Mühendislik Bilimleri ve Tasarım Dergisi
8(5), 271-289, 2020
e-ISSN: 1308-6693
Araştırma Makalesi

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.830587
Research Article

Özel Sayı: Uluslararası Mühendislikte Yapay Zeka ve Uygulamalı Matematik Konferansı (UMYMK 2020)
Special Issue: International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2020)

VİDEO ETİKETLEME UYGULAMALARINDA DERİN ÖĞRENME YAKLAŞIMLARININ KULLANILMASI ÜZERİNE KAPSAMLI BİR İNCELEME

Özlem Alpay[1*], M. Ali Akcayol[2]

[1] Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye
[2] Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye

Anahtar Kelimeler
Derin Öğrenme,
Video Etiketleme,
EAS

Öz

Video etiketleme, otomatik bir şekilde videolar için etiket oluşturma olarak tanımlanmaktadır. Hem bilgisayar görmesi hem de doğal dil yaklaşımlarını birlikte içerdiği için gittikçe ilgi çeken bir alan olmaktadır İfadeleri doğal dilde üretip ve onları görüntü çerçeveleri ile birleştirmek zorlu bir süreçtir. Bu sorunu çözmek için çeşitli yaklaşımlar geliştirilmiştir. Bu çalışmada, video etiketleme araştırmalarındaki gelişmeler hakkında bir literatür çalışması sunulmuştur. İncelenen çalışmalar kullanılan yöntemlere göre farklı kategorilerde incelenmiştir. Yöntemler özetlenmiş, güçlü ve sınırlı yönleri analiz edilmiştir. Derin öğrenme, bu konuda kullanılan en yaygın yöntemlerden biridir. Video etiketleme sistemlerinde derin öğrenme yaklaşımlarının uygulanabilirliği üzerine araştırmalar yapılmıştır. Bu konuda kullanılan veri setleri, performans değerlendirme kriterleri karşılaştırılarak analiz edilmiştir. Derin öğrenme yöntemlerindeki gelişmeler video etiketleme konusunda yeni yaklaşımlar sağlamıştır. Video etiketleme konusunda yapılan çalışmalarda derin öğrenme yöntemlerinin kullanılması ile başarılı sonuçlar elde edilmiştir

A COMPREHENSIVE REVIEW ON USING OF DEEP LEARNING APPROACHES IN VIDEO CAPTIONING APPLICATIONS

Keywords
Deep Learning,
Video Captioning,
CNN

Abstract

Video captioning is defined as creating captions for videos by automatically. It is an area of increasing interest as it includes both computer vision and natural language approaches. Producing expressions in natural language and combining them with image frames is a difficult process. Many approaches have been developed to solve this problem. In this study, a literature study about the developments in video captioning research is presented. The studies examined are examined in different categories according to the methods used. The methods are summarized and their strengths and limitations are analyzed. The methods are summarized and their strengths and limitations are analyzed. Deep learning is one of the most common methods used in this regard. Research has been conducted on the applicability of deep learning approaches in video labeling systems. Used datasets and performance evaluation criteria are analyzed. The developments in deep learning methods have provided new approaches to video captioning. The successful results have been observed with the use of deep learning methods in studies on video captioning.

Alıntı / Cite

Alpay, Özlem., Akcayol, M. Ali., (2020). Video Etiketleme Uygulamalarında Derin Öğrenme Yaklaşımlarının Kullanılması Üzerine Kapsamlı Bir İnceleme, Mühendislik Bilimleri ve Tasarım Dergisi, 8(5), 270-288.

Yazar Kimliği / Author ID (ORCID Number)
Ö. Alpay, 0000-0002-5432-4102
M.A. Akcayol, 0000-0002-6615-1237

Makale Süreci / Article Process	
Başvuru Tarihi / Submission Date	24.11.2020
Revizyon Tarihi / Revision Date	26.12.2020
Kabul Tarihi / Accepted Date	27.12.2020
Yayım Tarihi / Published Date	29.12.2020

* İlgili yazar / Corresponding author: sozlemalpay@gmail.com

1. Giriş (Introduction)

Artan video kullanımı miktarı ve sayısı ile videolara, video yazısı ekleme ve erişilebilirliği problemi ortaya çıkmaktadır. Bununla birlikte görsel içerik anlama, videoya alt yazı ekleme ve içerik özetlemesi için teknolojik gereksinimler artmaktadır. Geniş miktarda metnin kullanılabilirliğini gerçekleştiren Doğal Dil İşleme (DDİ-Natural Language Processing - NLP) teknolojisi, bu konuda önemli bir role sahiptir. Görsel içerikleri doğru ve özlü metinsel açıklamalara dönüştürebilen tekniklerin bu bilgilerin düzenlenmesinde bir araçtır. Bu konu altındaki teknikler "Görselden Metne" olarak adlandırılmaktadır.

Video anlama ve doğal dil cümlesi oluşturmadaki temel zorluklar şu şekildedir;

- Video içeriğinin ince hareket detaylarını ve ayrıca farklı nesnelerin etkileşimlerini anlamak,
- Video ve dil alanı arasındaki ilişkiyi öğrenmek,
- Videoda tanımlanan etkinliklerin sıralanması (Li ve diğ., 2019)

Nesne tanıma (Wu ve diğ., 2020), (Yuan ve diğ., 2019), özniteliklerin sınıflandırılması (Trabelsi ve diğ., 2019), (Zhao and Li 2017), eylemlerin sınıflandırılması (Jegham ve diğ., 2020), (Liu ve diğ., 2020) görüntülerin sınıflandırılması (Smirnov ve diğ., 2014) ve sahne tanıma (Zhou ve diğ., 2014), (Nan ve diğ., 2019) gibi çeşitli bilgisayar görme problemlerinde büyük ilerleme kaydedilmiş olmasına rağmen bir bilgisayarın kendisine iletilen bir görüntüyü otomatik olarak tanımlamak için insan benzeri bir cümle oluşturabilmek nispeten yeni bir alandır.

Video etiketleme, doğal dil kullanarak video içeriğini otomatik olarak tanımlamadır. Video etiketleme, metin tabanlı video alma, görme engelli kullanıcılar için erişilebilirlik ve multimedya önerisi gibi çeşitli uygulamalara sahiptir. Videolar için video yazısı kullanmadan önce, resimler için resim yazısı kullanılıp tek bir cümle ile görüntü içeriği açıklanmaktadır. Derin evrişim sinir ağının hızlı gelişmesi ile resim yazısı konusunda önemli ilerlemeler olmuştur. Daha sonra bu yaklaşım, büyük bir olayı içeren kısa videolara uygulanmıştır.

Video etiketleme de temel olarak iki sorun bulunmaktadır. Birincisi, framelerdeki görsel kavramların doğru bir şekilde analiz edilmesidir. Görsel öğeleri tespit edebilmek için yaygın olarak kullanılan VGG (Yang ve diğ., 2018) veya GoogleNet (Tang ve diğ., 2017) gibi Evrişimsel Sinir Ağı (ESA - Convolutional Neural Network-CNN) modelleridir. Oluşturulan ağ yapısı genellikle büyük ölçekli bir görüntü sınıflandırma veri seti üzerinde önceden eğitilir böylece giriş görüntüsünden birçok görsel kavram keşfedebilir. Ancak, videoda neler olduğunu anlamak istediğimizde bazı kavramlar diğerlerinden daha önemlidir. Geleneksel ESA modelinde, farklı görsel kavramların önemi belirlenememektedir. Bu nedenle, doğru ve özlü açıklamalar üretilmesine yardımcı olabilecek baskın görsel kavramları keşfetmek için bir yöntem tasarlanmalıdır. Bir videoda, bitişik çerçeveler arasında çok fazla bilgi vardır, bu nedenle kompakt bir dizi bilgilendirici çerçeve bulmak ve bunlara dikkat etmek daha iyidir. İkincisi de doğal dil işleme ile elde edilen görüntüler arasındaki ilişkiyi kurabilmektir. Video karelerine, açıklayıcı etiketleri otomatik olarak ekleyebilmek için uygun bir algoritma tasarlanmalıdır. En çok kullanılan bir yöntem, her sözcük üretilirken video karelerine ağırlık atamak için sinir ağlarını kullanan geçici odak yaklaşımıdır (Yao ve diğ.,2015). Görüntü karelerinin ağırlıkları hesaplanırken, her bir karenin tüm ham pikselleri ile sinir ağı beslenir, bu videoda çok sayıda ilgisiz veya dağınık arka plan bilgisi olduğunda iyi bir performans sağlamamaktadır. Çünkü bu tür arka plan bilgileri istediğimiz bilgileri içermemektedir ve yanlış görüntü karelerinin ağırlıklarına ve açıklamalarına neden olmaktadır (Chen ve diğ., 2018).

Problemin zorluğuna rağmen, video etiket yazısı oluşturmaya yönelik çeşitli çalışmalar yapılmıştır. Temel olarak makine çevirisindeki son gelişmelerden esinlenmiştir. Genel olarak, derin görsel temsiller elde etmek için 2B ve / veya 3B ESA'lardan faydalanırlar ve cümle veya kelime üretmek için Uzun Kısa Süreli Bellek (UKSB - Long Short Term Memory - LSTM) kullanılır (Wu ve diğ., 2016).

Video etiketleme problemi, çoğu zaman yapay makine çevrimi problemine benzetilebilir. Video etiketleme otomatik çeviri problemi olarak ele alındığında girdi cümlesine karşılık gelen yapı görüntü dizisi olarak düşünülebilir. Bu sebeple video etiketleme yaklaşımlarında kodlayıcı- çözücü (encoder - decoder) yapısına sahip diziden diziye (sequence-to-sequence) modeller yaygın olarak kullanılmaktadır. Burada çoğu zaman, ESA ve Tekrarlayan Sinir Ağları (TSA - Recurrent Neural Network - RNN) görüntüleri kodlamak için birlikte kullanılmakta ve açıklama üretme kısmında ise ayrı bir yinelemeli sinir ağı devreye girerek; doğal dilden bir cümle oluşturulmaktadır (Li ve diğ., 2019).

Bu makale aşağıdaki gibi düzenlenmiştir. Giriş bölümünde video etiketleme kavramı ele alınmıştır. Bölüm 2'de bu konuda yapılan çalışmalar karşılaştırmalı bir şekilde açıklanmıştır. Bölüm 3'de video etiketlemede kullanılan yaklaşımlar incelenmiştir. Bölüm 4'te video etiketlemede kullanılan derin öğrenme mimarileri, bölüm 5'te

kullanılan veri setleri ve bölüm 6'da değerlendirme ölçütleri açıklanmıştır. Son bölümde bu konuda yapılan çalışmalar detaylı bir şekilde incelenerek önerilerde bulunulmuştur.

2. Literatür Araştırması (Literature Survey)

Literatürde ve uygulamada derin öğrenmenin kullanıldığı çok sayıda çalışma yapılmıştır. Doğal dil işleme, görüntü ve video işleme, biyomedikal sinyal ve görüntü işleme, nesne tanıma, robotik, kimya, reklam, finans, arama motorları, otonom araç sistemleri gibi çok çeşitli konularda derin öğrenme uygulamaları geliştirilmektedir (Şeker ve diğ., 2017). Bu bölümde, derin öğrenme yöntemlerinin etiketleme sistemlerinde kullanıldığı çalışmalar analiz edilmiş ve etiketleme sürecinde kullanılan yöntemler kapsamlı bir şekilde incelenmiştir. Çalışmaların literatürü Tablo 1'de özetlenmiştir.

Tablo 1. Literatür Özeti (Literature Survey)

Referans	İçerik	Metodoloji	Veri Seti	Ölçütler
Özer ve diğ.,	Gerçek zamanlı görüntüleri analiz ederek etiket oluşturma	EAS ve UKSB	MSVD	BLEU 4 - ROUGE-L CIDEr - METEOR
Baraldi ve diğ.	Videolardaki süreksizlik noktalarını tanımlayarak etiket oluşturma	UKSB	M - VAD MPII – MD M - VAD	BLEU - ROUGE, METEOR - CIDEr
Wang ve diğ.,	Videonun sıralı karelerini bir zaman çizelgesinde uzamsal-zamansal şekilde göstererek etiketleme	İki katmanlı Geçitli Tekrarlayan Birim (GTB - Gated recurrent unit - GRU	MSVD	BLEU 4 - METEOR CIDER
Li ve diğ.	Video etiketleme	UKSB	MSVD MSR - VTT	BLEU - METEOR
Wu ve diğ.	Video etiketleme	EAS ve TSA	MSVD MSRV - TT	BLEU - METEOR CIDEr
Ding ve diğ	Video bölümleme ve anlamsal metin oluşturma	UKSB	VideoSet	BLEU - METEOR ROUGE
Xiao ve Shi	Video etiket oluşturma	EAS ve UKSB	MSVD MPII - MD	BLEU – METEOR CIDEr ve ROUGE
Gao ve diğ	Doğal cümleleri videolara aktararak etiket oluşturma	UKSB	MSVD MSR - VTT	BLEU- METEOR.
Pan ve diğ.	Videoların zamansal bilgisinden yararlanarak etiket oluşturma	Hiyerarşik Tekrarlayan Nöral Kodlayıcı (HTNK-Hierarchical Recurrent Neural Encoder - HRNE)	MSVD M - VAD	BLEU - METEOR, CIDEr - ROUGE
Venugopalan ve diğ.	Video karesi dizisini bir kelime dizisiyle ilişkilendirerek etiket oluşturma	UKSB	MSVD MPII - MD M-VAD	METEOR
Yao ve diğ	Zamansal yapıyı kullanarak zaman tabanlı etiket oluşturma	EAS ve UKSB	Youtube2Text, DVS	BLEU - METEOR
Xu ve diğ.	Görüntülerin içeriğini otomatik olarak tanımlama	EAS ve UKSB	Flickr8k, Flickr30k MS COCO	BLEU - METEOR

Bir görüntünün içeriğini otomatik olarak tanımlamak yapay zekâda bilgisayar vizyonunu ve doğal dil işlemeyi birbirine bağlayan temel bir sorundur. Vinyals ve diğ. (2015) bilgisayar görmesi ve makine çevirisindeki son gelişmeleri birleştiren ve bir görüntüyü tanımlayan doğal cümleler oluşturma da kullanılabilen derin tekrarlayan bir mimariye dayanan üretken bir model sunmuşlardır. Model, görüntülerden açıklamalar üretmek için sinirsel ve olasılıklı bir çerçeve kullanmaktadır. Bu modeller, değişken uzunluk girişini sabit boyutlu bir vektöre kodlayan ve bu gösterimi istenen çıkış cümlesinde "kodu çözmek" için tekrarlayan bir sinir ağı kullanır. Görüntüye verilen doğru tanım olasılığını doğrudan en üst düzeye çıkarmak için Eşitlik (1) kullanılmaktadır.

$$\theta^* = \text{argmax}_\theta \sum_{(I,S)} \log_{P(S|I;\theta)} \quad (1)$$

Burada θ modelin parametreleri, I bir görüntü ve S doğru transkripsiyonu. S herhangi bir cümleyi temsil ettiğinden, uzunluğu sınırsızdır.

$$\log_{p(S|I)} = \sum_{t=0}^{N} \log_{P(S_t|\, I,S0,\ldots S_{t-1})} \quad (2)$$

You ve diğ. (2016) semantik odak yaklaşımını kullanarak yukarıdan aşağıya, aşağıdan yukarıya yaklaşımları birleştiren yeni bir yaklaşım önermiştir. Yukarıdan aşağıya yaklaşıma dayanarak, evrişimli bir sinir ağı görsel özellikleri ayıklar ve ayrıca görsel kavramları (nesneler, bölgeler, özellikler vb.) tespit etmektedir. Aşağıdan yukarıya yaklaşımı, birkaç aday konsepti için TSA yinelemelerine göre dikkat ağırlıklarını değiştirmektedir. Anlamsal odak yaklaşımı, TSA kullanılarak görüntünün açıklamasını üretmek için hem görüntü niteliklerini hem de görsel kavramları birleştirmek için önerilmiştir.

Nabati ve Behrad (2020), UKSB ağları kullanarak video etiketleme için yeni güçlendirilmiş paralel bir mimari önermişlerdir. Önerilen mimari iki UKSB katmanından ve bir kelime seçim modülünden oluşur. Çoğu dizi öğrenme modeli gibi, önerilen şema kodlama ve kod çözmeyi içeren iki aşamadan oluşmaktadır. Üst UKSB katmanı, kodlama aşamasında bir video klibin görsel bilgilerini kodlama özelliğine sahip bir UKSB ağından oluşur. Bu UKSB katmanı, hem kare tanımlayıcılarını hem de zamansal ve kareler arası bağımlılıkları dikkate almak için sıralı modda derin bir ESA tarafından çıkarılan yüksek düzeyli kare özelliklerinden yararlanmaktadır Bir video klibin tüm örneklenmiş kare özelliklerini üst UKSB katmanına besledikten sonra, kod çözme aşaması sırasında videonun görsel içeriğine karşılık gelen olası kelimeleri üretmek için Artırılmış ve Paralel (AP - Boost Paralel - BP) UKSB'ler (AP - UKSB'ler) kullanılır. Önerilen mimaride ikinci katman olarak AP -UKSB'ler, bir giriş videosunun olası kelimelerini eşzamanlı olarak üreten birkaç paralel UKSB 'den oluşur. Eğitim aşaması sırasında, bu UKSB'ler, belirli bir AdaBoost algoritması kullanılarak AP- UKSB ağına yinelemeli olarak eklenir. AP- UKSB leri önerilen mimarinin üçüncü katmanı olarak Maksimum Olasılık Kelime Seçimi (MOKS - Maximum Probability Word Selection (MPWS) modülü takip eder. MOKS modülü, AP - UKSB'ler tarafından üretilen en olası kelimeyi seçmek için maksimum olasılık ölçütünü kullanmaktadır. Belirli bir AdaBoost algoritması kullanılarak düzenlenen ve eğitilen birkaç tamamlayıcı UKSB ağından oluşan bir komite makinesidir. Bu topolojide, görsel bilgileri kodlayan üst UKSB katmanı tüm AP- UKSB'ler arasında paylaşılır. Ancak cümleleri deşifre eden ikinci katmanda; oluşturulan cümlelerin verimliliğini artırmak için birçok güçlendirilmiş UKSB kullanılmıştır.

Test aşaması sırasında, AP - UKSB'lerin çıkışları aynı anda maksimum olasılık kriteri ve kelime seçim modülü kullanılarak birleştirilmiştir. Önerilen algoritma iyi bilinen iki video altyazı veri kümesiyle test edilmiş ve sonuçlar bazı algoritmalarla karşılaştırılmıştır. AP - UKSB'ler AdaBoost algoritmasına ve eğitim videolarının METEOR puanına göre eğitilmiştir. Yapılan karşılaştırmaların sonuçları, önerilen yaklaşımın üretilen cümlelerin verimliliğini ve değerlendirme puanlarını önemli ölçüde artırdığını ortaya koymuştur. Ayrıca, sonuçlarda AP-UKSB ağında beş UKSB ağının kullanılmasının sırasıyla CIDEr ve METEOR puanlarını sırasıyla yüzde 9,7 ve 2,4 oranında artırdığını göstermiştir.

Xu ve diğ. (2016), görüntülerin içeriğini otomatik olarak tanımlamayı öğrenen odak tabanlı bir model geliştirmişlerdir. Etiket oluşturma yaklaşımına, iki çeşit odak şeklini dâhil etmişlerdir bunlar; “sert” odak yaklaşımı ve “yumuşak” odak yaklaşımıdır.

Yumuşak deterministik odak yaklaşımı, standart geri yayılma yöntemleri ile eğitilebilir. Sert stokastik odak yaklaşımı, yaklaşık bir varyasyon alt sınırını veya eşdeğer şekilde maksimize ederek eğitilebilir. Dikkatin “nereye” ve “neye” odaklandığını görselleştirerek bu modelin sonuçları nasıl kavrayabildiği ve yorumlayabildiği gösterilmiştir. Görselleştirme yoluyla, modelin çıkış sırasında karşılık gelen sözcükleri üretirken bakışlarını dikkat çekici nesneler üzerinde nasıl otomatik olarak öğrenebileceği de gösterilmektedir. Flickr8k, Flickr30k ve MS COCO veri setleri, BLEU ve METEOR metrikleri kullanarak model değerlendirilmiştir. Modelin, Google NIC, Log Bilinear, TSA gibi modellerle karşılaştırıldığında daha başarılı olduğu gözlenmiştir.

Yao ve diğ. (2015), küresel zamansal yapıyı kullanmak için zamansal bir odak yaklaşımı geliştirmişlerdir. Model, görünüm özelliklerini yerel zamansal yapıyı kodlayan eylem özellikleriyle güçlendirilmiştir. Eylem özellikleri, uzaysal-zamansal evrişimli sinir ağından türetilmiştir. Zamansal odak yaklaşımı makine çevirisi bağlamında başarıyla kullanılan yakın zamanda önerilen bir yumuşak hizalama yöntemine dayanmaktadır. Bir açıklama oluştururken, zamansal odak yaklaşımı seçici olarak küçük bir çerçeve alt kümesine odaklanır ve kelime üretecinin yalnızca bu alt kümedeki nesneleri ve / veya faaliyetleri tanımlamasını mümkün kılmaktadır. Kullanılan 3B EAS videonun hem geçici hem de uzamsal olarak yerel hareket tanımlayıcılarından başlar ve hiyerarşik olarak daha soyut eylemle ilgili özellikler elde eder. Bu özellikler, açıklama oluşturucu tarafından kullanılmak üzere videoya gömülü önemli yerel yapıyı koruyup ve vurgulamıştır.

Geçici yapıdan yararlanmak için önerilen yaklaşımlar, video başına birden fazla açıklama içeren 1970 video klipten oluşan Youtube2Text veri kümesi adı verilen en yaygın kullanılan açık alan video açıklaması veri kümesinde değerlendirilmiştir. Ayrıca önerilen yaklaşımlar, 49000 video klip içeren DVD filmlerde açıklayıcı Tanımlayıcı Video Hizmeti (TVH - Description Video Server (DVS) parçalarına dayanan çok daha büyük ve yakın zamanda önerilen bir veri kümesinde test edilmiştir.

Çalışma şu katkıları sağlamaktadır:

1) Yerel uzay-zamansal bilgileri yakalayan yeni bir 3B EAS, TSA kodlayıcı-kod çözücü mimarisinin kullanılabileceği önerilmiştir. Statik çerçeve EAS, TSA video açıklama yöntemlerini kullanarak yerel geçici yapıyı kullanmanın gerçekten önemli olduğu sonucuna ulaşılmıştır.
2) Video boyunca statik çerçevelerin küresel analizi yoluyla elde edilen özelliklerin video tanımlama üretimi için daha etkin kullanılmasına izin verdiği görülmüştür.
3) Küresel ve yerel zamansal bilgileri kullanarak getirilen iyileştirmelerin ücretsiz olduğunu ve hem zamansal odak yaklaşımı hem de 3B EAS birlikte kullanıldığında iyi bir performans olduğu gözlenmiştir.

Gerçek dünya videoları genellikle karmaşık dinamiklere sahiptir ve açık alanlı video açıklamaları üretme yöntemleri zamansal yapıya duyarlı olmalı ve değişken uzunlukta hem girdi (kare sırası) hem de çıktıya (kelime sırası) izin vermelidir.

Venugopalan ve diğ. (2015), çalışmalarında, bu soruna yaklaşmak için videolara etiketler oluşturmak üzere yeni bir uçtan uca sıradan sıraya model geliştirmişlerdir. Bunun için, görüntü altyazısı üretiminde son teknoloji performansı gösteren tekrarlayan sinir ağlarından, özellikle UKSB'lerden yararlanılmıştır. UKSB modeli, video-cümle çiftleri üzerinde eğitilmiştir ve video klipteki olayın bir açıklamasını oluşturmak için bir video karesi dizisini bir kelime dizisiyle ilişkilendirmeyi öğrenmiştir. Geliştirilen model, doğal olarak çerçeve dizisinin zamansal yapısını ve üretilen cümlelerin dizi modelini, yani bir dil modelini öğrenebilmektedir.

Pan ve diğ. (2016), videoların zamansal bilgisinden yararlanarak HTNK adlı yeni bir yaklaşım geliştirmişlerdir. HTNK, giriş bilgi akışının uzunluğunu azaltarak ve art arda çoklu girişleri daha yüksek bir seviyede birleştirerek video zamansal yapısını daha uzun bir aralıkta verimli bir şekilde kullanmaktadır. Daha fazla doğrusal olmayanlığa ulaşırken hesaplama işlemleri önemli ölçüde azalmaktadır. HTNK, farklı ayrıntı düzeylerine sahip kare parçaları arasındaki zamansal geçişleri ortaya çıkararak, çerçeveler arasındaki zamansal geçişleri ve segmentler arasındaki geçişleri modellemektedir. Model, zamansal bilgilerin çok önemli bir rol oynadığı video altyazısına uygulanmıştır. Deneylerde, MSVD ve M-VAD veri setleri kullanılmıştır. Özellikle, giriş olarak yalnızca Red – Green - Blue (RGB) akışı olan tek bir ağ kullanıldığında bile HTNK, RGB ConvNet, 3B ConvNet gibi birden fazla girişi birleştiren tüm yeni sistemlerden daha iyi olduğu kanıtlanmıştır. Yöntemin video altyazı ölçütlerinde en son teknolojiden daha iyi olduğu gözlenmiştir.

Gao ve diğ. (2017), videoları doğal cümlelere aktarmak için anlamsal tutarlılığa sahip odak tabanlı bir UKSB modeli olan a UKSB adlı yeni bir uçtan uca model geliştirmişlerdir. Bu model, videolardaki dikkat çekici yapıları yakalamak için odak tabanlı yaklaşımı UKSB ile bütünleştirerek zengin anlamsal içeriğe sahip cümleler oluşturmak için çok modlu temsiller (yani kelimeler ve görsel içerik) arasındaki ilişkiyi araştırır. İlk olarak, daha anlamlı uzamsal özellikler elde etmek için, GoogleNet'in genişletilmiş bir sürümü olan Inception-v3 sinir ağı temel alınmıştır. Zamansal bilgileri kullanmak için, bu uzamsal 2B EAS özellik vektörlerini kodlamak için tek katmanlı UKSB görsel kodlayıcı sunulmuştur. Bir UKSB kod çözücü, önemli kelimeleri üretmek için bu görsel özellikleri t zamanında ve t-1 zamanında kelime gömme özelliğini alır. Son olarak, cümle açıklamasının ve video görsel içeriğinin anlamsal tutarlılığını garanti etmek için görsel ve cümle özelliklerini ortak bir alana eşlemek için multimodal yerleştirme kullanılmıştır. Bu nedenle, UKSB'ler aynı anda UKSB ve görsel-anlamsal gömme öğrenmelerini keşfedebilir. Kıyaslama veri kümeleri üzerinde yapılan deneyler, tek bir özelliği kullanan yöntemin, hem BLEU hem de METEOR 'da video altyazısı için en yeni temel hatlardan rekabetçi veya daha iyi sonuçlar elde edebileceğini göstermektedir.

Xiao ve Shi (2019), odak tabanlı yaklaşım kullanarak video etiketleme yöntemi önermişlerdir. Odak tabanlı yaklaşım da giriş ve çıkış bağlamına göre giriş özellikleri seçici bir şekilde kullanılmıştır. Video özelliğini kodlamak için Bi - UKSB'yi kullanılmış ve aktivasyon vektörleri eşitlik 3'de verilmiştir.

$$h_t = h_t^{(f)} + h_t^{(b)} \quad (3)$$

(t, ileri ve geri gizlenmiş aktivasyon vektörleridir.)
Önerilen mimaride, EAS'den elde edilen görsel özellikleri kodlamak için bir Bi - UKSB ve kod çözme için üç tek yönlü UKSB katmanı (turuncu dikdörtgen) kullanılmaktadır. Birinci UKSB (yukarıdan aşağıya doğru turuncu dikdörtgen), daha önceki kelimelerin zengin bilgilerini içeren orijinal girdi dizilerini verimli bir şekilde kodlamak için kullanılır; bu, denetleyicinin tasarımına daha iyi bir potansiyel kazandırır. Daha sonra, orta UKSB, ilkinden içerik bilgisi çıktısını kodlar ve dil bilgisinin son temsilini üretir.

Ding ve diğ. (2019), hem video bölümleme hem de anlamsal metin oluşturma için yeni bir yöntem sunmuşlardır. En ilgi çeken kesimlerden elde edilen anahtar kareler, göze çarpma tespiti ve UKSB kullanılarak video başlığına dönüştürülmüştür. İlgili mekanizmaların getirilmesinin amacı, mevcut göreve çok sayıda bilgiden daha önemli bilgilerin seçilmesidir. Uzun videoların anlamsal özetlemesinin doğruluğunu arttırmak için daha doğru nesne tanıma ağı ve detaylı başlıkları olan veri setleri gerekebilir. Deney sonuçları, geliştirilen modelin daha makul video segmentleri ile doğru anahtar kareler ürettiğini göstermiştir.

Belirli bir görüntü içeriği ile cümle oluşturma süreci arasında sıkı bir bağlantı kurmaya çalışılmaktadır. Bu bağlamda Kulkarni ve diğ. (2013), Bu, bir görüntüde nesne niteleyicileri (sıfatlar) ve uzamsal ilişkileri (edatlar) tespit etmiştir. Bu tespitleri tanımlayıcı metinden önceden elde edilen istatistiksel olarak düzeltip daha sonra yumuşatılmış sonuçları cümle oluşturma için kısıtlama olarak kullanılmıştır. Cümle oluşturma ya n-gram dil modeli ya da basit bir şablon temelli yaklaşım kullanılarak gerçekleştirilmiştir. Genel olarak önerilen yaklaşım, çeşitli uzamsal ilişkilerdeki birkaç nesne sınıfının nispeten az sayıda örneğini bile oluşturarak oluşturulabilecek potansiyel olarak çok sayıda sahneyi ele alabilir. Her bir faktör için oldukça küçük sayılar için bile, bu tür düzenlerin toplam sayısının tamamen örneklenmesi mümkün değildir ve herhangi bir görüntü kümesinin belirli bir önyargısı oluşmaktadır. Böyle bir yanlılığı değerlendirmekten kaçınmak için, değerlendirmedeki tüm görüntü özelliklerinden veya sahne / bağlam tanımasından kasıtlı olarak kaçınılmak istenmiştir ancak sunulan modele bir sahne düğümünün ve uygun potansiyel işlevlerin dâhil edilmesinin açık olacağı sonucuna ulaşılmıştır.

Kuznetsova ve diğ. (2014), ağaç benzeri bir yapı ile görüntü tanımları üretimi için bir ağ yapısı sunmuşlardır. Bu yöntemler başlangıçta bir görüntünün içeriğini tanımlayabilse de, metin üretme konusunda yoğun şekilde elle tasarlanmış ve sabittirler. Bu sorunu çözmek için araştırmacılar, sabit görüntü altyazı oluşturma performansını artırmak için derin öğrenme modelleri kullanmışlardır.

Kiros ve diğ. (2014), iki multimodal sinir dili modeli geliştirmişlerdir. Bir image text multimodal sinir dili modeli ile karmaşık cümle sorguları içeren görüntüleri almak ve resimler üzerinde koşullandırılmış metin oluşturma da kullanılmaktadır. Görüntü-metin modellemesi durumunda, modeller, evrişimsel bir ağla birlikte eğiterek kelime temsillerini ve görüntü özelliklerini birlikte öğrenebileceği kanıtlanmıştır. Mevcut yöntemlerin çoğundan farklı olarak, yaklaşımda şablonlar, yapılandırılmış tahminler ve / veya sözdizimsel ağaçlar kullanılmadan görüntüler için cümle açıklamaları oluşturulmuştur. Image - text modellenmesinin yanında geliştirilen algoritma, ses gibi diğer yöntemlere de kolayca uygulanabilmektedir.

Wu ve diğ. (2019), yaptıkları çalışmada, video altyazısı için evrişimli bir yeniden yapılanma tosequence (ConvRS) çerçevesi önermişlerdir. Özellikle, bir video klibin kompakt temsillerini elde etmek için bir yeniden yapılandırma ağı (tek katmanlı evrişimsel eşleme ile daha kompakt özellikleri kodlayarak) kullanmışlardır. Kod çözme aşamasında önce görsel ve sözcüksel özelliklere birleştirip, hiyerarşik bir kod çözücü oluşturmak için çoklu genişletilmiş evrişim katmanları seçilmiştir. Uzun süreli bağımlılıklar hiyerarşik yapı boyunca daha kısa bir yolla yakalanabildiğinden, kod çözücü uzun vadeli bilgilerin kaybını azaltabilmiştir. MSVD ve MSRVTT veri setleri ile karşılaştırıldığında önerilen yöntemin en gelişmiş performansı elde edebileceği gösterilmiştir.

Li ve diğ. (2018), geliştirdikleri model ile bir videonun yapısını eksiksiz bir şekilde koruyarak video için doğru bir açıklama sağlamışlardır. İlk olarak, videonun ilgili içeriğini yakalamak için multimodal kod çözücüye bir odak yaklaşımı ve bellek ağları eklemişlerdir. Bir odak yaklaşımı daha önce oluşturulan kelimelere dayanarak ilgili çerçevelerin bir alt kümesini seçer bu süreç görsel kavramların çıkarılması için çok önemlidir. Uzun vadeli bilgileri ezberleme konusunda önemli bir avantaja sahip olan bellek ağları, multimodal kod çözücü için ek geçici dinamikler sağlamaktadır. Daha önce oluşturulan kelimelerin seçimi için daha ilgili özellikleri seçmek için ek olarak bir özellik seçim algoritması uygulanmıştır. Özellik seçimi algoritması model performansını daha da arttırmıştır.

Wang ve diğ. (2018), video etiketleme için videonun sıralı karelerini bir zaman çizelgesinde uzamsal - zamansal gösterimli iki katmanlı GTB kod çözücüyü benimseyen yeni bir SeqInSeq modeli önermişlerdir. Model, MSVD veri kümesi üzerinde değerlendirilmiştir.

Baraldi ve diğ. (2017), video veri setlerindeki süreksizlik noktalarını tanımlayabilen bir UKSB hücresi önermişlerdir. Bir video klibin aralarında süreksizlik noktaları olan birkaç sahne olabileceğinden, yazarlar bir sınır detektörü eğitmiş, böylece UKSB hücresi bir videoyu süreksizlik noktalarına bölmüş ve birden çok parçaya kodlayabilmiştir. Kısa vadeli bağlam vektörleri, her yığından çıktı olarak alıp son bağlam vektörünü kodlamak için sırayla başka bir UKSB 'e iletilmiştir. Devamsızlık noktaları tespit edilerek, ek bir UKSB 'de gizli özellikleri dâhil ederek UKSB'nin sınırlandırılmasını incelemişlerdir. Başka bir deyişle, UKSB'nin zamansal yakalama aralığının sınırlarını genişletmek için her bir zaman adımında gizli özelliklerin hiyerarşik yapısını yapılandırmışlardır.

Özer ve diğ. (2020), bu çalışmada, görme engelli bireylerin olayları gerçek zamanlı görüntülerle analiz etmeleri ve anlamlı cümleler halinde ifade etmeleri için bir video altyazı sistemi geliştirilmiştir. Görme engelli bireylerin günlük yaşamlarında yaşadıkları sorunları daha iyi anlamak amaçlanmaktadır. Bu nedenle Altı nokta Kör Derneği bünyesindeki engelli bireylerin görüş ve önerileri sorunlarına daha gerçekçi çözümler üretmek amacıyla toplanmıştır. Bu çalışmada, 1970 YouTube kliplerinden oluşan MSVD eğitim veri seti olarak kullanılmıştır. İlk olarak, tüm kliplerin sesi kapatıldı, böylece kliplerin sesleri cümle çıkarma sürecinde kullanılmadı. EAS ve UKSB mimarileri cümle oluşturmak için kullanılmış ve deneysel sonuçlar BLEU 4, ROUGE-L ve CIDEr ve METEOR kullanılarak karşılaştırılmıştır.

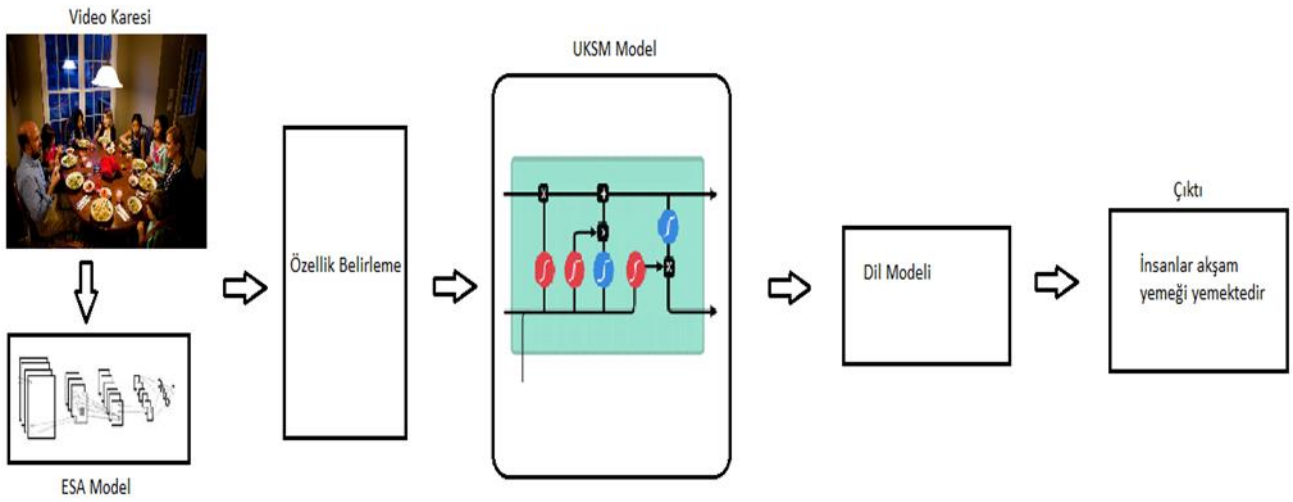
Video altyazılama olarak da adlandırılan, videoların doğal dilde açıklamalarının otomatik üretilmesi, yakın zaman önce çalışılmaya başlanan zorlu bir bütünleşik görme ve dil problemidir. Her ne kadar araştırmacılar İngilizce için çok sayıda çözüm önermiş, olsalar da, Türkçe video altyazı modellerini eğitmek için uygun veri kümelerinin bulunmamasından dolayı, Türkçe üzerinde henüz bir çalışma ortaya konmuş, değildir. Bu eksikliği gidermek için Çıtamak ve diğ. (2019), MSVD veri kümesindeki İngilizce açıklamaların Türkçeye çevrilmesiyle geniş çaplı bir Türkçe denektaşı veri kümesi oluşturulmuştur. Buna ek olarak, zamansal odak yaklaşımına sahip UKSB tabanlı diziden diziye mimarileri içeren çeşitli noral modellerin gerçekleştirimleri yapılmış, olup bu güçlü temel yöntemlerin veri kümesi üzerindeki başarımları gösterilmektedir. Veri kümesinin Türkçe altyazılama üzerine yapılacak gelecek çalışmalar için iyi bir kaynak oluşturulmuştur.

Görüntülerin dogal cümlelerle otomatik olarak tasvir edilmesi literatürde çok yakın zamanlarda incelenmeye başlanmış olan ve son derece zorlu kabul edilen bir araştırma problemidir. Bu problemin çözümüne yönelik ortaya konan yaklaşımların sayısının giderek artmasına rağmen bu alanda yaygın olarak kullanılan veri kümelerinin sadece İngilizce açıklamalar içermeleri nedeniyle bu çalışmalar büyük ölçüde tek dillidir ve İngilizce ile kısıtlı kalmıştır.

Ünal ve diğ. (2016), bu çalışmada, TasvirEt adı verilen bu veri kümesi üzerinde, yine literatürde ilk kez Türkçe görüntü altyazılama amacıyla kullanılabilecek iki yaklaşım da önerilmektedir. Bu veri kümesi, hâlihazırda İngilizce açıklamaları olan Flickr8K veri kümesinin Türkçe altyazılar ile zenginleştirilmiş halidir. Veri güdümlü yaklaşımların başarısı, veri kümesindeki altyazı miktarı ile doğru orantılı olduğu için, veri girişinin yaygınlaştırılması ve altyazıların çoğaltılması faydalı olacaktır. TasvirEt veri kümesi, Türkçe altyazılama amacı ile kullanılabileceği gibi, çok dilli yapısı sayesinde, altyazılama probleminde dillerin kullanımındaki farklılıkların da araştırılmasına olanak sağlayacak bir veri kümesidir.

3. Video Etiketleme (Video Captioning)

Video etiketleme, video dizisinin görsel içeriğini anlama ve uygun bir etikete dönüştürme sürecidir. Video, insanların bilgi edinmesi ve duyguları iletmesi için önemli bir kaynak olduğundan, otomatik video alt yazısı üretimi, video alma (Li ve diğ., 2019) ve otomatik video altyazısı (Yuan ve diğ., 2019), gibi günlük senaryolarda birçok pratik uygulamaya sahip olabilir. Örneğin, YouTube'a her dakika 300 saatlik video yüklenmektedir (Jegham ve diğ., 2020) ve dolayısıyla etkili bir video altyazı modeli çevrimiçi videolar için etiketleme, dizine ekleme ve arama kalitesini artırma potansiyeline sahiptir; konuşma sentez teknolojisi ile birlikte, video altyazısı kişilerin video içeriğiyle daha iyi etkileşim kurmaları için ek açıklamalar oluşturmaya yardımcı olabilir (Trabelsi ve diğ., 2019). Video etiketleme süreci Şekil 1'de gösterilmiştir.

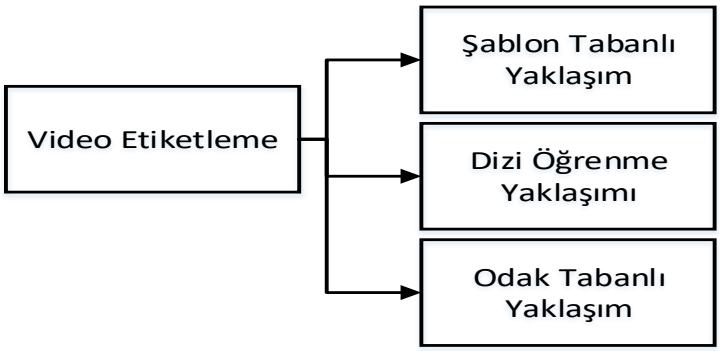


Şekil 1. Video etiketleme (Video captioning)

Ayrıca, video sınıflandırma görevinden farklı olarak, videoda komut dosyası oluşturma, videodaki göze çarpan özelliklerle temsil edilen anlamsal içeriğin kodunun çözülmesi ve bilginin tekrar dile dönüştürülmesi gerekmektedir.

3.1. Video Etiketleme Yaklaşımları (Video Captioning Approaches)

Video etiketleme da temel olarak 3 yaklaşım kullanılmaktadır. Bunlar şablon tabanlı yaklaşım, dizi öğrenme yaklaşımı ve odak tabanlı yaklaşımdır. Şekil 2'de baz alınan temel yaklaşımlar gösterilmiştir.



Şekil 2. Video etiketleme kategorileri(Video captioning categories)

3.1.1. Şablon Tabanlı Yaklaşım (Template Based Approach)

Şablon tabanlı yöntemlerde, bazı cümle oluşturma şablonları ve belirli gramer kuralları önceden tanımlanmaktadır. Kelimeler, önceden tanımlanmış şablonlarla son tanımı üretmek için görsel içerikten tespit edilmektedir. Bu yaklaşım, tanımlanmış şablonlara büyük ölçüde bağımlıdır, bu nedenle üretilen cümle her zaman şablonları takip eden sabit bir söz dizimsel yapıya sahip olacaktır. Konu – fiil - nesne üçlüsün çıkarmada nesne ve etkinlik belirteçlerinin çıktılarını gerçek dünya bilgisi ile birleştirilmektedir. Örneğin insan faaliyetleri, eylem ve kavram hiyerarşilerine dayanarak tanımlanmıştır. Farklı cümle parçaları arasındaki anlamsal ilişkiyi öğrenmek için anlamsal bir hiyerarşi tanımlanmıştır. Her iki istatistikten de büyük miktarda metin verisinin ayrıştırılması ve bilgisayarlı görüntünün tanıma algoritmalarından yararlanılmıştır (Li ve diğ., 2019).

Rohrbach ve diğ. (2013), geliştirdikleri model, ilk önce videodan anlamsal etiketlerin ara temsilini öğrenen ve daha sonra istatistiksel makine çevirisinden doğal dili çeviren iki aşamalı bir yaklaşım geliştirmişlerdir. İçeriğin doğal bir tanım oluşturması için model, görselde doğrudan bulunmamasına rağmen hangi kelimelerin eklenmesi gerektiğini öğrenmiştir. Eğitim için bir dizi video snippet'i ve cümle içeren paralel bir corpus olduğu varsayılmıştır. Video tanımlayıcı tarafından temsil edilen video snippet'leri bir cümle ile hizalanır, yani bir video snippeti ve ona karşılık gelen bir cümle ikilisi oluşturulmuştur. TACoS corpus'a dayanan bir mutfak senaryosunda insan etkinliği videolarını kullanarak önerilen yaklaşım değerlendirilmiştir. Otomatik değerlendirme ve insan kararlarını kullanarak, önceki çalışmaların motive ettiği çeşitli temel yaklaşımlar üzerinde önemli gelişmeler sağlanmıştır.

3.1.2. Dizi Öğrenme Yaklaşımı (Sequence Learning Approach)

Şablon temelli yöntemlerle karşılaştırıldığında, dizi öğrenme yaklaşımları görsel girdi ile ilgili cümle açıklamasını daha esnek söz dizimsel yapılarla doğrudan üretmeyi amaçlar (Wang ve diğ., 2018). Dizi öğrenme modelleri, önceden tanımlanmış şablonlara dayanmadan, video içeriğini doğrudan bir cümleyi tercüme eder.

Bu yöntem grubu genellikle EAS ve UKSB içeren iki yaklaşım temel almıştır. Bu model, bir kaynak dilin cümlelerini hedef dile çeviren makine çevirisi algoritmalarına dayanmaktadır. Bu yöntem de ilk olarak vektör özellikli bir çerçeve sırasını eşlemek için bir kodlayıcı kullanılır ve ardından kod çözücü tarafından çevrilmiş bir cümle oluşturulur.

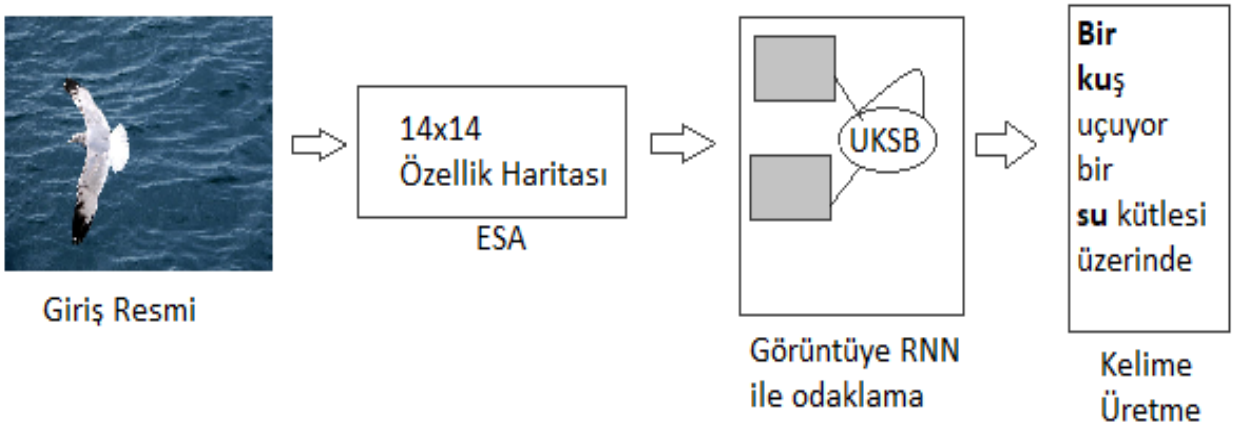
Kodlayıcı olarak genel de EAS kullanılmaktadır. Bir kodlayıcının amacı, kompakt ve temsili olan özellik vektörlerini hesaplamak ve kod çözücü için en alakalı görsel bilgiyi yakalayabilmektir (Song ve diğ., 2015). Büyük ölçekli görüntü tanıma, nesne algılama ve görsel altyazıda büyük başarı elde eden derin EAS'lerin hızlı gelişimi ile derin EAS'ler üst veya orta katmanlarından üst düzey özellikler çıkarabilir (Xu ve diğ., 2018). Kod çözücü olarak genel de TSA kullanılmaktadır.

Gan ve diğ, (2017), sorgu görüntüsü için metinsel bir açıklama oluşturmak üzere semantik kavramları kullanarak bir semantik bileşim ağı oluşturmuştur. Tüm etiketlerin olasılığı, topluluktaki UKSB ağırlık matrislerini işlemek için semantik kavram vektörünü oluşturmak için kullanılmıştır. Bununla birlikte görüntünün açıklamasını üretmek için işbirlikçi anlamsal kavram bağımlı ağırlık matrislerini öğrenmek için avantajı sağladığı gözlenmiştir.

3.1.3. Odak Tabanlı Yaklaşım (Attention Based Approach)

Yukarıdaki açıklanan yöntemler, esas olarak görüntüyü önceden eğitilmiş derin öğrenme mimarileri ile kodlar ve doğal dil cümlesi oluşturmak için kod çözme işlemi sırasında görüntü içeriğini sabit tutmaktadır. Ancak, gerekli tüm bilgileri tek bir dosyaya ayrıştırmak kolay bir iş değildir. Bu nedenle, bağlama göre farklı görüntü bölgelerine bakmak altyazı geliştirmek için yararlı olacaktır. Bu doğrultuda genellikle TSA kod çözücünün nereye ve neye katılması gerektiğini öğrenen görüntü altyazısı için attention based (odak tabanlı) yaklaşımı yaygın olarak kullanılmaktadır. Derin sinir ağlarındaki odak yaklaşımı, tahminlerde bulunmak için zaman içinde bilginin en alakalı kısımlarına odaklanan insanların dikkat etme özelliklerini ilham almaktadır. Mekânsal odak ve anlamsal odak dahil olmak üzere iki ortak odak yaklaşımı incelenmiştir (Li ve diğ., 2019).

Mekânsal Odak (Spatial Attention): Xu ve diğ. (2016), ilk olarak görüntü video etiketleme süreci için mekânsal odak yaklaşımını kullanmışlardır. Bu yaklaşım Şekil 3'de gösterilmiştir.



Şekil 3. Mekânsal odak yaklaşımı (Spatial attention approach)

Modellerinde, önceden eğitilmiş EAS'lerin son evrişimsel katmanını, tamamen bağlı bir katman kullanmak yerine görüntü kodlayıcı için kullanılmışlardır. Bu şekilde, görsel bilgi, verilen görüntünün farklı bölgelerine karşılık gelen bir dizi, örneğin, bir a = $\{\mathbf{a}_1, ..., \mathbf{a}_L\}$, $\mathbf{a}_i \in R^D$ olarak vektörleştirilir, ve dolayısıyla TSA kod çözücünün odak yaklaşımı altında farklı uzamsal görüntü bölgelerine katılmasına izin verir. Önceki çalışmalar gibi (Vinyals ve diğ., 2015) TSA kod çözücü de tek katmanlı bir UKSB olarak formüle edilmiştir. Bununla birlikte, görsel içeriği sabit tutmak yerine, her adımında UKSB'nin gizli durumunu aşağıdaki gibi hesaplamak için temel bir bağlam vektörü kavramı sunulmuştur.

$$h_t = \mathrm{LSTM}(x_{t-1}, h_{t-1}, Z_t) \qquad (4)$$

burada z_t, t zamanında bağlam vektörünü temsil eder ve her bir zaman adımında görüntü bölgeleri ve oluşturulan kelimeler arasındaki alaka düzeyini açıkça göz önünde bulundurmaktadır.

$$z_t = \varphi(\{\mathbf{a}_i\}, \{\alpha_i\}) \qquad (5)$$

burada α_i, her bir görüntü bölgesinin ağırlığını temsil eder ve φ (.), görüntü bölgeleri ve bunlara karşılık gelen ağırlıkları üzerinde bir tekli vektör temsilini hesaplayan bir füzyon fonksiyonu olarak çalışmaktadır. Her bir görüntü bölgesinin ağırlık a_i'si, dikkat fonksiyonunu temsil ettiği yerde elde edilmiştir.

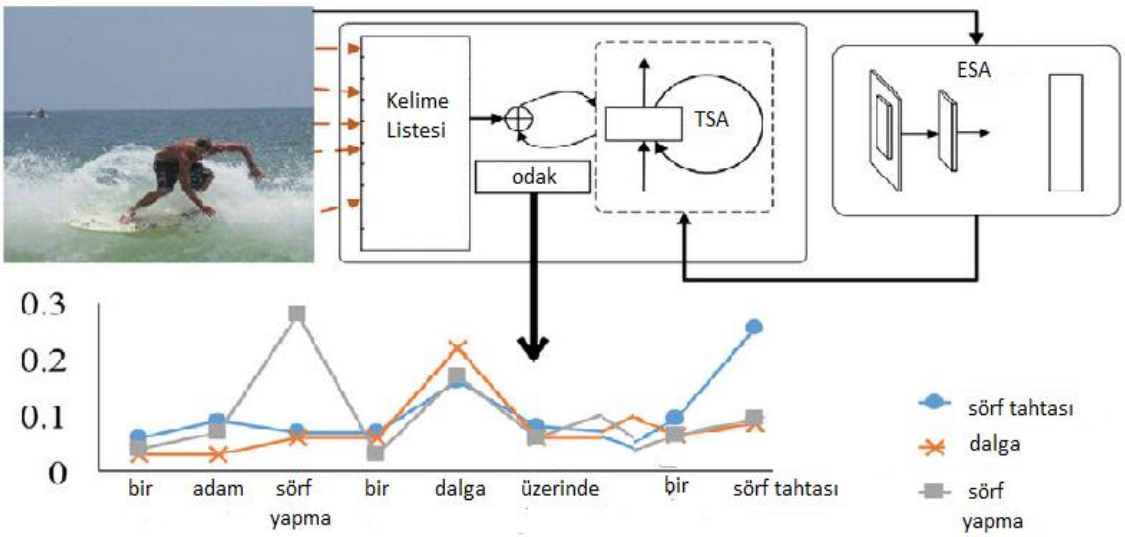
$$e_{ti} = f_{att}(a_i, h_{t-1}) \qquad (6)$$

$$a_{ti} = \exp(e_{ti}) \Big/ \sum_{k=1}^{L} \exp(e_{tk}) \qquad (7)$$

Çok katmanlı bir algılama ağı tarafından formüle edilmiş ve önceki gizli duruma göre koşullandırılmıştır.

Wang ve diğ. (2018), bir ReviewNet ile küresel bilgileri dikkate alarak bağlam vektörlerini düşünce vektörleri olarak genişletmeyi önermişlerdir. Hem TSA hem de EAS ağlarını, görüntü içeriğinin tüm görünümünü yakalamak ve ayrıca dizi olarak formüle etmek için görsel kodlayıcılar olarak benimsemişlerdir. UKSB katmanı tarafından formüle edilen ReviewNet, görsel diziyi mekânsal odakla yakalamak ve zamansal odak kullanarak düşünce vektörlerini oluşturmak için kullanılmaktadır. Bu inceleme süreci, bir veri çiftini formüle etmek için birkaç kez tekrarlanmıştır. (Yang ve diğ.,2016)'deki metin kod çözücü, girdi düşünce vektörlerine odak yaklaşımı yerleştirilmiş bir UKSB katmanını da benimsemektedir. Altyazı oluşturma işlemi sırasında görüntü özellikleriyle uzamsal odağı artıran (Xu ve diğ., 2016), 'den farklı olarak, ReviewNet nereye katılacağını "gözden geçirir" ve doğrudan kodlanmış görsel özellikleri kullanmaktan daha fazla küresel özelliği keşfetmeyi bekler.

Anlamsal Odak (Semantic Attention): Yüksek seviyeli semantik bilgilerin, metin kod çözücüye (Venugopalan ve diğ., 2014), görsel kodlayıcı mimarisine yardımcı olmak için yararlı olduğu gösterilmiştir; burada semantikler genellikle görsel nitelikler (You ve diğ.,2016) (Venugopalan ve diğ., 2014), (Yao ve diğ., 2016) olarak formüle edilir. Bu yöntemler genellikle metin kod çözücüye, hem kodlanmış görüntü içeriğini hem de algılanan görsel nitelikleri besleyerek üretim süreci için tamamlayıcı bilgi sağlamaktadır. Odak yaklaşımını mekânsal görüntü bölgelerinden "görsel kelimelere" aktarılırlar. Bu yaklaşım Şekil 4'de gösterilmiştir.



Şekil 4. Anlamsal odak yaklaşımı (Semantic attention approach)

Yüksek seviyeli semantiğin TSA kod çözücüsüne nasıl dâhil edileceğine dair kapsamlı bir araştırma Yao ve diğ. (2016), tarafından yapılmıştır. Bu çalışma odak yaklaşımını kullanmasa da, farklı özellikler arasındaki ilişkiyi açıkça ele almıştır.

Zamansal odak yaklaşımının bir tamamlayıcısı olarak, (Yu ve diğ., 2017) insan bakış verileri tarafından denetlenen kare başına mekânsal odak haritasını tahmin eden ve daha sonra her zamanki gibi yumuşak geçici odakları kullanılarak maskelenmiş görsel özellikler havuzundan resim yazısı oluşturan Gaze Encoder Attention Network (GEAN) önerilmiştir. Bir video izlerken bakış izlemenin özneler arasında oldukça kararlı olduğu, yani mekânsal odağın video altyazısına bağlanabileceği ve insan bakış verilerinin güçlü bir eğitim sinyali olarak kullanılabileceği gözlemlenmiştir. Reccurent Glance Prediction (RGP) modeli, GTB aracılığıyla zaman içinde çerçevelerin görsel özelliklerinin tarihini kaynaştırır ve kare başına bakış haritasını geçici sırayla tahmin eder. Bakış haritası daha sonra karşılık gelen 3B hareket özelliği haritasına ve fovea özellik haritasına uygulanır, çünkü yazarlar insanın odaklanmış bölgeleri daha fazla nöronla yüksek görme keskinliğinde algıladığını, periferik sahne alanları ise daha az nöronla düşük çözünürlükte algıladığını iddia eder. GEAN'ın geri kalanından bağımsız olarak, RGP, insan bakış verileri ile büyük ölçekli bir aktivite tanıma veri seti olan Hollywood2 EM veri seti ve kendi kendine oluşturulmuş bir oyuncak veri seti VAS ile eğitilmiştir.

Bu özelliklerin yanı sıra oluşturulan modeller videonun hangi bölümünün açıklama ile daha alakalı olduğunu belirleyebilmelidir. Çünkü altyazılar normalde videoda bulunan gerçeklerin küçük bir kısmına odaklanır. Daha da önemlisi, model açıklamayı oluşturmaya başladığında, anlamlı açıklamalar oluşturmak için yine de video kareleriyle birlikte hatırlatılmalıdır (Xiao ve Shi, 2019).

4. Derin Öğrenme ile Video Etiketleme (Video Captioning with Deep Learning)

EAS ve UKSB mimarisinin birleşimi, video etiketleme konusunda yapılan çalışmalarda kullanılan temel yapı haline gelmiştir. Girdi videosunun etiketleme kalitesi için hem özellik çıkarma hem de kelime oluşturma yöntemleri önemlidir.

Öncelikli olarak görüntü, hareket ve ses özellikleri de dahil olmak üzere video verilerinden genel geçici özellikler çıkarılır.

EAS, nesneleri tespit edebilme ve onları sınıflandırmada kullanılan önemli bir yaklaşımdır. EAS'lerin en büyük avantajı, önemli özelliklerin herhangi bir insan gözetimi olmadan otomatik olarak tespit edilebilmesidir. EAS; konvolüsyon, havuzlama ve tam bağlantılı katmanlarından oluşur. konvolüsyon ve havuzlama katmanı bir veya birden fazla olarak kullanılabilmektedir. EAS mimarisi, eylemleri ve görüntü bağlamına göre nesnelerle ilişkilendirir. Etkili bir video altyazısı için dikkatli bir EAS modeli seçimi önemlidir (Aafaq ve diğ., 2019). Nesne algılama da son zamanlarda en sık kullanılan EAS modelleri şunlardır; Inception ResNet V2 (Szegedy ve diğ., 2016), C3B (Tran ve diğ., 2015) ve I3B (Carreira ve diğ., 2017) Videoları tanımlama konusunda çalışan araştırmalar, birden çok özelliğin video altyazı modellerini iyileştirebileceğini kanıtlamışlardır (Yao ve diğ.,2015, Venugopalan ve diğ., 2015, Yingwei ve diğ., 2015, Yingwei ve diğ., 2016).

EAS (Liu ve diğ., 2020) kullanılarak video sınıflandırma ve görüntü altyazısında önemli ilerleme kaydedilmiş olsa da, video etiketleme görevi daha zordur ve hala zorlayıcı olmaya devam etmektedir. 4-10 saniyelik bir videoyu karakterize edip bir cümle oluşturmak için; 100-250 kareden oluşan klip, temel olarak sıralı görsel verilerin anlayıp aynı zamanda bu anlayışın doğal dile çevrilmesi için bir kapasite gerektirmektedir.

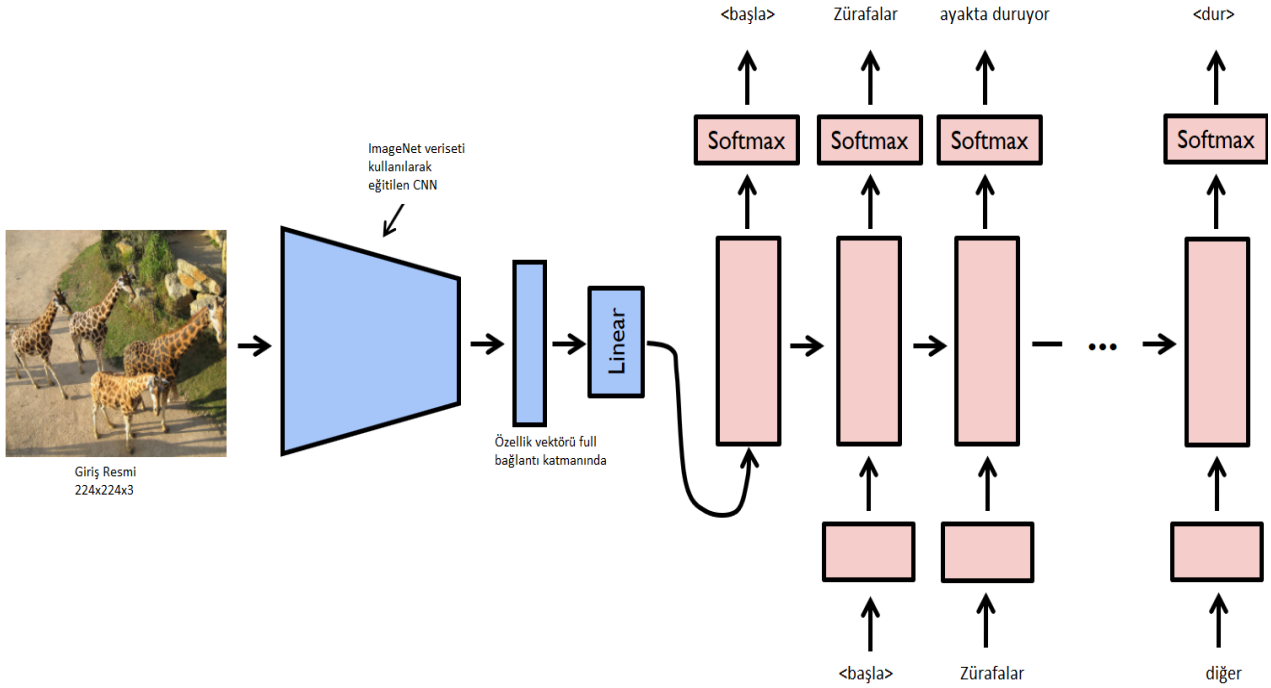
Geleneksel TSA'lar (Gers 2001), (Hochreiter ve Schmidhuber, 1997) girdi dizilerini gizli durum dizisine eşleyerek karmaşık zamansal dinamikleri öğrenmek için tasarlanmışlardır. Geleneksel TSA'ların metin oluşturma ve konuşma tanımada başarılı oldukları kanıtlanmıştır. Ancak TSA'larda, kaybolan gradyan sorunu nedeniyle uzun menzilli geçici bağımlılık videoları hakkında iyi işlem yapmak zordur (Yang ve diğ., 2018). Bu nedenle uzun vadeli hafızaya gereksinim olduğundan UKSB'ler ortaya çıkmıştır. Her UKSB katmanı, bir veya daha fazla tekrarlanan bağlı bellek hücresi, giriş, çıkış ve unutma kapılarını içerir. Doğal dil açıklamaları oluşturmak için kod çözücü olarak bir TSA (Elman, 1990), GTB (Cho ve diğ., 2014) veya UKSB ağları (Amaresh ve Chitrakala, 2019) kullanılır.

EAS, görüntülerdeki uzamsal verilerden özellik vektörleri oluşturmak için kullanılır ve vektörler, tam olarak bağlanmış doğrusal katman yoluyla TSA veya UKSB mimarisine beslenir ve sonuçta bir görüntünün tanımını üreten sıralı veri veya kelime dizisi oluşturulur. Kısacası giriş görüntüleri bir EAS tarafından işlenecek ve EAS çıkışını, açıklayıcı metinler üretebilmek için TSA girişine bağlanacaktır.

Bir açıklama üretmek için daha önceden tanımlı eğitilmiş mimariler ile bir EAS'ye belirli bir görüntü beslenmektedir. Bu ağın sonunda görüntüdeki uzamsal içeriği temsil eden bir dizi özellik aranmaktadır. EAS orijinal görüntüde bulunan büyük miktardaki veriden daha küçük bir gösterime sıkıştıran bir özellik çıkarıcısı olarak kullanıldığı için bu EAS'ye genellikle kodlayıcı denir, çünkü görüntünün içeriğini daha küçük bir özellik vektörüne kodlar. Daha sonra bu özellik vektörünü işleyebilir ve TSA'ya girdi olarak kullanılabilmektedir.

TSA'nın görevi, işlem vektörünün kodunu çözmek ve bir sözcük dizisine dönüştürmektir. Bu nedenle, ağın bu kısmına genellikle kod çözücü denir. TSA bileşeni de bölüm 5'de açıklanan veri setleri ile altyazılar konusunda eğitilir. Eğitimin amacı önceki kelimelere dayalı olarak bir cümlenin bir sonraki kelimesini tahmin etmektir. Geri yayılımı etkili bir şekilde gerçekleştirmek ve benzer çıktılar üretmeyi önlemek için iyi tanımlanmış bir sayısal kodlama ihtiyacı vardır. Bu yüzden, görüntüyle ilişkili altyazıları bir token sözcük listesine dönüştürülür.

Bir UKSB'nin gizli durumu, UKSB giriş belirtecinin bir fonksiyonudur ve önceki durum aynı zamanda yineleme fonksiyonu olarak da adlandırılır. Yineleme işlevi ağırlıklar tarafından tanımlanır ve eğitim işlemi sırasında bu model, UKSB hücreleri geçerli giriş sözcüğü verilen başlıkta doğru sonraki sözcüğü üretmeyi öğrenene kadar bu ağırlıkları güncellemek için geri yayılım kullanır. Model, her eğitimden sonra ağırlıklarını günceller, tek bir eğitim adımı sırasında ağ üzerinden gönderilen görüntü altyazısı çiftlerinin sayısıdır. Model eğitildikten sonra, birçok resim yazısı çiftinden öğrenmiş olup yeni resim verileri için resim yazıları oluşturulmaktadır. Şekil 5'de EAS (kodlayıcı) – TSA veya UKSB (kod çözücü) yapısında kullanılan mimari yapı gösterilmiştir.



Şekil 5. Encoder-Decoder yapısı (Encoder-Decoder structure) (Xu ve diğ., 2016)

5. Karşılaştırmalı Veri Setleri (Benchmark Datasets)

Video etiketleme, resim yazısı ile benzer sorunlara sahiptir, video yazısındaki ek bir zorluk da, zengin video açıklamaları ile gelen veri setlerinin azlığıdır. Bunun nedeni videoların seslendirilmesinin çok daha zor ve pahalı olmasıdır. Yaygın olarak kullanılan ve video altyazısı oluşturma işleminin öğrenmeye dayalı yöntemlerle sınırlı bir şekilde başarılmasını sağlayan birçok klasik kıyaslama veri seti oluşturulmuştur. Bunlar, video sayısı ile anlatım derecesi bakımından küçük, anlamsal ve görsel içerik açısından basitlerdir. Yakın zamana kadar, insan düzeyindeki video tanımlarını oluşturmada derinlemesine öğrenmeden tam anlamıyla faydalanabilmek için birçok büyük ölçekli, yoğun açıklamalı video yazısı veri setleri geliştirilmiştir (Su, 2018). En çok kullanılan veri setleri ve özellikleri Tablo 2'de verilmiştir.

Tablo 2. Veri Setleri (Datasets)

Veri Seti	İçerik	Kaynak	Videolar	Klipler	Video başına cümle sayısı	Kelime sayısı	Video süresi (saat)
TACoS	Cooking	AMTurker	127	7206	18227	146771	15.9
MSVD	Youtube	AMTurker	1970	1970	70028	607339	5.3
M-VAD	Movie	DVS	92	49000	55904	519933	84.6
MPII-MD	Movie	DVS+Script	94	68000	68375	653467	73.6
MSR-VTT	Open	AMTurker	7180	10000	200000	1856523	41.2
ActivityNet Captionings	Open	AMTurker	20000	20000	100000	1348000	849.0

TACoS veri seti (Regneri ve diğ., 2013), ilk çalışmalardan biridir ve iç mekanlardaki farklı etkinliklerin videolarını içermektedir. Videoların süresi uzundur, genellikle dakikalar civarındadır. Her videoya, hem geçici konumlara sahip etkinlik etiketleri hem de birden fazla geçici konumlarla ilgili açıklamalar eklenmiştir. 7206 benzersiz zaman aralığında toplam 18227 video cümle çiftine sahiptir. Veri kümesi ayrıca, her bir etkinliğin başladığı ve bittiği yaklaşık zaman damgaları elde ederek etkinlikleri tanımlayan cümlelerin hizalanmasını sağlamaktadır (Rohrbach, 2012).

TACoS - Multi Veri Kümesi, geçici segment başına paragraf açıklaması içeren veri kümesinin bir uzantısıdır. TACoS corpus'taki her video için aşağıdakileri içeren üç açıklama seviyesi toplanmıştır: (1) video başına en fazla 15 cümleyle ayrıntılı video açıklaması; (2) video başına 3-5 cümle içeren kısa bir açıklama ve son olarak (3) videonun tek bir cümle açıklaması. Verilere ek açıklama, nesne, etkinlik, araç, kaynak ve hedef gibi tuples şeklinde her zaman konu olan bir kişi ile sağlanır (Rohrbach ve diğ., 2014, Regneri ve diğ., 2013).

YouCook veri seti (Das ve diğ., 2013), farklı tarifler pişiren farklı kişilerin 88 YouTube pişirme videosundan oluşur. Videoların çoğunda arka plan (mutfak / sahne) farklı. Bu veri seti, aynı mutfakta ve aynı arka planda sabit kamera izleme noktasıyla kaydedilen MP-II Cooking (Graves ve Jaitly, 2014), veri setinden daha zor bir görsel problemdir. Veri seti, ızgara, fırınlama vb. Gibi altı farklı pişirme stiline ayrılmıştır. Makine öğrenimi için eğitim seti 49 video içerir ve test seti 39 video içerir. Eğitim videoları için nesnelerin ve eylemlerin kare şeklinde ek açıklamaları da sağlanır.

MPII Movie Description Corpus (MPII - MD) (Rohrbach ve diğ., 2015), M - VAD'ye benzer bir şekilde inşa edilmiş olan yeni bir büyük ölçekli film açıklaması veri setidir. 55 sesli açıklamadan, mevcut filmlerden yaklaşık 37000 film klibi ve 49 Hollywood filminden yaklaşık 31000 film klibi içerir. Her video klip, film betiklerinden bir cümle ve DVD açıklayıcı video servisinden (DVS) bir cümle ile donatılmıştır. Veri kümesinin ek açıklamaları yarı otomatik olarak bölümlenir ve elle kliplerle hizalanır. Manuel olarak düzeltildiğinden, video snippet'leri ve açıklamaları arasındaki hizalama bu durumda MVAD'dan daha doğrudur

Montreal Video Annotation Dataset (M - VAD) (Torabi ve diğ., 2015), DVD açıklayıcı video hizmeti anlatımlarından gelen büyük ölçekli bir film açıklaması veri setidir. Görme engelli insanlara yardım etmek için üretilmiş, filmlerin görsel öğelerini tanımlayan ses parçalarıdır. Veri setinde 92 DVD filmden 49000 video klip çekilmiştir. Her klip, yarı otomatik olarak kopyalanmış ve tek bir cümle anlatımına eşlik eder. Kelime kullanımı, ilgili filmin türüne göre değişir. Filmlerin görsel ve metin içeriğinin çeşitliliğinin yüksek olması nedeniyle video altyazı görevi için özellikle zorlayıcı bir veri kümesidir.

VideoStory (Gella ve diğ., 2018), 20k sosyal medya videoları içeren bir çoklu cümle açıklama veri kümesidir. Bu veri kümesinin, tek bir cümle ile yeterince gösterilemeyen uzun videoların hikâye anlatımı veya açıklama oluşturma konusunu ele alması amaçlanmıştır. Her video en az bir paragrafla eşleştirilir. Paragraf başına ortalama geçici olarak yerelleştirilmiş cümle sayısı 4.67'dir. Veri kümesinde 123k cümle içeren toplam 26245 paragraf vardır ve cümle başına ortalama 13.32 kelime vardır. Ortalama olarak, her paragraf video içeriğinin% 96,7'sini kapsar. Veri kümesi, birlikte meydana gelen olaylar arasında yaklaşık% 22 oranında geçici çakışma içerir. Veri seti, sırasıyla 17908, 999 ve 1011 videodan oluşan eğitim, validasyon ve test bölümlerine sahiptir ve ayrıca 1039 video içeren bir kör test seti önerir. Her eğitim videosuna bir paragraf eşlik eder, ancak doğrulama ve test setlerindeki videoların her biri değerlendirme için üç paragrafa sahiptir. Kör test için ek açıklamalar yayınlanmaz ve yalnızca farklı yöntemleri karşılaştırmak için sunucuda bulunur.

ActivityNet Captionings (Krishna ve diğ., 2017), yoğun resim yazısı olayları için özel olarak piyasaya sürülen büyük ölçekli bir kıyaslama veri kümesidir. 849 video saat tutarında 20000 video içermektedir. Videolar, çok çeşitli kategorileri kapsayan video arama motorundan toplanır. Ortalama olarak, her videoda toplamda 100000 cümle vardır. Her cümle videonun benzersiz bir bölümünü kapsar ve değişen zaman aralıklarında gerçekleşen bir olayı açıklar. Ortalama olarak, her cümlenin uzunluğu 13.48 kelimedir ve ilgili videonun 36 saniyesini yaklaşık % 31'ini açıklar. Zengin açıklama, zamansal yapıların açıkça keşfedilmesini sağlar.

ActivityNet Entities (ANet - Entities) veri kümesi, (Zhou ve diğ., 2018), varlıkları ve ek açıklamaları olan ilk video veri kümesidir. Bu veri kümesi, farklı altyazılarla birlikte, ActivityNet Altyazı veri kümesinin eğitim ve doğrulama bölümleri üzerine kuruludur.

Microsoft Video Description Corpus (MSVD) (Chen ve Dolan, 2011), aynı zamanda ilk çalışmalarında Youtube Veri Kümesi olarak da adlandırılır ve ilk açık dünya veri kümelerinden biridir. Tek bir olay içeren kısa kliplerden oluşur. Sonuç olarak, her klip, oldukça sabittir ve küçük geçici yapı karmaşıklığı ile 10 saniye ile 25 saniye arasında sürer. Toplamda 1.970 video klibi vardır Spor, hayvanlar ve müzik gibi geniş bir konu yelpazesini kapsar. Her klip, birçok dilde etiketlenmiş birden fazla paralel ve bağımsız cümle ile gelir. Özellikle İngilizce için, video başına kabaca 40

paralel cümleye sahiptir; toplamda 80000klip çifti elde edilir. 16000 benzersiz kelime listesi vardır; Her cümle ortalama 8 kelime içermektedir (Chen ve Dolan, 2011).

MSR Video-to-Text (MSR -VTT) (Xu ve diğ., 2016), piyasaya sürülen büyük ölçekli bir video resim yazısı ölçütüdür. Cümle sayısı ve kelime bilgisi bakımından büyük bir video resim yazısı veri kümesidir. Bir video arama motorundan taranan 10000 video klip, haber, spor vb. dahil video arama kategorisi 20'dir. Her bir klibin süresi 10 ila 30 saniye arasındadır, toplam süre ise 41,2 saattir. Her video klip, semantiğinin iyi bir şekilde ele alınmasını sağlayan birden fazla paralel ve bağımsız cümle ile açıklanmıştır. Toplam 29.000 klip kelimesiyle toplam 200.000 klip cümle çifti bulunmaktadır.

Charades (Sigurdsson ve diğ., 2016), bu veri seti günlük kapalı ev aktiviteleri hakkında 9848 video içermektedir. Bu videolar üç farklı kıtadan 267 AMT çalışanı tarafından kaydedildi. Onlara eylemleri ve nesneleri tanımlayan komut dosyaları verildi ve belirtilen nesnelerle eylemler gerçekleştirmek için komut dosyalarını takip etmeleri istendi. Komut dosyalarında kullanılan nesneler ve eylemler sabit bir sözcük dağarcığından gelir. Videolar 15 farklı iç mekan sahnesinde kaydedilir ve yalnızca 46 nesne ve 157 eylem sınıfı kullanması sınırlıdır. Veri kümesi 157 eylemi açıklayan 66500 ek açıklamadan oluşur. Ayrıca 46 nesne sınıfına 41104 etiket sağlar. Ayrıca, tüm videoları kapsayan 27847 açıklama içerir. Veri kümesindeki videolar, ortalama 30 saniye süren günlük yaşam etkinliklerini gösterir. Veri seti, eğitim ve test amacıyla sırasıyla 7985 ve 1863 videolara ayrılmıştır

Video Titles in the Wild (VTW) (Zeng ve diğ., 2016), içerisindeki Video Başlıkları, klip başına ortalama 1,5 dakika süren 18100 video klip içerir. Her klip yalnızca bir cümleyle açıklanır. Bununla birlikte, ortalama bir kümenin tüm veri kümesinde en fazla iki cümleyle göründüğü çeşitli bir kelime haznesi içerir. Video başına tek bir cümlenin yanı sıra, veri kümesi, klibin görsel içeriğinde bulunmayan bilgileri açıklayan eşlik eden açıklamalar da (artırılmış cümleler olarak bilinir) sağlar. Veri kümesi, video içeriği açıklamasının aksine video başlığı oluşturma için önerilir, ancak video sorusunu yanıtlama da dahil olmak üzere dil düzeyinde anlama görevleri için de kullanılabilir.

6. Değerlendirme Ölçütleri (Evaluation Metrics)

Video yazısı sonucu, doğal dil olarak doğruluk ve anlam bilimin ilgili video ile olan ilgisine dayanarak değerlendirilir. Yaygın olarak kullanılan değerlendirme ölçütleri şunlardır;

SVO Accuracy (Venugopalan ve diğ., 2014), erken dönem çalışmalarında, üretilen SVO (Konu, Fiil, Nesne) üçlemelerinin temel gerçeklerle tutarlı olup olmadığını ölçmek için kullanılır. Bu değerlendirme ölçütlerinin amacı geniş anlambilimin eşleştirilmesine odaklanmak ve görsel ve dil ayrıntılarını göz ardı etmektir.

BLEU (Park ve diğ., 2017), makine çevirisi alanındaki en popüler metriklerden biridir. Fikir, n-gram eşleşme sayımlarının geometrik ortalamasını hesaplayarak iki cümle arasındaki sayısal bir çeviri yakınlığını ölçmektir. Sonuç olarak, kelimelerin uyumsuzluğunu ayarlamak konusunda hassastır. Ayrıca, karmaşık içeriğe adapte olmayı zorlaştıran daha kısa cümleler kurabilir.

$$\mathrm{BLEU}-\mathrm{N}(c_i, S_i) = b\left(c_i, S_i\right) \exp\left[\sum_{n=1}^{N} \omega_n \log P_S\left(c_i, S_i\right)\right] \qquad (8)$$

$$b\left(c_i, S_i\right) = f(x) = \begin{cases} 1, & \text{if } I_C > I_S \\ e^{1-I_s/I_c}, & \text{if } I_C \leq I_S \end{cases} \qquad (9)$$

I_C, c_i aday cümlenin toplam uzunluğu; ls, corpus düzeyinde etkili referans uzunluğunun uzunluğudur. Bir aday cümle için birden fazla referans mevcut olduğunda, en yakın referans uzunluğu seçilir. ω̇n genellikle aşağıdakiler için sabit tutulur.

$$N = 1,2,3,4; \quad \frac{P_n\left(c_i, S_i\right) = \sum_k \min\left(h_k\left(c_i\right), \max h_k\left(s_{ij}\right)\right)}{\sum_k h_k\left(c_i\right)} \qquad (10)$$

BLEU-N kısa, bir aday cümle ile bir referans cümle veya bir dizi referans cümle arasında ortak olan n-gramın (4 grama kadar) fraksiyonunu ölçer (Ma ve Wang, 2017).

ROUGE, metin özetlerini değerlendirmek için 2004 yılında (Lin, 2004), metriği önerilmiştir. Oluşturulan cümlelerin hatırlama puanını hesaplar referans cümleleri ile üretilenler arasındaki n-gram örtüşen dizileri ölçmeleri bakımından BLEU puanına benzer. Aradaki fark ROUGE'in n-gram oluşumlarını toplam referans cümle sayısı toplamında, BLEU ise adayların toplamındaki oluşumları dikkate almasıdır. ROUGE metrik hatırlamaya dayandığından, uzun cümlelerde tercih edilir (Ma ve Wang, 2017).

$$\text{ROUGE} - \text{N} = \frac{\sum_{S\in\{\text{ReferenceSummaries}\}}\text{gram}_n\in S \sum \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S\in\{\text{ReferenceSummaries}\}}\text{gram}_n\in S \sum \text{Count}(\text{gram}_n)} \quad (11)$$

n''nin n-gram, gram_n ve $\text{Count}_{\text{match}}$ (gram_n) uzunluğunu temsil ettiği durumlarda, aday özeti ve bir dizi referans özeti içinde ortaya çıkan maksimum n-gram sayısıdır (Lin, 2004).

CIDEr (Park ve diğ., 2017)başlangıçta, aday resim yazısı ile insan açıklayıcılar tarafından sağlanan referans cümleler arasındaki fikir birliğini ölçen, resim için tanımlayıcı cümleleri değerlendiren bir metriktir. Bu nedenle, insan yargılarıyla yüksek oranda ilişkilidir. Bu ölçü, doğruluğu ve gramer doğruluğunu yakalaması anlamında diğerlerinden farklıdır.

$$\text{CIDEr}_{n(c_i,s_i)} = \frac{1}{m}\sum_j \frac{g^n(c_j).g^n(s_{ij})}{||g^n(c_j)||.\ ||g^n(s_{ij})||} \quad (12)$$

burada $g^n(c_j)$ tüm n-gram uzunlukları temsil eden bir vektördür. $||g^n(c_j)||$, $||\ g^n(c_j)\ ||$'nin büyüklüğünü gösterir. Aynı şey $||g^n(s_{ij})||$ için de geçerlidir. Ayrıca, CIDEr metnin dil bilgisel özelliklerini ve daha zengin anlambilimini yakalamak için daha yüksek n-gram (daha yüksek, daha uzun kelime sırası) kullanır (Vedantam ve diğ., 2015). Bu nedenle, aşağıdaki denklemi kullanarak farklı n-gramların puanlarını birleştirir.

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} \omega_n \text{CIDEr}_n(c_i, S_i) \quad (13)$$

METEOR (Ma ve Wang, 2017), belirli bir hipotez cümlesiyle bir dizi aday referansı arasındaki uyuma göre hesaplanmaktadır. METEOR, WordNet eş anlamlarını kullanarak, tam belirteç eşleşmeleri, köklü belirteçler, parola öbekleri ve semantik olarak benzer eşleşmeleri karşılaştırır. Kelime eşleme modülleri başlangıçta dizi çifti arasındaki tüm olası kelime eşleşmelerini tanımlar. Birden fazla maksimum kelime eşleşme hizalaması bulursa, iki dizgideki kelime sırasının en çok benzediği hizalamayı seçer. Modüllerin çalışma sırası sözcük eşleştirme tercihlerini yansıtır (Lavie ve Agarwall, 2007). METEOR 'un bu anlamsal yönü onu diğerlerinden ayrılmaktadır. Literatürde METEOR 'un her zaman BLEU ve ROUGE'den daha iyi olduğu ve referans sayısının az olduğu durumlarda CIDEr'den daha iyi olduğu görülmüştür.

6. Sonuçlar ve Tartışma (Results and Discussion)

Video etiketleme sistemleri, video görüntülerine, tasvirleme yöntemi ile altyazı ekleyerek kullanıcılara, görüntü karelerini anlama ve önemli noktalara odaklanmada kullanılmaktadır. Video etiketleme sistemlerinde, şablon temelli yaklaşım, dizi öğrenme yaklaşımı ve odak tabanlı yaklaşım gibi farklı yöntemler kullanılmaktadır. Şablon temelli yaklaşım belirli gramer kurallarına dayanmaktadır. Diğer yaklaşımlar derin öğrenme mimarileri kullanarak kodlayıcı- kod çözücü yapısı ile videolardaki dikkat çeken noktaları baz alarak gerçekleştirilmiştir. Şablon temelli yaklaşımlara göre daha doğru ve gerçekçi altyazı ürettiği gözlenmiştir. Her üç yaklaşıma ait örnekler karşılaştırmalı bir şekilde Tablo 3'te sunulmuştur.

İncelenen video etiketleme yaklaşımlarında kullanılmak üzere farklı kategorilerde veri setleri incelenmiştir. Bu konuda kullanılan ölçüt değerleri ile çalışmaların başarıları gözlenmiştir.

Geleneksel yöntemler, video arka plan farklarına göre, videonun özelliklerinden yararlanarak video etiketleme yapmaktadır. Makine öğrenme yaklaşımları için yeni ve popüler bir yaklaşımdır. Video etiketleme konusunda literatürde çok sayıda algoritma önerilmiş olsa da, video içeriklerinin farklılıkları nedeniyle tek bir yöntem tüm uygulamalarda tatmin edici performans sağlayamamaktadır.

Tablo 3' te her yaklaşımdan birer çalışma örnek olarak sunulmuştur. Video etiketleme süreci ilk olarak resim etiketleme olarak başladığı için yaklaşımlar da resim veri setleri de kullanılmıştır. Derin öğrenme yaklaşımında gerçekleşen yenilikler gelişmeler video etiketleme çalışmalarında da etkisini göstermiştir. İlk yaklaşım olan şablon temelli yaklaşım etiket (cümle) oluşturma aşamasında geleneksel yöntem olarak kabul ettiğimiz doğal dil üretme yaklaşımını kullanmıştır. Dizi öğrenme ve odak tabanlı yaklaşım hem görüntü işleme hem de dil (cümle) oluşturma aşamasında derin öğrenme yaklaşımlarını kullanmıştır. Şablon temelli yaklaşımda üretilen açıklamalar süreçteki şablona bağımlıdır. Kullanılan şablon yeteri kadar veri veya kural içermiyorsa üretilen alt yazı da o kadar doğru sonuç vermez. Dizi öğrenme yaklaşımı da gerek görüntü işleme gerek cümle üretme sürecinde oluşturulan ESA ve TSA veya UKSB mimarilerine bunların eğitim setlerine bağımlıdır. Odak tabanlı yaklaşımlarda kullanılan modeller, veri setleri ve bazı dikkat noktaları baz alınarak oluşturulurlar. Bu da videolardaki karelerin, oluşturulan cümleleri farklı derecede etkilemektedir.

Tablo 3. Yaklaşımların Karşılaştırılması ve Değerlendirilmesi (Comparing and Evaluating Approach)

Referans	Yöntem	İçerik	Sonuç
Rohrbach ve diğ.	Şablon tabanlı yaklaşım	TaCoS veri seti ve Doğal dil üretme ile Anlamsal gösterim yöntemleri	Otomatik değerlendirme ve insan kararlarını kullanarak, önceki çalışmaların motive ettiği çeşitli temel yaklaşımlar üzerinde önemli gelişmeler sağlanmıştır.
Kojima ve Tamura	Şablon tabanlı yaklaşım	Doğal dil üretme ile kavram hiyerarşilerine ve ilişkilerine bağlı olarak videoları çözümleme	Video görüntülerinden çıkarılan anlamsal özelliklerle ilişkilendirilebilen fiiller, nesneler vb. uygun sözdizimsel bileşenler belirlenir ve daha sonra doğal dil cümlelerine çevrilir. Ayrıca önerilen yöntemin performansını çeşitli deneylerle kanıtlanmıştır.
Ayers ve Shah	Şablon tabanlı yaklaşım	Kavram ile nesneleri izleme ve tanıma ile uygun kareleri seçme	Çalışma, içeriğe dayalı video sıkıştırması olan video dizilerinden bir dizi anahtar çerçeve oluşturmuştur. Birkaç video dizisi (manuel olarak oluşturulmuş) üzerinde test edilmiş ve iyi performans göstermiştir.
Gan ve diğ,	Dizi öğrenme yaklaşımı	COCO, Flickr30k ve Youtube2Text veri seti ve TSA, UKSB yöntemleri	Üç görsel altyazı veri kümesi üzerinde yapılan deneyler, önerilen yaklaşımın üstünlüğünü doğrulamıştır.
Xu ve diğ.	Dizi öğrenme yaklaşımı	MSVD, MSR-VTT ve MPII-MD veri seti ile TSA	Üç video altyazı veri setinde başarılı sonuçlar elde edilmiştir.
Song ve diğ	Dizi öğrenme yaklaşımı	TSA, UKSB yöntemleri ve MSR – VTT ve MSVD veri seti	Model, video altyazısının performansını artırabildiği ve farklı faktörleri dikkate alarak bir videoyu açıklamak için birden çok cümle oluşturabilmektedir.
Xu ve diğ.	Odak tabanlı yaklaşım	ESA, UKSB yöntemleri ve Flickr8k, Flickr30k ve MS COCO veri seti	Öğrenilen eşleştirmelerin insan sezgisine çok iyi karşılık geldiği gösterilmiştir.
Yu ve diğ	Odak tabanlı yaklaşım	Hollywood2 ve VAS veri seti GEAN (ESA, RGP) yöntemi	Daha iyi etiket üretmek için birden fazla odak tabanlı altyazı oluşturma yönteminin insan bakışının etkilerinden yararlandığı kanıtlanmıştır.
Li ve diğ	Odak tabanlı yaklaşım	UKSB yöntemi ve MSR-VTT veri seti	Geliştirilen modelin, etiket oluşturma sürecine geriye doğru akışı getirip kodlayıcı tarafından çıkarılan anlamsal parçaların daha iyi kullanılmasına yardımcı olabileceği doğrulanmıştır.

Gelecek çalışmalar için farklı video türlerinde, farklı ve küçük boyutlu ağların eğitilmesi ile video etiketleme yapılması önerilmektedir. Video etiketleme de üretilen cümlelerin verimliliğini arttırmak için de giriş videosu türünün, derin bir video sınıflandırıcısı tasarlanarak da belirlenebilir olması önerilmektedir.

Video etiketlemede kullanılan veri setlerinde Türkçe metin özetleme veri seti bulunmadığı için Türkçe üzerinde tamamlanmış bir çalışma mevcut değildir. Video etiketleme üzerine gerçekleştirilecek tez çalışmasında bu eksikliği gidermeye yönelik çalışmalar yapılacaktır.

Çıkar Çatışması (Conflict of Interest)

Yazarlar tarafından herhangi bir çıkar çatışması beyan edilmemiştir. No conflict of interest was declared by the authors.

Kaynaklar (References)

Aafaq, N., Akhtar, N., Liu, W., Mian, A., 2019. Empirical Autopsy of Deep Video Captioning Frameworks, arxiv.org/pdf/1911.09345v1

Amaresh, M. and Chitrakala, S., 2019. Video Captioning using Deep Learning: An Overview of Methods, Datasets and Metrics, International Conference on Communication and Signal Processing, India

Ayers, D. and Shah, M. 2001. Monitoring human behavior from video taken in an office environment. Image and Vision Computing, 19(12),833–846.

Baraldi, L., Grana, C. and Cucchiara, R., 2017. Hierarchical Boundary-Aware Neural Encoder for Video Captioning, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA

Carreira, J., Zisserman, A., 2017. Quo Vadis, Action Recognition? A New Model and The Kinetics Dataset. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA

Chen, D., Dolan, W., 2011, Collecting Highly Parallel Data For Paraphrase Evaluation. In ACL: Human Language Technologies, 1, 190-200.

Chen, Y., Zhang, W., Wang, S., Li, L., Huang, Q., 2018. Saliency-Based Spatiotemporal Attention for Video Captioning, International Conference on Multimedia Big Data (BigMM), Xi'an, China

Cho, K.,Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning Phrase Representations Using Rnn Encoder-Decoder for Statistical Machine Translation. arXiv:1406.1078,

Çtamak,B., Kuyu, M., Erdem, A., Erdem, E., 2019. MSVD-Turkish: A Large-Scale Dataset for Video Captioning in Turkish, 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Türkiye

Das, P., Xu, C., Doell, R. F., Corso. and J. J., 2013. A Thousand Frames in Just a Few Words: Lingual Description of Videos Through Latent Topics and Sparse Object Stitching. 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA

Ding, S., Qu, S., Xi, Y., Wan, S., 2019. A Long Video Captioning Generation Algorithm for Big Video Data Retrieval, Future Generation Computer Systems 93, 583–595

Elman, J. L., 1990. Finding Structure in time. Cognitive Science, 14(2), 179–211

Gan, Z., Gan, C., Hez, X., Puy, Y., Tranz, K., Gaoz, J.,Cariny, L., Dengz, L., 2017. Semantic Compositional Networks for Visual Captioning, arXiv:1611.08002v2,

Gao, L., Guo, Z., Zhang, H., Xu, X., Shen, H., 2017. Video Captioning With Attention-Based LSTM And Semantic Consistency, IEEE Transactions Multimedia, 19(9), 2045–2055

Gella, S., Lewis, M. and Rohrbach. M., 2018. A Dataset for Telling the Stories of Social Media Videos. In Proc of the 2018 Conference on Empirical Methods in Natural Language Processing. 968-974

Gers, F., Long Short-Term Memory in Recurrent Neural Networks, Ph.D. dissertation, Dept. Comput. Sci., Univ. Hannover, Hannover, Germany, 2001.

Graves, A., Jaitly, N., 2014. Towards End-To-End Speech Recognition With Recurrent Neural Networks.31st International Conference on Machine Learning (ICML-14). 1764- 1772.

Hochreiter S. and Schmidhuber, J., 1997. Long Short-Term Memory,Neural Computer, 9(8), 1735–1780.

Jegham, I., Khalifa, A.B., Alouani, I.,Mahjoub, M.A., 2020. Vision-Based Human Action Recognition: An Overview and Real World Challenges, Forensic Science International: Digital Investigation, 32, 200901

Kiros, R., Salakhutdinov, R., Zemel, R., 2014. Multimodal Neural Language Models, Proceedings of the 31st International Conference on Machine Learning, PMLR, 32(2), 595-603

Krishna, R., Hata, K., Ren, F., Fei-Fei, L. and Niebles. J. C., 2017.Dense-Captioning Events in Videos. In Proceedings of the IEEE International Conference on Computer Vision , Venice, Italy

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S.,Choi, Y., 2013. Babytalk: Understanding and Generating Simple Image Descriptions, IEEE Transactions On Pattern Analysis and Machine Intelligence, 35 (12)2891–2903.

Kuznetsova, P.,Ordonez, V., Berg, T.L., Choi, Y., 2014. TREETALK: Composition and Compression of Trees for Image Descriptions, Transactions of the Association for Computational Linguistics 2 (1), 351–362.

Kojima A., and Tamura, T.,2002, Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions, International Journal of Computer Vision 50(2), 171–184

Lavie, A., Agarwall, A., 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments,2007, Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic

Li, H., Song, D., Liao, L., Peng, C., 2019. Revnet: Brıng Revıewıng Into Vıdeo Captioning for a Better Descrıptıon, IEEE International Conference on Multimedia and Expo (ICME) Chiana

Li, S., Tao, Z., Li, K., Fu, Y., 2019. Visual to Text: Survey of Image and Video Captioning, IEEE Transactions on Emerging Topics in Computational Intelligence, 3(4), 297-312.

Li, W., Guo, D., Fang, X., (2018). Multimodal Architecture for Video Captioning with Memory Networks and an Attention Mechanism, Pattern Recognition Letters 105, 23–29

Lin, C.Y., 2004. ROUGE: A Package for Automatic Evaluation of Summaries, In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain

Liu, J., Wang, Z., Liu, H., 2020. HDS-SP: A Novel Descriptor For Skeleton-Based Human Action Recognition, Neurocomputing, 385,22-32

Ma, M., Wang, B., 2017. A Grey Relational Analysis based Evaluation Metric for Image Captioning and Video Captioning, 2017 International Conference on Grey Systems and Intelligent Services (GSIS), Stockholm, Sweden

Nabati, M., Behrad, A., 2020. Video Captioning Using Boosted And Parallel Long Short-Term Memory Networks, Computer Vision and Image Understanding, 190, 102840.

Nan, W., Zhigang, Z., Huan, L., Jingqi, M., Jiajun, Z., Guangxue, D., 2019. Gesture Recognition Based on Deep Learning in Complex Scenes, 2019 Chinese Control And Decision Conference (CCDC). Nanchang, China, China

Özer E.G., Karapınar İ.N., Başbuğ S., Turan S., Utku A., Akcayol M.A., 2020. Deep learning based new model for video captioning, International Journal of Advanced Computer Science and Applications, 11(3), 1-6.

Pan, P., Xu, Z., Yang, Y., Wu, F.,Zhuang, Y., 2016. Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA

Park, J., Song, C., Han. J-H., (2017), A Study of Evaluation Metrics and Datasets for Video Captioning. International Conference Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B. and Pinkal. M., 2013. Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics (TACL) 1, 25–36,

Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., Schiele, B., 2014. Coherent Multi-Sentence Video Description with Variable Level of Detail. Pattern Recognition, 184-195, Germany

Rohrbach, A., Rohrbach, M., Tandon, N. and Schiele. B., 2015. A Dataset for Movie Description. arXiv.org/abs/1501.02530

Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., Pinkal, M. and Schiele. B., 2012. Script Data for Attribute-Based Recognition of Composite Activities. In European Conference on Computer Vision, 144-157, Springer

Rohrbach, M., Qiu, W., Titov, I., 2013. Translating Video Content to Natural Language Descriptions, 2013. IEEE International Conference on Computer Vision, Sydney, NSW, Australia

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A., 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. arxiv.org/abs/1604.01753

Smirnov, E.A., Timoshenko, D.M., Andrianov, S.N., 2014. Comparison of Regularization Methods for ImageNet Classification with Deep Convolutional Neural Networks, AASRI Procedia, 6,89-94

Song, J., Guo, Y., Gao, L.,Li, X., Hanjalic, A.,Shen, H.T., (2015), From Deterministic to Generative: Multimodal Stochastic RNNs for Video Captioning, Journal Of Latex Class Files, 14(8),1-10.

Su, J., 2018. Study of Video Captioning Problem

Szegedy, C., Ioffe, S., Vanhoucke, S. Alemi, A., 2016. Inception-v4, Inception-Resnet And The Impact Of Residual Connections on Learning. /arxiv.org/abs/1602.07261

Şeker, A., Diri, B., Balık, H.H., 2017. Derin Öğrenme Yöntemleri ve Uygulamaları Hakkında Bir İnceleme, Gazi Mühendislik Bilimleri Dergisi, 3(3). 47-64

Tang, P., Wang, H., Kwong, S., 2017. G- MS2F: Googlenet Based Multi-Stage Feature Fusion Of Deep CNN For Scene Recognition, Neurocomputing, 225,188-197

Torabi, A., Pal, C., Larochelle, H. and Courville. A., 2015. Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research. arXiv:1503.01070

Trabelsi, A., Elouedi, Z., Lefevre, E., 2019. Decision Tree Classifiers For Evidential Attribute Values And Class Labels, Fuzzy Sets and Systems, 366,46-62

Tran, D., Bourdev, L., Fergus, R., Torresani,L., Paluri, M., 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. arxiv.org/abs/1412.0767

Unal, ME, Citamak, B., Yagcioglu, S., Erdem, A., Erdem, E., Cinbis, N.I., Cakici, R., 2016. Tasviret: A Benchmark Dataset for Automatic Turkish Description Generation From Images, 2016 24th Signal Processing and Communication Application Conference (SIU), Zonguldak, Turkey

Vedantam, R., Zitnick, C.L., Parikh, D., 2015. Cider: Consensus-Based İmage Description Evaluation, IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Boston, MA, USA.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K., 2015. Sequence To Sequence—Video To Text, 2015 IEEE International Conference Computer Vision, Santiago, Chile

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R. and K. Saenko. 2014. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. arXiv preprint arXiv:1412.4729, 2014

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and Tell: A Neural Image Captioning Generator, 28th IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA

Wang, B., Ma, L., Zhang, W., Liu, W., 2018. Reconstruction Network for Video Captioning, IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City USA

Wang, H., Gao, C., Han, Y., (2018). Sequence in Sequence for Video Captioning, Pattern Recognition Letters, 130, 327-334

Wu, A., Han, Y., Yang, Y., Hu, Q., Wu, F., 2019. Convolutional Reconstruction-to-Sequence for Video Captioning, IEEE Transactions on Circuits and Systems for Video Technology, 30(11), 4299 - 4308

Wu, X., Sahoo, D., Hoi, S.C.H., 2020. Recent Advances in Deep Learning For Object Detection, Neurocomputing, 396,39-64

Wu, Z., Yao,T., Fu, Y., Jiang, Y.-G., 2016. Deep Learning for Video Classification and Captioning, Frontiers of Multimedia Research 3-29

Xiao, H., Shi, J., 2019. Video Captioning with Adaptive Attention and Mixed Loss Optimization, IEEE Access, 7, 135757-13769.

Xu, J., Mei, T., Yao, T., Rui, Y., 2016. Msr-vtt: A Large Video Description Dataset for Bridging Video and Language. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., 2016. Show, Attend and Tell: Neural Image Captioning Generation with Visual Attention, Proceedings of the 32nd International Conference on Machine Learning, (PMLR) 37, 2048-2057,

Xu, N., Liu, A., 2018. Dual-Stream Recurrent Neural Network for Video Captioning, IEEE Transactions On Circuits And Systems For Video Technology, 29(8), 2482-2493

Yang, Y., Zhou, J., Jiangbo A., Bin, Y., Hanjalic, A., Shen, H.T., Ji, Y., 2018. Video Captioning by Adversarial LSTM, IEEE Transactions on Image Processing, 27(11), 5600-5611

Yang, Z., Yuan, Y., Wu, Y., Salakhutdinov, R., Cohen, W. W., 2016. Review Networks for Caption Generation, 30th International Conference on Neural Information Processing, Barcelona, SPAIN

Yang, Z., Yue, J., Li, Z., Zhu, L., 2018. Vegetable Image Retrieval with Fine-tuning VGG Model and Image Hash, IFAC-PapersOnLine, 51(17), 280-285.

Yao, L., Cho, K., Ballas, N., Pa´ı, C., Courville, A., 2015. Describing Videos By Exploiting Temporal Structure, International Conference on Computer Vision (ICCV), Santiago, Chile

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A., 2015. Describing Videos By Exploiting Temporal Structure, 2015 IEEE International Conference Computer Vision, Santiago, Chile

Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T., 2016. Boosting İmage Captioning with Attributes, in Proc. IEEE Int. Conference Computer Vision, Venice, Italy

Yingwei, P., Mei, T., Yao,T., Li, H., Rui. Y., 2015. Jointly Modeling Embedding and Translation to Bridge Video and Language. arxiv.org/abs/1505.01861

Yingwei, P., Yao, T., Li, H., Mei. T., 2016. Video Captioning with Transferred Semantic Attributes. arxiv.org/abs/1611.07675

You, Q., Jin, H., Wang, Z., Fang, C., Luo J., 2016. Image Captioning with Semantic Attention, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA

Yu, Y., Choi, J., Kim, Y., Yoo, K., Lee, S.-H., Kim, G., 2017. Supervising Neural Attention Models for Video Captioning by Human Gaze Data. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, Hawaii

Yuan, J., Xiong, H-C., Xiao, Y., Guan, W., Wang, M., Hong, R., Li, Z.Y., 2019. Gated CNN: Integrating Multi-Scale Feature Layers For Object Detection, Pattern Recognition 105, 107131

Zeng, K., Chen, T., Niebles, J. C., Sun, M., 2016. Title Generation for User Generated Videos. arxiv.org/abs/1608.07068

Zhao, H., Li, X., 2017. A Cost Sensitive Decision Tree Algorithm Based On Weighted Class Distribution With Batch Deleting Attribute Mechanism, Information Sciences, 378, 303-316

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A., 2014. Learning Deep Features For Scene Recognition Using Places Database, Proceedings of the Advances in Neural Information Processing Systems (NIPS). 487–495.

Zhou, L., Kalantidis, Y., Chen, X., Corso, J. J., Rohrbach, M., 2018. Grounded video description. arxiv.org/abs/1812.06587