TÜRKİYE CUMHURİYETİ
YILDIZ TEKNİK ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

İNSAN AKTİVİTELERİ İÇEREN VİDEOLARDAN YOĞUN VİDEO ÖZETİ OLUŞTURMA

20011044 — Yusuf Enes KURT
20011045 — Muhammed Ali LALE

BILGISAYAR PROJESİ

Danışman
Doç. Dr. Ali Can KARACA

Ocak, 2024

TEŞEKKÜR

Öncelikle proje seçiminde, projenin tanıtımında, proje esnasında karşımıza çıkan çeşitli problemlerde verdiği desteklerle içimizi rahatlatan, bize yol gösteren sevgili danışman hocamız Doç. Dr. Ali Can Karaca'ya sonsuz teşekkürlerimizi ve saygılarımızı sunarız.

Ayrıca bugünlere gelmemizde büyük emeği olan Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği bölüm hocalarımıza teşekkür ederiz.

Yusuf Enes KURT
Muhammed Ali LALE

İÇİNDEKİLER

KISALTMA LİSTESİ

AI	Artificial Intelligence
ANN	Artificial Neural Networks
BLEU	BiLingual Evaluation Understudy
CIDEr	Consensus-based Image Description Evaluation
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DL	Deep Learning
DLNN	Deep Learning Neaural Networks
GloVe	Global Vectors
GPU	Graphics Processing Unit
I3D	Inflated 3D ConvNet
LCS	Longest Common Subsequence
LSTM	Long Short-Term Memory
METEOR	Metric for Evaluation of Translation with Explicit Ordering
ML	Machine Learning
NLP	Natural Language Processing
RNN	Recurrent Neural network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
TPU	Tensor Processing Unit
VGG	Visual Geometry Group

ŞEKİL LİSTESİ

TABLO LİSTESİ

ÖZET

İNSAN AKTİVİTELERİ İÇEREN VİDEOLARDAN YOĞUN VİDEO ÖZETİ OLUŞTURMA

Yusuf Enes KURT
Muhammed Ali LALE

Bilgisayar Mühendisliği Bölümü
Bilgisayar Projesi

Danışman: Doç. Dr. Ali Can KARACA

Bu çalışmada videoların görsel ve ses özelliklerini dikkate alan, bunları birbiriyle entegre eden ve bu bilgiler ışığında detaylı video açıklamaları üreten bir sistem sunulmaktadır. Mevcut çalışmalar genellikle sadece görsel ya da sadece ses verilerini analiz etmektedir. Bu projede her iki veri türünü de kapsayan bir çalışma yapılmıştır.

Model eğitirken insan beynini taklit eden makine öğrenimi ve doğal dil işleme yaklaşımları tercih edilmiştir. Videonun özelliklerinin işlenmesi, videonun bölümlere ayrılması ve cümleler üretilmesi için Bi-modal Transformer modeli kullanılmıştır.

Bu proje; görme ve işitme engelli bireylere, sosyal medya platformlarına ve video içerik sağlayıcılarına yara sağlayabilecek bir çalışmadır.

Veri seti olarak insan aktivieleri içeren videolar kullanılmıştır. Yazımızda Bi-modal Transformer'ın ayrıntılı içeriğinden, "caption" ve "proposal" modellerinin oluşumundan ve videoda metin yazdırmak için nasıl kullanıldığından bahsedilmiştir.

Anahtar Kelimeler: Yoğun video özetleme, Bimodal Transformer, makine öğrenimi, derin öğrenme, CNN, LSTM, özellik çıkarma, encoder, decoder, Proposal Generation

ABSTRACT

DENSE VIDEO CAPTIONING FROM HUMAN ACTIVITY VIDEOS

Yusuf Enes KURT
Muhammed Ali LALE

Department of Computer Engineering
Computer Project

Advisor: Assoc. Prof. Dr. Ali Can KARACA

This paper presents a system that takes into account the visual and audio features of videos, integrates them and generates detailed video descriptions based on this information. Existing studies usually analyse only visual or only audio data. In this project, both types of data are analysed.

Machine learning and natural language processing approaches that mimic the human brain are used to train the model. The Bi-modal Transformer model was used to process the features of the video, segment the video and generate sentences.

This project is a work that can benefit visually and hearing impaired individuals, social media platforms and video content providers.

Videos containing human activities were used as a data set. In this paper, the detailed content of the Bi-modal Transformer, the formation of the "caption" and "proposal" models and how it is used to print text in the video are mentioned.

Keywords: Dense video captioning, Bimodal Transformer, machine learning, deep learning, CNN, LSTM, feature extraction, encoder, decoder, Proposal Generation

1
Giriş

Projenin içeriğini anlamak için ilk başta “Makine Öğrenimi”, “Derin Öğrenme” ve “Doğal Dil İşleme” gibi konularda gerekli bilgiler verilecektir.

1.1 Makine Öğrenimi

Makine öğrenimi (ML), verilerden elde edilenleri kullanıp tahminde bulunmayı amaçlayan, bilim ve teknolojinin birlikte kullanıldığı bir çalışma alanıdır. Bilgisayarların bir insan gibi öğrenmesini sağlama düşüncesiyle ortaya çıkmıştır. Makine öğreniminin temel adımlarını şu şekilde sıralayabiliriz [1]:

- **Veri Toplama:** Model eğitimi ve testi için gerekli olan uygun ve çeşitli veriler temin edilir.
- **Veride Ön İşlemlerde Bulunma:** Alınan ham veriler, model eğitimi için uyarlanır. Bu kısımda azaltma, normalizasyon ve dönüştürme gibi işlemler yapılır.
- **Model Belirleme:** Proje çalışmasını uygun algoritmayı sağlayan model seçilir.
- **Modeli Eğitme:** Ön işlemden geçirilmiş veriler ile model beslenir ve eğitilir. Bu kısımda model; verilerin özelliklerini ve arasındaki bağları, yakınlık derecelerini öğrenmeye çalışır.
- **Değerlendirme:** Eğitilmiş model, test için ayrılmış veriler ile test edilir ve performansı çeşitli ölçütler kullanılarak ölçülür.
- **Hiper Parametre Ayarı:** Çeşitli hiper parametrelerde değişiklik yaparak modelin farklı performansları ölçülür ve en iyi sonucu veren model seçilir.

1.2 Derin Öğrenme

Derin öğrenme, makine öğreniminin alt dalıdır. İnsan beyninin davranışını taklit ederek en az üç katmana sahip bir sinir ağı ile büyük miktarda veriyi öğrenmeyi sağlar. İnsana ihtiyaç duymadan analitik ve fiziksel görevleri yerine getirir. Sinir ağının katmanlarını arttırmak, doğru sonuçlar almaya ve iyileştirmeye yardımcı olabilmektedir.[2]

Derin öğrenme sinir ağları (DLNN), tahmin veya doğru kategorik ayırma için her biri önceki katman üzerine inşa edilen, birbirine bağlı düğümlerden oluşan, çok katmanlı bir yapıdır. Çeşitli hesaplamaların sinir ağı boyunca ilerlemesine “ileri yayılım” denir. Ağın en sol tarafındaki giriş ve en sağ tarafındaki çıkış katmanlarına “görünür katmanlar” denir. Giriş katmanından veri alınır, çıkış katmanından tahmin elde edilir. Tahmindeki hataları hesaplamak için “geriye yayılma” yöntemi kullanılır.[2] Belirli sorunları veya veri kümelerini ele almak için farklı sinir ağı türleri vardır. Bunlara CNN ve RNN örnek verilebilir. Bu algoritmalar raporun "Sistem Tasarımı" kısmında ayrıntılı bir şekilde anlatılmıştır. Günümüzde derin öğrenme; finansal hizmetler, müşteri hizmetleri ve sağlık hizmetleri gibi alanlarda kullanılmaktadır.

1.3 Doğal Dil İşleme

Doğal Dil İşleme, bilgisayarın oluşturduğu verileri veya çıkardığı sonuçları insanların anlayabileceği dile dönüştürmeyi amaç edinmiş bir bilim dalıdır. Problemlerin çözümleri için makine öğrenmesi ve derin öğrenme yaklaşımlarıyla ortak bir şekilde çalışır. Bilgisayarların gerekli görevleri yerine getirip doğal dilleri anlaması için araştırmacılar, insanların dili nasıl anladığını ve kullandığını derinlemesine incelemişlerdir. Doğal dil işlemenin temelleri, bilgi bilimleri, dilbilim, matematik, yapay zekâ, psikoloji vb. bir dizi disiplinde yatmaktadır.[3]

Bilgisayar terminolojisinde “dil” kavramı denilince akla C, Java gibi programlama dilleri geldiğinden bu bilim alanına insanların birbirleri ile iletişim için kullandıkları “doğal dil” denmiştir.

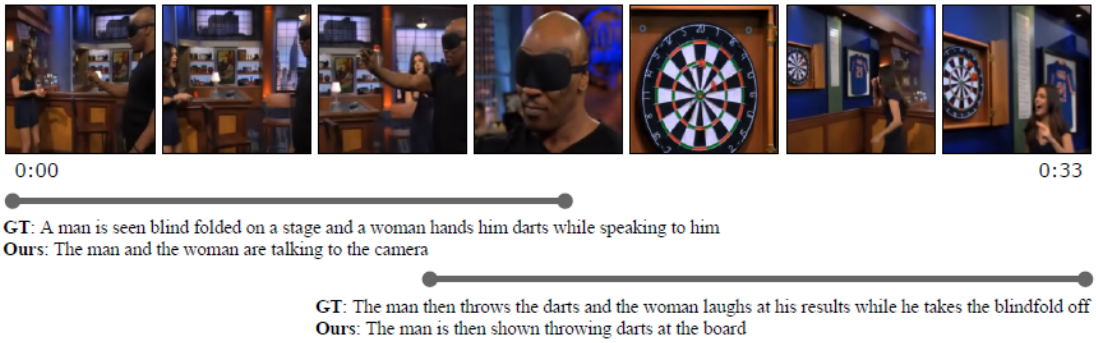
Doğal dil işlemenin; makine çevirisi, metin işleme ve özetleme, diller arası bilgi erişimi, konuşma tanıma, ses ve görsellerden metin üretme vb. çalışma alanları bulunmaktadır.

1.4 Projenin Amacı

Bu projede ana hedef, videolarda bulunan görsel ve ses verilerini etkili bir şekilde analiz edip yorumlamaktır. Günümüzde bulunan çalışmalar sadece görsel veya sadece ses verilerini dikkate alan çalışmalardır. Fakat bir video analizinde ikisini de dikkate almak büyük önem arz etmektedir. Bir insan, dış dünyadaki verileri iyi algılayabilmesi için birden fazla duyu organını kullanır ve böylelikle daha anlamlı çıkarımlarda bulunur. Bu çalışma, insan davranışlarını örnek alarak makine öğrenimi, doğal dil işleme gibi teknolojileri "Bi-modal Transformer" modelini kullanarak bir araya getirmiştir. Bu model her iki tür veriyi (video-ses) de entegre eder ve daha anlamlı çıkarımlarda bulunur.

Bi-modal Transformer, video içerisindeki her sahne veya olayı tanımlayarak bunlardan anlamlı cümleler üretmeyi hedefler. Böylelikle görme engelli bireyler için video içeriğini anlamakta, veri analistleri için büyük video verileri yerine “caption”ları veri kümesi olarak kullanıp videoları daha hızlı analiz etmesinde büyük yararı dokunacaktır.

Projenin hedeflediği sonuç örneği Şekil 1.1’de gösterilmiştir.



Şekil 1.1 Caption Example [4]

2
Ön İnceleme

2.1 Benzer Çalışmalar

Son yıllarda videolardan özellik çıkarıp bu özelliklerle video içeriği özetleme çalışması yapan birçok proje bulunmaktadır. “ActivityNet Captions” verileri, bu projeler için veri kümesi olarak kullanılmıştır.

Daha öncesinde "Dense-Captioning Events in Videos" adında videonun sadece görsel verilerini alarak videodaki olayları yoğun bir şekilde altyazılandıran, videodaki birden fazla olayı tespit eden bir çalışma mevcuttur. [5] Benzer olarak "Reconstruction Network for Video Captioning" adında sadece görsel verileri işleyen ve tek bir cümle üreten çalışma da mevcuttur. [6] Bizim çalışmamızda olduğu gibi videoda dikkat mekanizmasının önemsendiği, LSTM ile videonun önemli kısımlarına dikkat çeken "Video Captioning With Attention-Based LSTM and Semantic Consistency" adında ve buna benzer çalışmalar bulunmaktadır. [7]

2.2 Projeye Olan İhtiyaç?

- Projede üretilen altyazıları sesli açıklayarak video içeriğinin görme engelli bireylere erişilebilir hale getirebiliriz. Ayrıca görsel verinin yanında ses verisi de işlendiğinden işitme engelli bireyler de videonun ses içeriğinden haberdar olabilmektedir. Bu yüzden hem görme hem de işitme engelli bireylere video içeriğini anlama konusunda büyük fayda sağlamaktadır.

- Eğitim içerikli videolarda özet etiketlerinin (captions) bulunması, öğrencilerin video içeriğini kavramlara dökmesini ve konuyu daha iyi anlamasını kolaylaştırabilir.

- İçerik sağlayıcıları; kendilerinde bulunan video arşivlerini, içerik etiketleri sayesinde daha kolay sınıflandırabilir ve izleyicilerinin ne tür içeriklerle ilgilendiğini daha kolay anlayabilir. Böylelikle izleyicilerine tamamen hitap eden video içerikleri oluşturabilirler.

- Hırsızlık olaylarının geçtiği bir video veri setinin hazırlanması ile güvenlik kameraları tarafından kaydedilen videoların analizinde önemli olaylar otomatik olarak tanımlanabilir. Böylece güvenlik görevlileri hırsızlık vb. olaylara hızlıca müdahale edebilir.
- Sosyal medya platformları, bu tür çalışmalarla izleyicilerin ana sayfasına onlara daha çok hitap eden videolar tanımlayabilir.

2.3 Proje Kapsamı

Projemizden beklenilen işlemler;

- Görsel ve ses verilerinin ayrı ayrı analiz edilmesi ve analiz edilen bilgilerin birbiriyle entegrasyonu,
- Videolarda sahne değişiminin doğru bir şekilde tespit edilmesi ve bu sahnelerdeki her önemli an için anlamlı cümle üretilmesi,
- Sistemin performansının BLEU, METEOR, CIDEr vb. ölçütlerle değerlendirilmesi ve sonuçlara göre sistemin kalitesinin iyileştirilmesidir.

2.4 Proje Gereksinimleri

Projede modeli eğitmek için cümleler ile etiketlenmiş veri kümesine ihtiyaç duyulmaktadır. Bu sebeple ActivityNet Captions adında insan aktiviteleri içeren video kümesine gereksinim duyulmuştur.

Verilerin çeşitli makine öğrenimi ve derin öğrenme yöntemleriyle modellenmesi için PyTorch, NumPy ve Pandas gibi araçlara ve kütüphanelere ihtiyaç vardır.

Bu tür kütüphanelerin kullanılabilmesi ve gerekli algoritmaların model üzerinde kodlanarak uygulanması için ise Anaconda Navigator, Jupyter Notebook ve Colab gibi platformlara ihtiyaç duyulmuştur.

2.5 Veri Kümesi

Bu çalışmada ActivityNet Captions veri kümesi kullanılmıştır. Bu özel olarak geliştirilen geniş kapsamlı bir karşılaştırma veri kümesidir. 849 saatlik video içeriğinde 20.000'e yakın video bulunmaktadır. Bu videolar, geniş bir kategori yelpazesini kapsayan video arama motorlarından derlenmiştir. Ortalama olarak her bir videoda 4 cümle bulunmaktadır. Her cümle, video içindeki özgün bir olayı ve bu olayın farklı zaman aralıklarındaki açıklamasını içerir. Cümlelerin ortalama uzunluğu yaklaşık olarak 13.48 kelimedir.

Her bir videonun kaç cümle ile özetlendiğini, bu cümlelerin hangi saniyeleri temsil ettiğini, videonun uzunluğunu, videonun id'sini, videonun hangi fazda olduğunu (train ya da validation) ve video cümlesinin indeks numarasını tutan .csv dosyaları bulunmaktadır. Bu dosyalar, hem video feature dosyalarına erişimde hem de cümleler ile featureları birbiriyle ilişkilendirmede yardımcı olur. Bu csv dosyalarının içeriğine dair bir kısım Şekil 2.1'de gösterilmiştir.

video_id	caption	start	end	duration	phase
v_QOISCBRmfWY	A young woman is se	0.83	19.86	82.73	train
v_QOISCBRmfWY	The girl dances arou	17.37	60.81	82.73	train
v_QOISCBRmfWY	She continues danci	56.26	79.42	82.73	train
v_ehGHCYKzyZ8	The video starts with	0	2.78	61.72	train
v_ehGHCYKzyZ8	A man and woman a	3.09	61.72	61.72	train
v_ehGHCYKzyZ8	The woman lays on t	15.43	55.24	61.72	train
v_ehGHCYKzyZ8	The man starts point	17.59	54	61.72	train
v_ehGHCYKzyZ8	The woman begins t	39.81	54.62	61.72	train

Şekil 2.1 Veri Kümesi Yapısı

Video featureları görsel ve ses olarak iki ayrı klasörde tutulmaktadır. Görsel featurelar, "i3d-25fps-stack64step64-2stream-npy" adlı klasörde bulunmaktadır. Her bir videonun "flow" ve "rgb" türünde 2 çeşit feature dosyası bulunmaktadır. Ses featureları ise "vggish-npy" adlı klasörde bulunmaktadır.

3
Fizibilite

3.1 Teknik Fizibilite

Bu bölümde projenin teknik, iş gücü ve zaman, yasal ve ekonomik fizibilitelerinden bahsedilmiştir.

3.1.1 Yazılım Fizibilitesi

Proje kapsamında programlama dili olarak Python 3.7 kullanılmıştır. Python dili makine öğrenimi ve görüntü işleme projelerinde yaygın olarak kullanılmasının yanı sıra basitliği ve okunabilirliği açısından da avantajlı görüldüğü için seçilmiştir. Veri bilimi ve makine öğrenimi projeleri için tasarlanmış Anaconda kullanılarak sanal ortamlar yaratılıp kullanılmıştır. Bunlar sayesinde güçlü ve esnek bir yazılım ekosistemi oluşturulmuştur. Numpy, Pandas, Argparse gibi yardımcı kütüphanelerle birlikte derin öğrenme teknikleri için PyTorch, Tensorboard, Scikit-learn ve özel olarak doğal dil işleme teknikleri için Spacy ile Torchtext kütüphaneleri kullanılmıştır. Bu kütüphanelerin tamamının açık kaynak kodlu olup ücretsiz kullanılabilmeleri de kütüphane tercihleri sırasında göz önünde bulundurulmuştur.

3.1.2 Donanım Fizibilitesi

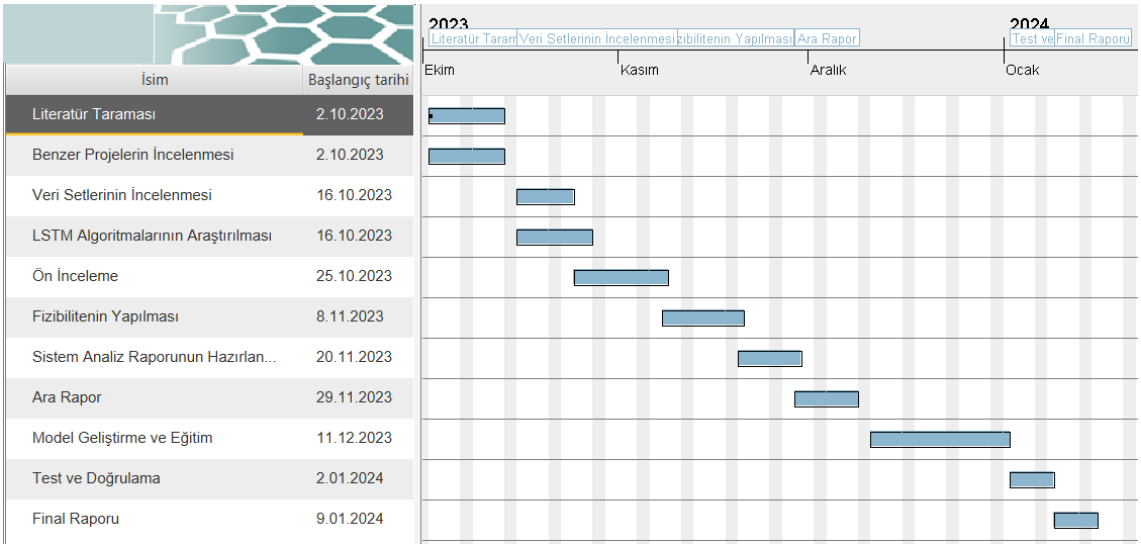
Projemiz için donanım bileşenleri olarak Monster Abra A7 ve Lenovo Ideapad Gaming 3 Model bilgisayarlar kullanılmıştır. İki bilgisayar da 16 GB RAM ve 1 TB SSD kapasitesiye sahiptir. Bilgisayarların GPU'ları 4 GB VRAM'e sahip NVIDIA GeForce GTX1650 olup Ryzen 5 5600H ve Intel I5 10300H CPU'ya sahiptirler. Bilgisayarlar Windows 11 Pro (6400TL) işletim sistemine sahiptir ancak Pro sürümü gerekli olmayıp başka bilgisayarlarda da Windows olduğu varsayılarak fizibiliteye eklenmemiştir.

Proje yüksek GPU ve TPU içeren sistemlerin varlığını gerektirdiğinden, train aşamasında bilgisayarların kendi işlemci ve grafik üniteleri kullanımının verimsiz olacağından mütevellit projenin derlenip çalıştırılması için T4, V100, A100 gibi farklı

işlem güçlerine sahip grafik kartları ile işlemlerin daha hızlı bir şekilde yapılmasına olanak sağlayan Google Colab ortamı seçilmiştir. Aylık maliyeti 165,60TL'dir. Sonuçların değerlendirilmesi için evaluate aşamasında yine Google Colab veya bilgisayarların kendi işlemcileri kullanılabilir. Daha yüksek GPU ve CPU gücüne sahip bilgisayarların varlığında Google Colab ortamı kullanımına gerek kalmayacaktır. Bununla birlikte Colab kullanımında da daha düşük işlem gücüne sahip bilgisayarlar kullanılabilir.

3.2 İş Gücü ve Zaman Fizibilitesi

Planlanan işilerin tamamlanması için gereken zaman Şekil 3.1'de gösterilmiştir.



Şekil 3.1 Gantt Diyagramı

3.3 Yasal Fizibilite

Kullandığımız veri kümesi ActivityNet Captions Dataset, ActivtyNet Event Dense-Captioning Challenge kapsamında akademik amaçlarla dense-video captioning için kullanıldığı müddetçe ücretsiz ve erişilebilir bir veri kümesidir. Projemizde bu veri kümelerinin kullanımı karşısında herhangi bir yasal engel yoktur. Projede Python diliyle birlikte kullanılan kütüphaneler de açık kaynaklı ve ücretsizdir. Projede herhangi bir yasa veya yönetmeliği ihlal edecek bir etken bulunmamaktadır.

3.4 Ekonomik Fizibilite

Kullanılan geliştirme ortamı Google Colab'in aylık ücreti 165,60TL'dir. Kullandığımız kütüphane ve frameworkler ücretsizdir. Ticari bir amaçla kullanılmadığı takdirde kullandığımız veri kümesi de ücretsizdir.

Maliyet tablosu Tablo 3.1'de gösterilmiştir.

Tablo 3.1 Ekonomik Fizibilite

Açıklama	Adet/süre	Fiyat	Toplam
Google Colab Pro/ay	2 ay	165.60TL	331.20TL
Kullanılan kütüphane ve diller	2 ay	0TL	0TL
Lenovo Ideapad Gaming 3	1	20000TL	20000TL
Monster Abra A7	1	20000TL	20000TL

4
Sistem Analizi

Bu kısımda, sistem oluşturan bileşenler teker teker ele alınmıştır.

4.1 Gereksinimler

- İnsan aktivitleri içeren videolardan görsel ve işitsel featurelar oluşturulması gerekmektedir.
- İşlenen videolardaki farklı sahnelerin tespit edilip her sahne için farklı captionlar üretilmesi (dense video captioning) sağlanmalıdır.
- CNN ve LSTM algoritmaları kullanılarak görüntü işleme ve doğal dil işleme teknikleri bir arada kullanılmalıdır. Bu tekniklerle sadece görsel, sadece işitsel veya hem görsel hem işitsel featurelar kullanılarak bir video captioning modeli tasarlanıp eğitilmelidir.
- İnsan aktiviteleri içeren bir videonun eğitilmiş sistemde test edilmesi ve captionlar içeren bir çıktı alması gerekmektedir.

4.2 Hedefler

Çalışmamızda birincil hedeflerimiz şunlardır:

- Videolar içerisindeki frameler incelenerek nesne tanıma, sahne segmentasyonu gibi görevler yerine getirilip videolar doğru şekilde analiz edilmeli ve doğru şekilde bölütlenmelidir.
- İnsan aktiviteleri içeren videolar için anlamlı cümleler içeren captionlar üretilebilmelidir.
- Üretilen bu cümleler videonun belirtilen süreler arasında mevcut bulunan sahnesiyle ilişkili olup bu sahneyi açıklamalıdır.

- Captionların içeriği video içeriği ile uyuşmalıdır.
- Bimodal analiz ile görsel ve işitsel veriler etkili bir şekilde birleştirilmeli ve videolar için daha kapsamlı analizler gerçekleştirilmelidir.
- Sadece görsel, sadece işitsel veya hem görsel hem işitsel veriler kullanılarak üç farklı model eğitilip bu modellerin vereceği sonuçlar karşılaştırılıp yorumlanmalıdır.

4.3 Performans Kriterleri

Image captioning modellerinde üretilen captionların doğruluk oranlarını ölçmek için bazı yaygın metrikler kullanılır. Projemizde bizim kullandığımız metrikler: BLEU, METEOR, ROUGE-L, CIDEr.

4.3.1 BLEU

Oldukça yaygın kullanılan bir istatistik olan BLEU (Bilingual Evaluation Understudy) üretilen captionların veri kümesinin asıl captionlarına (ground truth) kıyasla ne kadar doğru olduğunu değerlendirir. Tahmin edilen ve olması gereken captionlar arasındaki uyum BLEU score olarak belirtilen bir puan ile gösterilir. Daha yüksek puan daha yüksek eşleşme anlamına gelmektedir. [8]

4.3.2 METEOR

Bir başka metrik olan METEOR (Metric for Evaluation of Translation with Explicit Ordering), daha popüler olan BLEU metriğinde bulunan bazı sorunları gidermek için tasarlanmıştır. Tam kelime eşlemenin yanı sıra kök ve eş anlamlı kelimeleri eşleştirme gibi diğer metriklerde bulunmayan özellikleri vardır. [9]

4.3.3 ROUGE-L

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) doğal dil işlemede makine tarafından üretilen metinler ile referans olan gerçek metinleri karşılaştırmada kullanılan bir metriktir. ROUGE metrikleri 0 ile 1 arasında değişir ve daha yüksek puan daha yüksek bir uyuşmayı gösterir. ROUGE-L ile LCS (Longest Common Subsequence) problemi ile ortak alt dizinlerin uzunluğunu da hesaplar ve referans ve tahmin sonuçlarının kaç n-gram ortak olduklarını hesaplar. [10]

4.3.4 CIDEr

CIDEr (Consensus-based Image Description Evaluation) istatistiği oluşturulan tahminlerin çeşitliğini, referansları ile uyumu dikkate alarak captionların orijinalliğini değerlendirir.

5
Sistem Tasarımı

Projemiz videolarda frameler arası değişiklikleri tespit edip videoları sahnelere ayırmaya odaklanır. Belirlenen bu sahneler için video özetini içeren captionlar doğal dil işleme yöntemleri kullanarak üretilir. Hem videoların analizi sırasında hem de captionların üretimi sırasında sinir ağları önemli bir rol oynamaktadır. Görüntülerin analiz edilip caption üretildiği projelerde CNN (convolutional neural network) videodan özellik çıkarımında, RNN (recurrent neural network) ise sıralı veri işlemede çok başarılıdır.

5.1 Sinir Ağları

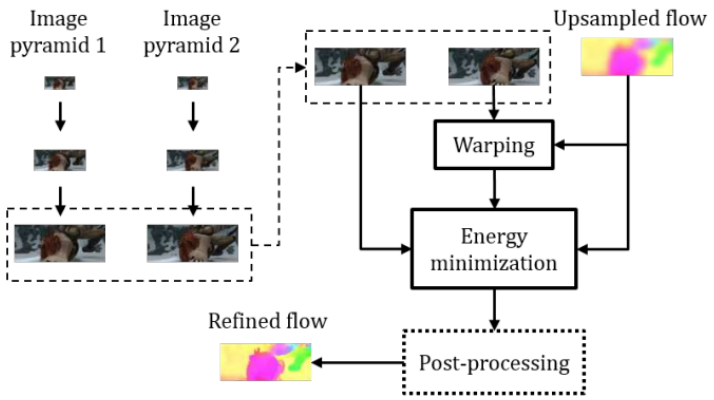
Sinir ağları, insan beynindeki sinirlerden esinlenerek tasarlanan bir yapay zeka çalışmasıdır. Karmaşık verileri işlemek ve öğrenmek için kullanılmaktadır. Bu sinir ağları girdi, gizli ve çıktı katmanları olmak üzere üç ana katmandan oluşmaktadır.

Projemizde görüntülerden özellik çıkarmak ve caption üretmek için CNN ve RNN'den yararlanılmıştır. RNN'ler, CNN'lerin görüntüden çıkardığı özelliklere dayanarak görüntünün metinsel bir açıklamasını oluşturur. Bu aşamada model; görsel ve metinsel özellikler arasındaki ilişkiler gözetilerek, veri kümesi kullanılarak eğitilir.

5.1.1 CNN

CNN, görüntü tanıma ve işleme amacıyla tasarlanmış olan bir sinir ağı yapısıdır. Özellikle 2 veya 3 boyutlu verilerle çalışmada kullanılır. Görüntünün desenlerini tanımak için özelleşmiş konvolüsyon katmanları içerirler. Bu katmanlar gerekli filtreleme işlemlerini yaparlar.

CNN, video içeriği üzerinde etkili analizler yapmamıza olanak sağlar. Video captioning projemizde her bir karedeki önemli görsel özellikleri çıkarmak için CNN'i kullandık. Bu işlemin bir tasviri Şekil 5.1'de gösterilmiştir. CNN, her bir katmanında farklı özellik haritaları oluşturarak görsel özellikleri önemlerine göre temsil eder.



Şekil 5.1 CNN ile Özellik Çıkarımı [11]

Görsel analiz aşamasında I3D (Inflated 3D ConvNet) modelini kullandık. I3D üçüncü bir boyut olarak zaman boyutunu da içeren CNN mimarisi ile video verilerini işlemek üzerine tasarlanmıştır.

Projemiz sadece görsel verilerle değil aynı zamanda işitsel verileri de analiz ederek bu bilgileri entegre bir şekilde kullanmayı amaçlar. Görüntü ve ses verilerinin birleştirilmesi video içeriğinin daha doğru bir şekilde anlaşılmasına olanak sağlar. Özellikle video içerisindeki konuşmalar veya diğer ses verileri videoyu anlayıp daha doğru captionlar üretebilmek için kritik öneme sahiptir.

İşitsel özellikle hazırlarken CNN mimarisini temel alan önceden eğitilmiş VGGish derin öğrenme modelini kullandık. VGGish ses verileri üzerinde etkili bir biçimde çalışabilmesi için özel olarak geliştirilmiştir.

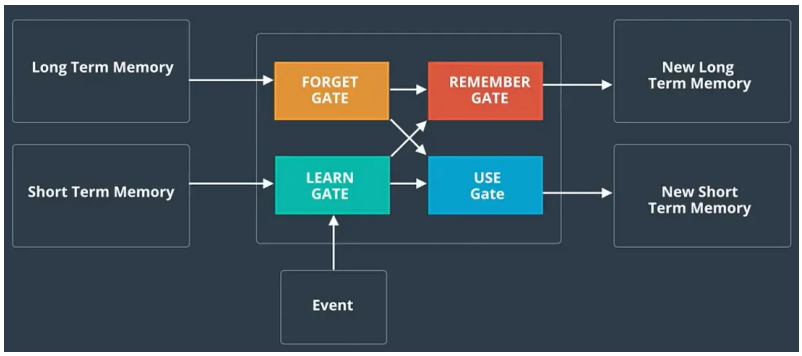
5.1.2 RNN

RNN, genellikle önceden işlenmiş olan veriler ile bir sonraki adımı tahmin etmek için kullanılan bir derin öğrenme algoritmasıdır. Sıralı verileri işleyebilmesi için geçmiş zaman adımlarındaki bilgileri saklayan ve ilerideki adımlara ileten bağlantılara sahiptir. Yapılan tahmin ile gerçek değer karşılaştırılır ve bu işlemden sonra bir hata değeri elde edilir. Bu hata değeri kullanılarak düğümlerde gradient hesaplanır ve geriye yayılım (back propagation) gerçekleştirilir.

Geleneksel RNN algoritmalarında back propagation adımında ağırlıkların yok olması bir problemdir. Bu problemi çözmek için LSTM algoritması tasarlanmıştır.

5.1.2.1 LSTM

LSTM algoritmaları, geleneksel RNN algoritmalarından farklı olarak bilgi akışını düzenleyen kapılara (gates) sahiptirler. Şekil 5.2'de gösterilen bu kapılar şu işlemleri görmektedirler:



Şekil 5.2 LSTM Akış Şeması [12]

Unutma Kapısı (Forget Gate): Önceki adımlardan uzak tutmak için neyin uygun olduğuna karar verir.

Hatırlama Kapısı (Remember Gate): Mevcut adımdan hangi bilgilerin ekleneceğine karar verir.

Öğrenme Kapısı (Learn Gate): Yeni olayları (event) ve kısa süreli hafızaları nasıl birleştireceğini belirler.

Çıktı Üretme: Bir sonraki durumun ne olacağını belirler.

Projemizde ise LSTM, videolardaki önemli aktiviteleri bulmakta ve bu aktiviteler için uygun ve anlamlı cümleler oluşturmaktadır. Video içindeki kritik anları belirlerken zaman serisi analiz yeteneğinden faydalanmaktadır. Sahne değişimini ise görsel veya işitsel verilerin zamana göre değişiminde tespit etmektedir.

5.2 Transformer

Transformer kavramı, literatüre Google'ın 2017'de yayınladığı "Attention Is All You Need" makalesiyle girmiştir. Doğal dil işleme konusunda devrim niteliğinde bir yapay zeka modelidir. Transformer yaklaşım modelinin temel özellikleri şu şekildedir: [13]

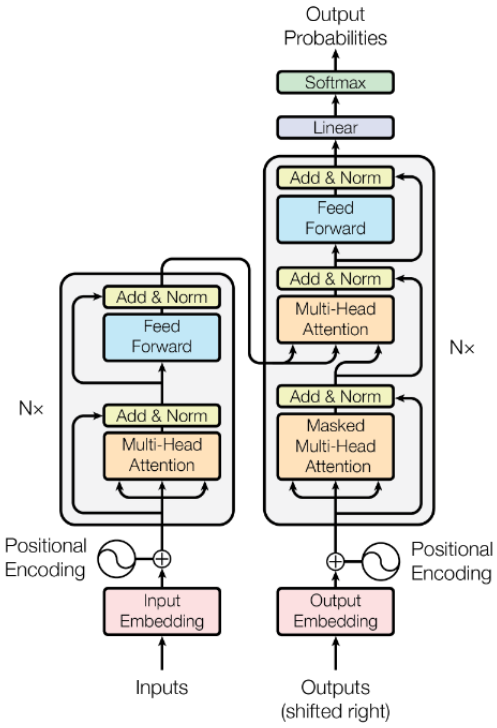
Attention Mekanizması: Transformer, featurelardaki bilgiyi işlerken verinin tamamına eş zamanlı olarak odaklanır. Transformer öncesi modellerde bu durum sıralı bir şekilde olmaktaydı. Fakat Transformer bu yaklaşımıyla veri içindeki önemli kısımlara odaklanmayı kolaylaştırır. Böylece, verinin farklı kısımları arasındaki ilişkiyi

daha iyi anlayabilmektedir.

Parallel Process: Transformer, veriyi sıralı bir şekilde değil, paralel işleyerek model eğitme süresini kısaltmıştır. Böylelikle, daha büyük veri kümelerini kullanmayı sağlamıştır.

Esneklik: Transformer, resimden kelime dizisi, kelime dizisinden kelime dizisi, kelime dizisinden vektör modelleri gibi geniş bir doğal dil işleme görevlerini kapsamaktadır.

Transformer mimarisi Şekil 5.3'te gösterilmiştir.



Şekil 5.3 Transformer Mimarisi [13]

5.2.1 Bi-Modal Transformer

İki modlu Transformerlar 2 adet veri türünü aynı anda işleyebilmektedirler. Bu durum, verilerden daha anlamlı ve zengin çıkarım sağlar. Bizim projemiz kapsamında Transformer, "Encoder"da görsel ve ses bilgilerini birbiriyle entegre edip analiz eder ve "Decoder"da bu analizleri metinler ile bağlaştırarak modeli eğitir. Projemiz kapsamında Bi-modal Transformer'da "Encoder" ve "Decoder" yapısının ayrıntılı içeriği "Uygulama" kısmında açıklanmıştır.

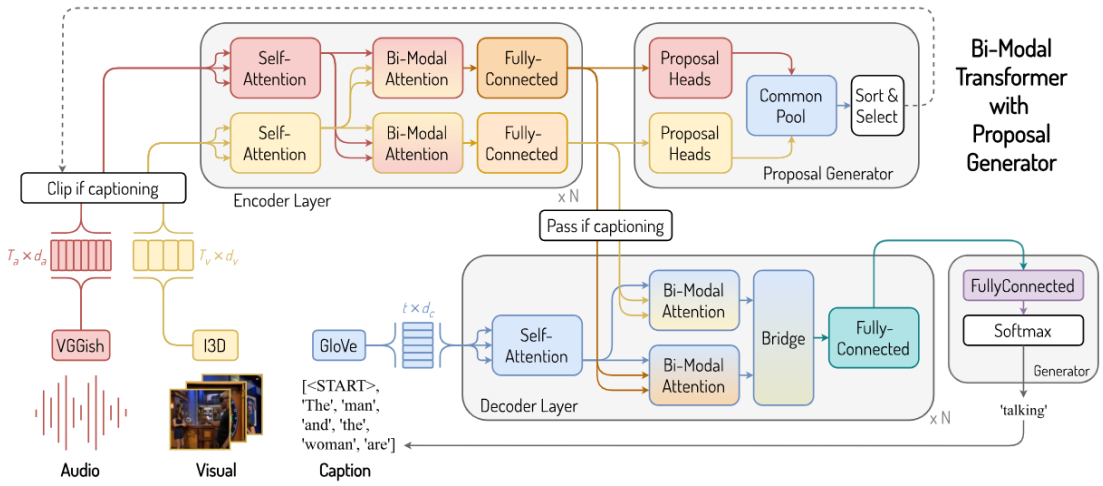
6
Uygulama

6.1 Programın Çalışma Mantığı

Ses ve görsel featurelar VGGish ve I3D olarak, captionlar GloVe olarak hazır çıkarılmış featurelardır. ActivityNet Captions'ın bu hazır çıkarılmış featureları kullanılmıştır.

Daha sonrasında bu featurelar "Encoder"da işlenir ve "Decoder" ve "Proposal Generator"a gönderilir. "Decoder", özellikleri yazılar ile bağlama; "Proposal Generator", videonun önemli anlarını düzgün bir şekilde tespit etme işleminden sorumludur.

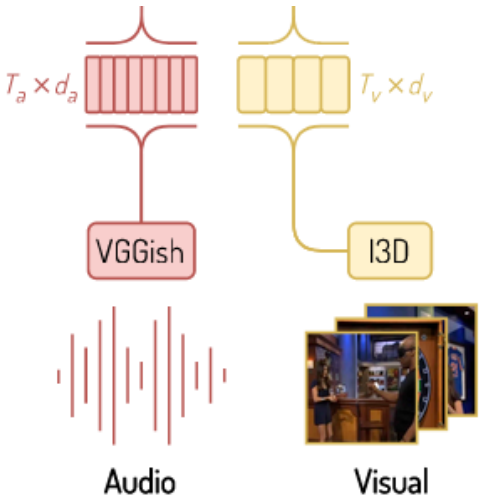
Bu modüllerin birlikte çalışmaları Şekil 6.1'de gösterilmiştir.



Şekil 6.1 Bimodal Transformer [4]

"VGGish" ve "I3D" (Inflated 3D ConvNets), özellikle ses ve video işleme alanlarında kullanılan iki farklı türde derin öğrenme modelidir. VGGish, görsel nesne tanıma için tasarlanmış popüler VGG modelinin bir varyasyonudur. Derin katmanları ve etkili özellik çıkarımı ile bilinir. I3D, başlangıçta 2D ConvNet mimarilerinin video analizi için uyarlanmış bir versiyonudur. Model, 2D konvolüsyon katmanlarını şişirerek (inflate) 3D konvolüsyon katmanlarına dönüştürür. Bu, modelin zaman boyutunu

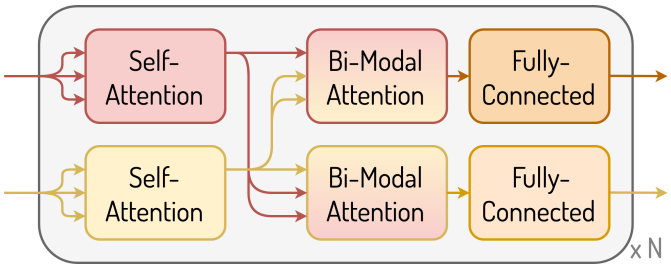
da dikkate alarak video verisini daha etkili bir şekilde işlemesini sağlar. Projemizde "Feature Çıkarma" train işleminin başında yapılmamış olup hazır çıkarılmış featurelar kullanılmıştır. Bu işlem, "Singe Video Prediction" kısmında yapılmıştır.



Şekil 6.2 VGGish & I3D Feature Çıkarma [4]

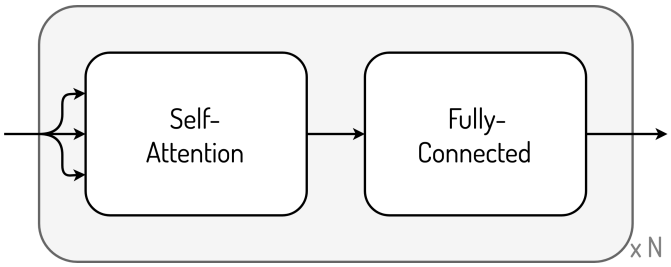
Şekil 6.2'de de görüldüğü gibi model, görsel (video) ve işitsel (ses) verileri ayrı ayrı işler.

6.2 Ses ve Görüntü İşleme



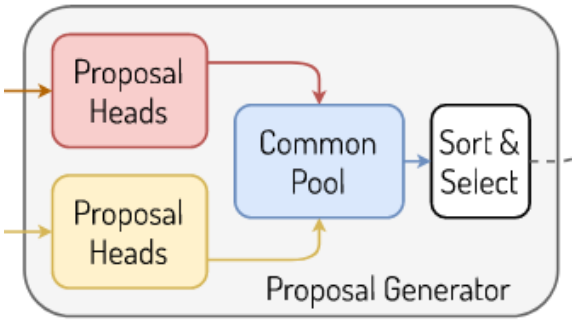
Şekil 6.3 Bimodal Encoder [4]

Şekil 6.3'te görülen Encoder; iki farklı özellik türünü (ses ve görsel) girdi olarak alır. Daha sonrasında bu girdileri bir dizi katmandan geçirir. Attention mekanizmasında videonun hem ses hem de görsel özelliklerine odaklanır. Daha sonrasında bu özellikler birbiriyle ilişkilendirilir ve birleştirilir. Böylece görsel kalıtımlı ses ve ses kalıtımlı görsel özellikler çıkarılır. Fully connected katmanında ise özelliklerin boyutları dönüştürülür ve bu verilerin "Decoder"a aktarılması için uygun hale getirilir.



Şekil 6.4 Encoder [4]

Şekil6.4'te ise Encoder'ın tek modal versiyonu gözükmektedir. Tek modal versiyonuna Vanilla Transformer's Encoder denmektedir. Vanilla Transformer's Encoder'da sadece görsel özellik üzerinden caption oluşturulur. Ses kullanılmaz.

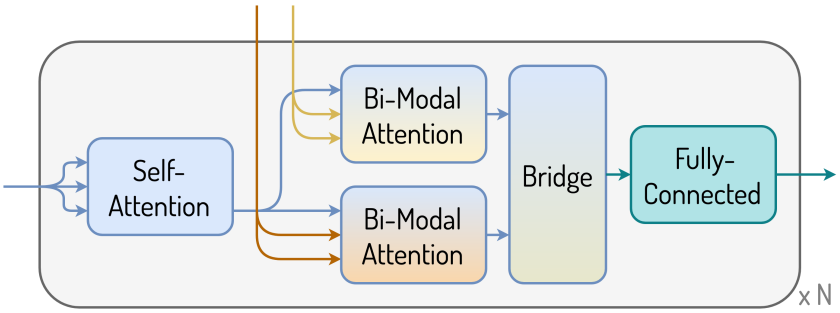


Şekil 6.5 Proposal Generator [4]

Şekil 6.5'te Multi-headed Proposal Generator modülü, videodaki olayları tespit etmek için kullanılır. Video içindeki potansiyel olayların zaman aralıklarını ve önemini tahmin eder. Bu işlemi "Encoder"dan aldığı görsel kalıtımlı ses ve ses kalıtımlı görsel özellikleri alarak yapar. Daha sonrasında iki özellikten çıkardığı tahminleri ortak bir havuzda birleştirir. Bu havuzda her iki mod için de öneriler bulunmaktadır. Daha sonrasında bu öneriler güven skorlarına göre sıralanır ve en yüksek ilk 100 öneri seçilir. Son olarak bu öneriler "caption" modülüne geri gönderilir ve videodaki önemli olayları tanımlamak için kullanılır.

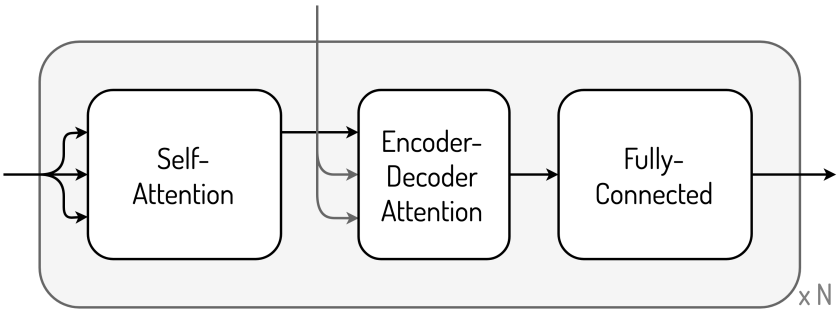
Genel olarak "Proposal Generator" modülü, hangi video bölümlerinde caption olacağını belirlemesine yardımcı olur.

6.3 Caption Oluşturma



Şekil 6.6 Bimodal Decoder [4]

Şekil 6.6'da görülen Bi-modal Decoder, "Encoder" tarafından üretilen birleştirilmiş görsel ve ses özelliklerini işler ve daha önceden yazılmış "caption"larla birlikte değerlendirir. "Decoder" bu bilgileri kullanarak cümlenin bir sonraki kelimesini nasıl üreteceğini öğrenir. [14]



Şekil 6.7 Decoder [4]

Şekil 6.7'de ise Decoder'ın tek modal versiyonu olan Vanilla Transformer's Decoder gözükmektedir.

Böylelikle "Caption" ve "Proposal" olmak üzere 2 adet model eğitilmiş olur.

6.4 Single Video Prediction

Single video prediction, bir videon için açıklamalar üretme işlemidir. Bir mp4 dosyasından, oluşturulmuş "caption" ve "proposal" modelleriyle cümleler oluşturulur. Bu işlem aşağıdaki aşamalardan oluşmaktadır.

- Öncelikle videonun görsel ve ses öğeleri analiz edilir.
- Sonrasında CNN tekniğiyle videodan görsel ve ses özellikleri çıkarılır ve bir dosyaya kaydedilir.

- Kaydedilen özellik dosyası modele verilerek videonun içeriğiyle belirli sahneler için metinler oluşturulur.
- Son olarak oluşturulan bu metinlerin anlamlı olması kontrol edilir ve video içeriğiyle uyumluluğu karşılaştırılır.

7
Deneysel Sonuçlar

Bu bölümde projemizde kullanılan veri kümesi, kullanılan teknikler ve farklı evaluation kriterlerinin projenin başarısına ne derecede etki ettiği anlatılmıştır.

Değiştirlen metrikler ile yapılan farklı deneylerin ne tür farklı sonuçlar verdikleri, eğitilen modellerle yapılan testlerin çeşitli metrikler ile nitelendirilmesi vasıtasıyla yapılmıştır.

Sonuçlara etkisini gözlemlemek için değiştirdiğimiz en büyük etkenler veri kümesi boyutu ve model eğitiminde kullanılan özelliklerin türleridir (görsel veya işitsel).

7.1 Veri Kümesi

Kullandığımız veri kümesi yaklaşık 20.000 video içeren ActivityNet Captions veri kümesidir. Bu veri kümesi kaynaklarımız için çok yüksek boyutlu olup eğitim süresinin çok uzayacağı ve bu süreçte hataların artabileceğinden dolayı veri kümesi küçültülmüştür. Farklı boyutlardaki verilerin sonuçlara etkisini gözlemlemek amacıyla kaynaklarımızın işleyebileceği 2 farklı boyutta veri kümeleri oluşturulmuştur.

İki veri kümesinde de train/toplam veri oranı %75 seviyesinde tutulmuş. 2.000 ve 4.000 videoluk veri kümeleri oluşturulmuştur. 1500 adet train, 500 adet validation ve 3000 adet train, 1000 adet validationdan veri kümelerimizden raporun ileri bölümlerinde "2k" ve "4k" isimlendirmeleri ile bahsedilmiştir.

Veri kümesi validation ve train olmak üzere 2 bölüme ayrılmış, test için ayrı bir veri kullanılmamıştır. Validation kısmı val1 ve val2 isminde 2 bölüme ayrılmış ve earlystop mekanizmasının daha doğru çalışması için 2 küme üzerinde validation işlemi yapılıp sonuçlar 2 kümeden gelen sonuçların ortalamasına göre hesaplanmıştır. Evaluation kısmında alınan sonuçlar da bu validation kümeleri ile hesaplanmıştır.

7.2 Özelliklerin Kullanımı

Projede sonuçların değerlendirilmesi noktasında en büyük hedef görsel ve işitsel özelliklerin modelin başarım sonucuna etkisinin gözlemlenmesidir.

Modellerin eğitiminde önceden eğitilmiş üç boyutlu I3D modeli ile çıkarılmış görsel özellikler ile önceden eğitilmiş VGGish modeli ile çıkarılmış işitsel özellikler kullanılmıştır. Sadece görsel özellikler, sadece işitsel özellikler ve hem görsel hem işitsel özellikler kullanılarak her veri kümesi için üç farklı model eğitilmiş. Farklı boyutlardaki iki farklı veri kümesi üzerinde yaptığımız çalışmalar sonucu toplam altı farklı model eğitilmiştir.

7.3 Sonuçlar

Farklı veri kümesi boyutu ve farklı türde özelliklerin kullanımı ile eğitilen modellerin doğruluklarının çeşitli metriklerle karşılaştırılması Tablo 7.1'de verilmiştir.

Tablo 7.1 Veri Kümesi Boyutu ve Özellik Çeşitlerine Göre Sonuçlar

Veri Kümesi	Kullanılan Özellik	BLEU_3	BLEU_4	METEOR	ROUGE-L	CIDEr
2k video	Audio	1.9981	0.7491	7.3333	17.4502	19.3073
	Video	2.3170	1.0875	9.1743	18.7456	23.8733
	Audio&Video	3.4680	1.2396	8.8500	19.6310	28.5315
4k video	Audio	2.8560	1.1180	7.0387	16.6512	24.3962
	Video	2.4567	1.0219	8.3700	17.9581	25.3352
	Audio&Video	3.4613	1.4402	8.5688	18.2218	27.3613

8
Performans Analizi

Bu bölümde Tablo 7.1'deki sonuçların analizi yapılmıştır.

8.1 Veri Kümesi

ActivtyNet Captioins veri kümesinden 2k ve 4k olmak üzere iki ayrı veri kümesi elde etmiştik. Bu veri kümelerinin BLEU, ROUGE, METEOR ve CIDEr metrikleri ile ölçülmüş sonuçları Tablo 7.1'de gösterilmiştir. Bu sonuçlara göre audio, video ve audio&video özellikleriyle eğitilen modellerde BLEU ve CIDEr metriklerine göre daha fazla veri kullanımında genellikle daha iyi performans alındığı gözlemlenmiştir. Ancak beklenmedik şekilde METEOR ve ROUGE-L metriklerinde veri sayısı arttırıldığında çok az bir oranda da olsa performans düşüşü tespit edilmiştir.

Bu sonuçlar, veri kümesi genişletilmesinde genel bir artış beklentisinin her zaman geçerli olmayacağını gösterir.

8.2 Özelliklerin Kullanımı

VGGish modeli ile elde edilen işitsel özellikler ve I3D modeli ile elde edilen görsel özellikler olmak üzere iki farklı tür özellik kullanılmıştır. Bu özellikler tek başlarına veya birlikte kullanılarak üç farklı model elde edilmiştir.

İşitsel (audio) özelliklerin tek başına kullanımında genel performansın daha düşük olduğu gözlemlenirken görsel (video) özelliklerin kullanımında genelde daha iyi bir performans alınmıştır. En yüksek performans ise her iki özelliğin birlikte kullanılarak eğitildiği modellere alınmıştır.

Bu sonuçlarda daha derinlemesinde analiz yapıldığında bazı model ve metriklere göre bu genel performansın dışında kalan durumlar tespit edilmiştir.

8.3 Sonuç

Sonuç olarak elde edilen bulgularda özelliklerin kullanım tercihinin performansa etkisinin genel bir çizgi içserisinde olabileceği ancak veri kümesi boyutunun performansa etkisi gözlemlenirken daha karmaşık bir ilişki olduğu söylenebilir. Daha büyük veri kümeleri üzerinde yapılan çalışmalar daha sağlıklı sonuçlar, bu ilişkilerin daha iyi analiz edilip anlaşılmasına yardımcı olabilir.

9
Sonuç

Bu bölümde elde edilen deneysel sonuçlar özetlenmiş ve gelecek çalışmalar hakkında çeşitli düşünceler dile getirilmiştir.

9.1 Veri Kümesi Büyüklüğünün Etkisi

Yapılan deneylerde veri kümesinin büyüklüğü artıkça BLEU ve CIDEr metriklerinde belirgin bir performans artışı tespit edilmiştir.. Ancak ROUGE-L ve METEOR metriklerinde minimal düzeyde bir düşüş gözlemlenmiştir. Bu durum veri kümesi boyutunun performans ile doğru orantılı olabileceğini ancak sonuçların metriklere göre değişebileceğini göstermektedir.

9.2 Özellik Çeşitlerinin Performansa Etkisi

Görsel ve işitsel özelliklerin ayrı ve birlikte kullanımlarıyla eğitilen modeller üzerine yapılan analizde genellikle en düşük performansın audio modellerinde ve en yüksek performansın audio&video modellerinde alındığı gözlemlenmiştir. Bu durum görsel özelliklerin videolardaki olayları anlamada daha büyük bir öneme sahip olduğunu, işitsel özelliklerin de buna ekstra katkı sağlayabildiğini gösterir.

Daha derinlemesine bir analizde ise bazı modellerin bazı metriklere göre her zaman bu genellemeye orantılı bir performans vermediği gözlemlenmiştir. . Örneğin, 4k video modeli BLEU metriğinde audio modeline göre düşük bir performansa sahipken, 2k audio&video modeli METEOR metriğinde video modeline göre düşük bir performansa sahiptir.

9.3 Kullanılan Metrikler Hakkında

Raporun "Sistem Tasarımı" bölümünde kullanılan metrikler ve özellikleri hakkında bilgi verilmiştir. Bu metriklerin bazı durumlarda birbirlerine paralel performans gösterip bazı durumlarda ise aksiliklerle karşılaşıldığı da bir gerçektir.

Genellikle BLEU ve CIDEr metriklerinin birbirlerine paralel, METEOR ve ROUGE-L metriklerinin de kendi arasında paralel olduğu ve beklenmedik sonuçlarda çoğunlukla METEOR ile ROUGE-L metriklerinin beklenin aksine sonuç verdiği gözlemlenmiştir. Bu durum metriklerin hesaplanma şeklinin sonuçların hesaplanmasına olan etksini net bir şekilde gösterir.

9.4 Single Video Prediction

9.4.1 Video Özelliklerinin İşlenmesi

Tek bir videodan özellik çıkarılması işleminde video hakkında elde edilen captionların video ile tutarlı olması dolayısıyla VGGish ve I3D modelleri kullanılarak işitsel ve görsel özelliklerin başarı ile çıkarıldığı söylenebilir.

9.4.2 Model Tahmini

Elde edilen özellikler ile eğitilen model video içeriği ile uyumlu tahminler üretebilmiştir. Ancak bir sahnede aynı captionı tekrar tekrar vermek, sahneleri kendi içinde fazla bölüntülemek gibi bazı problemler gözlemlenmiş ve üretilen captionlar üzerinde çeşitli filtreleme teknikleri uygulanarak bu problemler minimuma indirgenmeye çalışılmıştır.

İki tenis oyuncusunun maçını içeren bir videonun modele verilmesi ile alınan sonuç Şekil 9.1'de görüldüğü gibidir.


```
{'start': 0.3, 'end': 10.1, 'sentence': 'A man is seen standing in a court holding a tennis ball'}
{'start': 0.0, 'end': 3.8, 'sentence': 'A person in a white shirt is standing in a court'}
{'start': 0.0, 'end': 1.4, 'sentence': 'A person is seen standing in a court holding a tennis ball'}
{'start': 2.4, 'end': 4.2, 'sentence': 'A man in a white shirt is standing in a court'}
{'start': 6.8, 'end': 11.6, 'sentence': 'A man in a white shirt throws a ball onto the field'}
{'start': 3.4, 'end': 5.2, 'sentence': 'A player in a white shirt is playing tennis'}
{'start': 3.7, 'end': 7.2, 'sentence': 'A man in a white shirt is playing tennis'}
```


Şekil 9.1 Sonuç

9.4.3 Sonuç

Projenin single video prediction bölümünde denenen videolar ve alınan sonuçlar incelendiğinde üretilen tahminlerin içeriğinin çoğunlukla video içeriği ile uyumlu olduğu gözlemlenmiştir. Lakin cümlelerde dil bilgisi hataları, captionların zamanlamalarında yanlışlıklar veya tekrar eden captionlar gibi problemlerle karşılaşılmıştır. Bu problemlerin olası çözümlerine ilişkin öneriler "Gelecek Çalışmalar" bölümünde tartışılmıştır.

9.5 Gelecek Çalışmalar

9.5.1 Veri Boyutu ve Kalitesi

Gelecek çalışmalarda aynı kaynaktan oluşturulmuş daha fazla video içeren veri kümeleri veya farklı kaynaklardan alınmış yeni videolar kullanılabilir. Daha kaliteli içerikler bulmak modelin videoları anlama yeteneğini arttırabilirken daha büyük veri kümeleri nadir görünebilecek durumlar için de geçerli tahminler yapılabilmesine olanak sağlar.

9.5.2 Yeni Özelliklerin Eklenmesi

İşitsel ve görsel özelliklere ek olarak farklı özellikler de içeren modellerin geliştirilmesi düşünülebilir. Örneğin konuşma verisinin de bulunduğu bir veri setiyle Speech Recognition (konuşma algılama) yeteneğine sahip bir model eğitilebilir.

9.5.3 Hiperparametre Ayarları

Model performansını arttırmak için çeşitli optimizasyon çalışmaları ve hiperparametre ayarları üzerinde çalışmalar yapılabilir.

9.5.4 Tahminlerin İncelenip Doğrulanması

Modelin ürettiği captionların doğrulanması aşamasında kullanıcılardan da geri bildirim alınması ve ranking benzeri bir uygulama şekliyle modelin ne tür captionlar üretmesi gerektiği konusunda daha bilgili olması sağlanabilir.

Bu öneriler projemizin devamında yapılacak çalışmalar için gelecekteki araştırmacılara rehberlik edebilir.

Referanslar

[1] A. C. M. S. Guido, "Introduction to machine learning with python: A guide for data scientists," 2016.

[2] IBM. "What is deep learning?" (), [Online]. Available: `https://www.ibm.com/topics/deep-learning` (visited on 01/17/2024).

[3] G. Chowdhury, "Natural language processing. annual review of information science and technology," 2003.

[4] V. Iashin and E. Rahtu, "A better use of audio-visual cues: Dense video captioning with bi-modal transformer," in *British Machine Vision Conference (BMVC)*, 2020.

[5] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Densecaptioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[6] B. Wang, L. Ma, W. Zhang, and W. Liu, "Reconstruction network for video captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7622–7631.

[7] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Jul. 2002, pp. 311–318.

[9] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Jun. 2005, pp. 65–72.

[10] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," Jun. 2004.

[11] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.

[12] M. Nguyen. "Illustrated guide to lstm's and gru's: A step by step explanation." (), [Online]. Available: `https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21` (visited on 01/18/2024).

[13] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[14] V. Iashin and E. Rahtu, “Multi-modal dense video captioning,” in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2020.

Özgeçmiş

BİRİNCİ ÜYE

İsim-Soyisim: Yusuf Enes KURT
Doğum Tarihi ve Yeri: 24.11.2000, İstanbul
E-mail: enes.kurt1@std.yildiz.edu.tr
Telefon: 0551 112 98 59
Staj Tecrübeleri:

İKİNCİ ÜYE

İsim-Soyisim: Muhammed Ali LALE
Doğum Tarihi ve Yeri: 30.11.2002, Yozgat
E-mail: ali.lale@std.yildiz.edu.tr
Telefon: 0543 283 64 65
Staj Tecrübeleri:

Proje Sistem Bilgileri

Sistem ve Yazılım: Windows İşletim Sistemi, Python
Gerekli RAM: 12GB
Gerekli Disk: 20GB