. IEEE Computer Society, (2015) doi: https://doi.org/10.48550/arXiv.1411.4555

Q. Li, (2023). Application of Computer Vision Technology in Environmental Art Design, 2023 2nd International Conference on Artificial Intelligence and Computer Information Technology (AICIT), Yichang, China, 2023, pp. 1-5, doi: 10.1109/AICIT59054.2023.10277738.

Topbaş A., Jamil A., Hameed A. A.,Ali S. M., Bazai S. and Shah S. A. (2021).Sentiment Analysis for COVID-19 Tweets Using Recurrent Neural Network (RNN) and Bidirectional Encoder Representations (BERT) Models, 2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), Quetta, Pakistan, pp. 1-6, doi: 10.1109/ICECube53880.2021.9628315.

Wang Y., (2020). Basic Methodologies Used in NLP Area. 2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, 2020, pp. 505-511, doi: 10.1109/AUTEEE50969.2020.9315550.

Wang C., Gu X. (2022) Dynamic-balanced double-attention fusion for image captioning, Engineering Applications of Artificial Intelligence, Volume 114, 2022, 105194, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2022.105194

Wei J., Zou K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv preprintarXiv:1901.11196, 2019.

Xiao C., Xu S.X., Zhang K. (2023). Multimodal Data Augmentation for Image Captioning using Diffusion Models, arXiv:2305.01855v1

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., et al. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the international conference on machine learning (pp. 2048–2057)

Xu, H., Saenko, K., 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: Proceedings of the European Conference on Computer Vision. pp. 451–466

Xu K., Ba J.L., Kiros R., Cho K., Courville A., Salakhutdinov R., Zemel R.S., Bengio Y. (2015) Neural image caption generation with visual attention (2015) 32nd International Conference on Machine Learning, ICML 2015, 3, pp. 2048 - 2057, Cited 5924 times.

Xu J., Efstratios G., Basura F., Tinne T.(2015). Guiding long-short term memory for image caption generation. In Proceedings of IEEE International Conference on Computer Vision (ICCV), 7–13 Dec. 2015.

Yidan Xu, Jiaqing Liang, Yaoyao Zhuo, Lei Liu, Yanghua Xiao, Lingxiao Zhou. (2024). TDASD: Generating medically significant fine-grained lung adenocarcinoma nodule CT images based on stable diffusion models with limited sample size, Computer Methods and Programs in Biomedicine, Volume 248, 108103, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2024.108103.

Young P., Lai A., Hodosh M., Hockenmaier J. (2014). Flickr30K, Son erişim tarihi: 13.08.2024

Yu, L., Zhang, W., Wang, J. and Yu, Y. (2017). SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient.

In: Singh, S. and Markovitch, S., (eds.). AAAI Conference on Artificial Intelligence. AAAI Press, 2852–2858. Available at: http:// aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14344.

Zeng Y., Lu H., Borji A. (2019). Statics of Deep Generated Images" IEEE Transactions on Image Processing, arXiv:1708.02,2019.

Zhang, P., Xu, T., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X. (2021). "Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks" , Computer Vision Foundation, IEEE Department of Research and Development, vol.201,2021

Zhang J., Li K., Wang Z., Zhao X., Wang Z. (2021) Visual enhanced gLSTM for image captioning, Expert Systems with Applications, Volume 184, 2021, 115462, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.115462.

Zhu X., Fan J., Yan X., Mao T., Huang J., Zhang Y. (2024) "SDD: Strawberry Disease Detection Framework Based on Computer Vision Neural Networks," 2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL), Zhuhai, China, 2024, pp. 1077-1082, doi: 10.1109/CVIDL62147.2024.10603536.

Zhu J.Y., Park T., Isola P., Efros A.A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2017.

TEŞEKKÜR

Bu çalışma boyunca, değerli fikir ve önerileriyle doğru yönde ilerlememde katkısı bulunan danışmanım sayın Prof. Dr. Aybars Uğur'a ve çalışma süresince sağladıkları tüm destek için değerli aileme teşekkürü bir borç bilirim.

12/02/2025

İmzası

Ad Soyad

ÖZGEÇMİŞ

İlayda Yıldız, lisans eğitimini 2017 yılında, EGE Üniversitesi Bilgisayar Mühendisliği Bölümü'nde tamamladıktan sonra İzmir'de iş hayatına başlamıştır. Aynı zamanda öğrenim hayatına, EGE Üniversitesi FBE'nde yüksek lisans eğitimi ile devam etmiştir. 2020 yılında yüksek lisans eğitimini de tamamladıktan sonra ara vermeden doktora programına aynı üniversitede devam etmiştir. 2022 yılında, kariyer hayatına yurt dışı tecrübesini de ekleyebilmek için Münih'e taşınmış ve hayatını burada devam ettirmektedir.

Yayınlar:

1. Yildiz I., UĞUR A., 2023, "Küçük Görüntü Veri Setleri İçin Derin Öğrenme, Veri Artırımı ve Dikkat Mekanizması Kullanılarak Altyazı Üretimi", 2. ULUSLARARASI MÜHENDİSLİK VE FEN BİLİMLERİ KONGRESİ (16-17 Aralık 2023), İstanbul-Türkiye