Fırat Üniversitesi Müh. Bil. Dergisi
34(2), 829-837, 2022

Araştırma Makalesi
https://doi.org/10.35234/fumbd.1148296

Görüntülerden Derin Öğrenmeye Dayalı Otomatik Metin Çıkarma: Bir Görüntü Yakalama Sistemi

Zeynep KARACA[1*], Bihter DAŞ[2]
[2] Yazılım Mühendisliği, Teknoloji Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye
[*1] krczeynep1996@outlook.com, [2] bihterdas@gmail.com

(Geliş/Received: 25/07/2022; Kabul/Accepted:26/09/2022)

Öz: Bilgisayarlı görme ve doğal dil işlemenin çalışma alanlarından biri olan görüntüden metin üretme (image capturing), doğal bir dil kullanarak görüntü içeriğini otomatik olarak tanımlama görevidir. Bu çalışmada, MS COCO veri seti üzerinde İngilizce dili için encoder-decoder tekniğine dayalı bir otomatik altyazı oluşturma yaklaşımı önerilmiştir. Önerilen yaklaşımda, görüntü özniteliklerini çıkarmak için encoder olarak Evrişimli Sinir Ağı (CNN) mimarisi ve görüntülerden altyazı oluşturmak için bir decoder olarak Tekrarlayan Sinir Ağı (RNN) mimarisi kullanılmıştır. Önerilen yaklaşımın performansı BLEU, METEOR ve ROUGE_L değerlendirme kriterleri kullanılarak değerlendirilmiş ve her bir görüntüden 5 cümle elde edilmiştir. Deneysel sonuçlar, modelin görüntülerdeki nesneleri doğru bir şekilde algılamada tatmin edici olduğunu göstermektedir.
Anahtar kelimeler: Doğal Dil İşleme, Görüntüden Metin Üretme, Metin Tarama, Metin Tahmini, Derin Öğrenme

Automatic Text Extraction Based on Deep Learning from Images: An Image Capture System

Abstract: Image capturing, one of the working areas of computer vision and natural language processing, is the task of automatically identifying image content using a natural language. In this study, an encoder-decoder technique-based automatic captioning approach is proposed for the English language on the MS COCO dataset. In the proposed approach, Convolutional Neural Network (CNN) architecture is used as an encoder for extracting image features, and Recurrent Neural Network (RNN) architecture is used as a decoder for creating subtitles from images. The performance of the proposed approach was evaluated using BLEU, METEOR, and ROUGE_L evaluation criteria, and 5 sentences were obtained from each image. The experiment results show that the model is satisfactory in correctly detecting the objects in the images.

Keywords: Natural Language Processing, Image Capturing, Text detection, Text Prediction, Deep Learning

1. Giriş

Dijital ortamlardaki görüntü sayısındaki artış her geçen gün artmaktadır. Bu artış, görüntü işleme, örüntü tanıma ve görüntü altyazısı üzerine yapılan çalışmalarda büyük ilgi uyandırmaktadır. Doğal bir dil kullanarak bir görüntünün içeriğinin otomatik olarak belirlenmesi ve insan benzeri bir doğal açıklamanın otomatik olarak çıkarılması, görüntüden metin üretilmesi olarak adlandırılmaktadır [1,2]. Bir görüntünün metinsel tanımını oluşturma, görüntüdeki nesneleri tanıma, anlamsal ilişkiler, arka plan sahnesini anlama ve bu bilgiyi sözdizimsel olarak doğru cümlelere dönüştürme süreci image capturing çalışma konularıdır. Verilen görüntüyü anlamsal olarak en iyi tanımlayan cümleyi üretmek amaçlanmaktadır [3]. Resim yazısı sorunu hem bilgisayarlı görmenin hem de doğal dil işlemenin bir parçası olarak görülmektedir [4,5]. Bu bağlamda metin üretme (image capturing) için yalnızca görüntünün içeriğini anlamak değil, aynı zamanda söz dizimsel ve anlamsal olarak doğru bir alt başlık bulmak, nesneler arasındaki anlamsal ilişkileri çıkarmak, resimdeki detayları yakalamak, konuyu anlamak için yeterlidir. Arka plan sahnesi ve bu bilgiyi insanların doğal bir konuşması olarak tanımlamak oldukça önemlidir [6,7]. Görüntüden metin üretme işleminde sistem görüntüdeki nesneleri tanımlar, görüntünün göze çarpan özelliklerini bulmaya çalışır ve bilgileri çıkardıktan sonra görüntü için en anlamsal ve sözdizimsel olarak en uygun ve kısa özet cümleyi üretmelidir [8]. Oluşturulan açıklama bağımsız ve anlamsal olarak doğruysa, kullanıcıya anlamlı bilgileri belirsizlik olmadan iletecektir [9]. Bir görüntüyü anlama ve o görüntüyle ilgili metin oluşturma süreci bilgisayarlar için çok karmaşıktır. Metin üretilmesi sistemi, görüntüyü en iyi tanımlayan cümleyi oluşturmak için yapay zekanın doğal dil işleme ve bilgisayarla görme yeteneklerini kullanmaktadır. Bunu yaparken sistem, görüntü içeriğini en iyi şekilde anlamak için görüntü içeriğini cümlelere çevirmek için bir kod çözücü kullanmaktadır. Derin öğrenme yöntemlerinden olan Evrişimli Sinir Ağları (ESA) ve Tekrarlayan Sinir Ağları (TSA) bu alanda yaygın olarak kullanılmaktadır. ESA yöntemleri, görüntü altyazıları için kodlayıcı olarak

* Sorumlu yazar: krczeynep1996@outlook.com. Yazarların ORCID Numarası: [1] 0000-0002-7751-8567, [2] 0000-0002-2498-3297

kullanılırken [10-12], TSA, görüntü içeriğini açıklama semantiği olarak veren bir kod çözücü olarak kullanılır [13,14]. TSA, anlamlı ve tutarlı cümleler oluşturmak için görsel semantik birimlerden görüntüler alır ve ardından görsel birimleri metinsel kelimelere dönüştürmek için generator tasarlar. Bu çalışma, resim yazısı problemlerini çözmek için sinir ağı tabanlı bir model önermektedir. Önerilen model, bir kodlayıcı ve kod çözücü kullanan bir sistemi tanıtmaktadır. Sistem, görüntülerden özellikleri çıkarmak için ESA yöntemini ve ilgili alt başlığı ve metni oluşturmak için TSA yöntemini kullanır.

Makalenin geri kalanı şu şekilde düzenlenmiştir. Bölüm 2, literatürde bu alanda yapılan çalışmaların kısa bir özetini sunmaktadır. Üçüncü bölümde deneysel yöntemde kullanılan veri seti, ESA modeli, kodlayıcı-kod çözücü ve performans değerlendirme kriterlerinden bahsedilmiştir. 4. bölümde önerilen yöntemle ilgili bulgulara yer verilmiş, çalışmanın genel katkıları Sonuç bölümünde sunulmuştur.

2. Literatür taraması

Literatürde otomatik olarak görüntü alt yazısı oluşturmaya yönelik çeşitli çalışmalar bulunmaktadır. Bu çalışmaların çoğunda makine öğrenmesi ve derin öğrenme yöntemleri kullanılmıştır. Bai et al. ESA+ESA tabanlı altyazıyı geliştirmek için MS COCO veri kümesinde Multi-model Graphical Convolutional Network (MGCN) yöntemini kullandı. Arkaplan grafiğini işlemek ve görüntüler üzerinde özellik çıkarımı için nesneler arasındaki görsel ilişkileri göstermek için bir iyileştirme yöntemi geliştirdiler [15]. Kılıçkaya et al. Bir görüntünün içerdiği nesnelere bağlı olarak görüntünün daha yüksek düzeyde bir temsilini elde etmek ve görüntünün tanımını ondan otomatik olarak çıkarmak için Im2Text yöntemini önermişlerdir. Veri seti olarak Pascal cümleleri kullanılmıştır [16]. Lu et al. güzel sanatlara yönelik resim içeriğini tanımlamak için bir model önermişlerdir. Önerilen yöntemin performansı, genel bir altyazı veri seti olan ArtCap veri seti [17] üzerinde değerlendirilmiştir. Yang et al. insan davranışını bir görüntüde tanımlamak için yeni bir konu, insan merkezli altyazı oluşturmayı amaçlamıştır. İnsanlara odaklanan yeni bir COCO veri seti oluşturmuşlardır. Çalışmada, görüntü özelliklerini hiyerarşik hale getirmek için yeni bir üç dallı başlık modeli kullandılar. Çalışma, mevcut çalışmalara göre bazı iyileştirmeler yapmış olsa da daha ayrıntılı açıklamalar üretmek için yeterli değildi [18]. Agrawal ve ark. bir kodlayıcıya ve dikkat tabanlı bir kod çözücüye dayalı bir altyazı oluşturucu modeli önerdi. Çalışmanın Inception V3 modelinde, ESA kodlayıcıdan sonra tanıtılan dikkat mekanizması ile giriş sahnesi görüntüsündeki en alakalı bilgilere dikkat edildi. Böylece kod çözücü, altyazıyı oluşturmak için görüntünün yalnızca belirli kısımlarını kullandı. Veri seti olarak MS-COCO kullanıldı. Önerilen model, altyazıyı geleneksel kodlayıcı-kod çözücü tabanlı modelden daha fazla geliştirdi [19]. Li ve ark. resim yazısı için yeni çok seviyeli benzerlik odaklı anlamsal eşleştirme yöntemi önerdiler. Çalışmada, görsel ve metinsel anlam birimleri arasındaki ilişkiyi ölçmek için yerel bir anlamsal benzerlik değerlendirme mekanizması tasarlanmıştır. Yöntem, MS COCO veri kümesi üzerinde test edilmiştir. Bu amaçla görüntünün görsel anlam birimleri ile oluşturulan cümlenin metinsel anlam birimleri eş zamanlı olarak çıkarılmıştır [20]. Tablo 1, görüntüden metin üretme için mevcut çalışmaların bazılarını listeler.

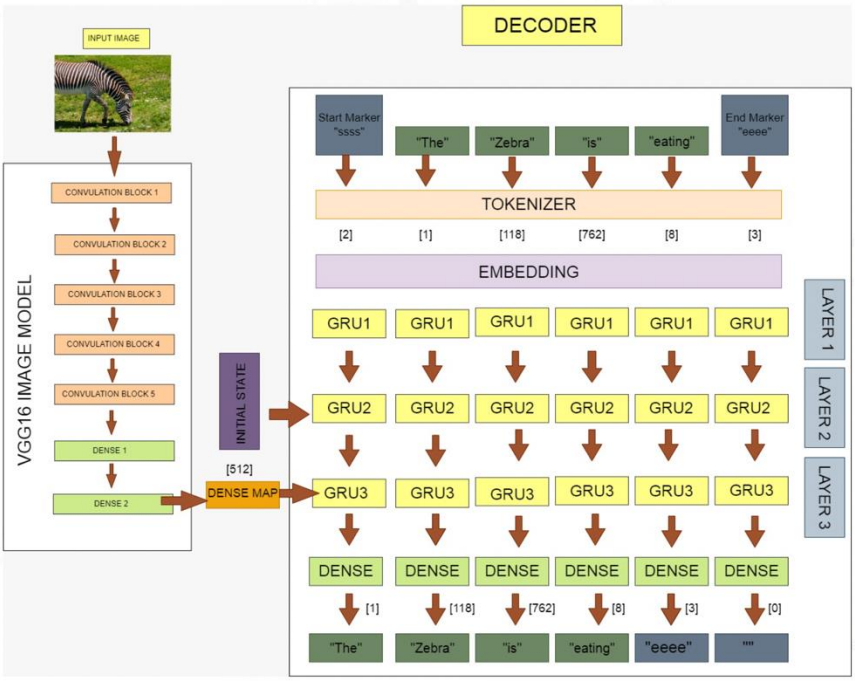
Tablo 1. Görüntüden metin üretimi için yapılmış çalışmalardan bazıları

Yazar	Veriseti	Yöntem	Sonuçlar
Bai ve ark. [15]	MC-COCO	MGCN	BLUE2=0.567 METEOR=0.281
Kilickaya ve ark. [16]	Pascal cümleleri	Im2Text	BLUE1=0.11
Lu ve ark. [17]	MSCOCO+ArtCap	Faster R-CNN	BLUE1=0.508 METEOR=0.1317
Yang ve ark. [18]	HS-COCO	HCCM	BLUE1=0.839 METEOR=0.304
Agrawal ve ark. [19]	MS-COCO	Inception V3	-----
Li ve ark. [20]	MS-COCO	CNN+RNN	-----

3. Materyal ve Yöntem

Bu çalışmada otomatik resim alt yazıları oluşturan bir model sunulmuştur. Çalışmadaki deneysel uygulamada, görüntülerin özniteliklerinin çıkarılmasından sorumlu ESA mimarisine sahip bir kodlayıcı ve altyazıların

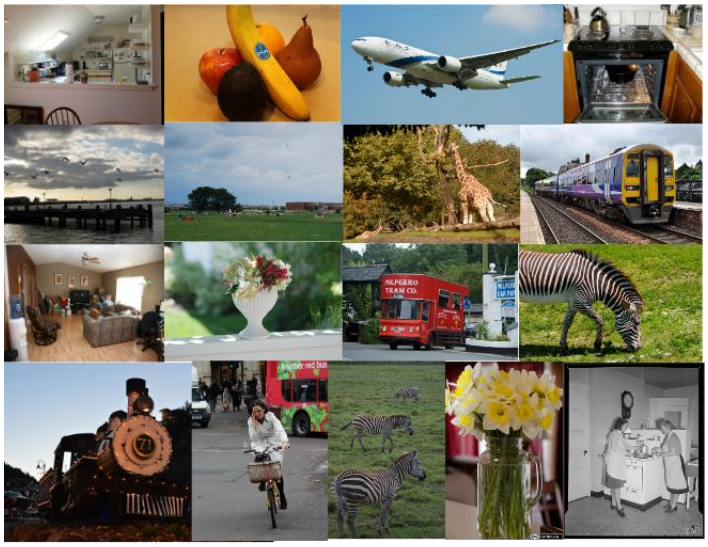
oluşturulmasından sorumlu TSA mimarisine sahip bir kod çözücü kullanılmıştır. Bu birleşik sistemle, ingilizce metin üretme kodlayıcı-kod çözücü modeli, MS-COCO veri kümesi üzerinde test edilir. Bu model, görüntüleri tanımada çok iyi olan VGG16 veri modelini kullanmaktadır. VGG16 modeline görüntü verildikten sonra 5 convolution bloğu ve 2 dense katmanından geçirilerek düşünce vektörü (thought vector) üretilir ve görüntünün içeriği bu thought vektörüne yazılır. Görüntünün içeriğini açıklayan thought vektörü ve doğru cümle, kod çözücüye girdi olarak verilir. Kod çözücüye verilen cümlenin başına başlangıç belirteci, sonuna ise bitiş belirteci eklenir. Kod çözücü, çıktı olarak kelimeler üretir. Son cümle, bu kelimelerin gerçekte olması gereken kelimelerle karşılaştırılmasıyla üretilir. Önerilen sistemin mimarisi Şekil 1'de verilmiştir.



Şekil 1. Önerilen sistemin mimarisi

3.1 Veri seti

Bu çalışmada, Tsung-Yi Lin ve diğerleri tarafından oluşturulan Microsoft Common Objects in Context (MS COCO) veri seti [10] için kullanıldı. MS COCO veri kümesi, çoğunlukla sinir ağları olmak üzere bilgisayarla görme için zorlu, yüksek kaliteli görsel veri kümeleri içerir. Büyük veri seti, doğal bir bağlamda ortak nesnelerin ve günlük sahnelerin açıklamalı fotoğraflarından oluşmaktadır. Bu nesneler, "sandalye" veya "muz" gibi önceden tanımlanmış sınıflar kullanılarak etiketlenmektedir. Çalışmada eğitim için 118.287 veri kullanılırken, test için 5000 görüntü ve toplam 123.287 görüntü kullanılmıştır. Şekil 2, MS COCO veri tabanından bazı etiketlenmiş görüntüleri göstermektedir.



Şekil 2. MS COCO veri setinden bazı orijinal görüntüler

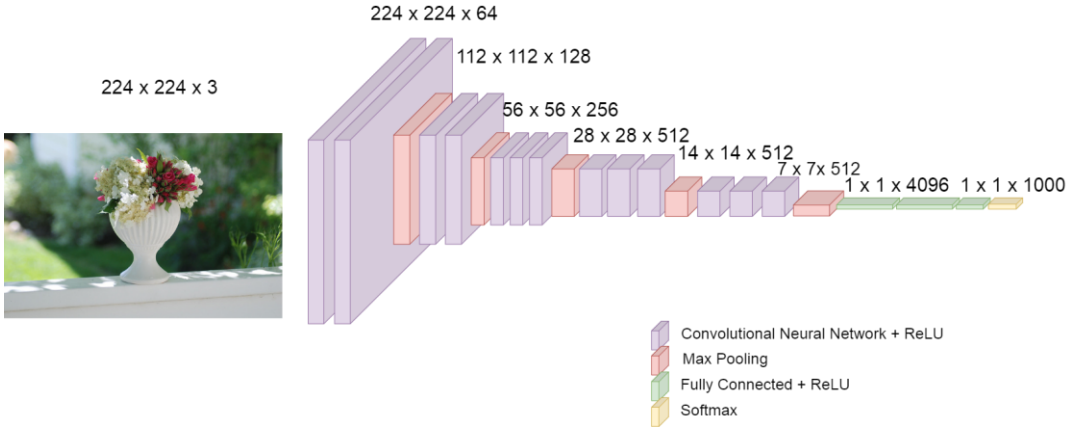
3.2 Evrişimsel Sinir Ağları

Evrişimsel Sinir ağları (ESA-Convolutional Neural Networks-CNN), görüntü tanıma ve sınıflandırmada kullanılan bir derin öğrenme ağları sınıfıdır. ESA yönteminde girdi olarak verilen görüntünün bilgisayarlar tarafından tanınması ve işlenmesi için matris formatına dönüştürülmesi gerekmektedir. Aynı görünen iki görüntü arasındaki farkı söylemek oldukça zordur. Ancak matris formatına dönüştürülen görüntüler arasındaki detaylar kolaylıkla ayırt edilebilmektedir. Matrislerdeki farklılıklara göre oluşturulan sistem bu şekilde hangi görüntünün hangi etikete ait olduğunu belirlemektedir. Kenar algılama, giriş görüntüsünün yüksek frekanslı bölgelerini temsil etmektedir. Bu öznitelik bilgisini elde etmek için yatay ve dikey olmak üzere iki filtre kullanılır. Ortaya çıkan çıktı, görüntünün kenar bilgilerini göstermektedir. Kenarlar genellikle bir evrişimli ağ modelinin ilk katmanlarında hesaplanmaktadır [21,22].

3.2.1 VVG16 modeli

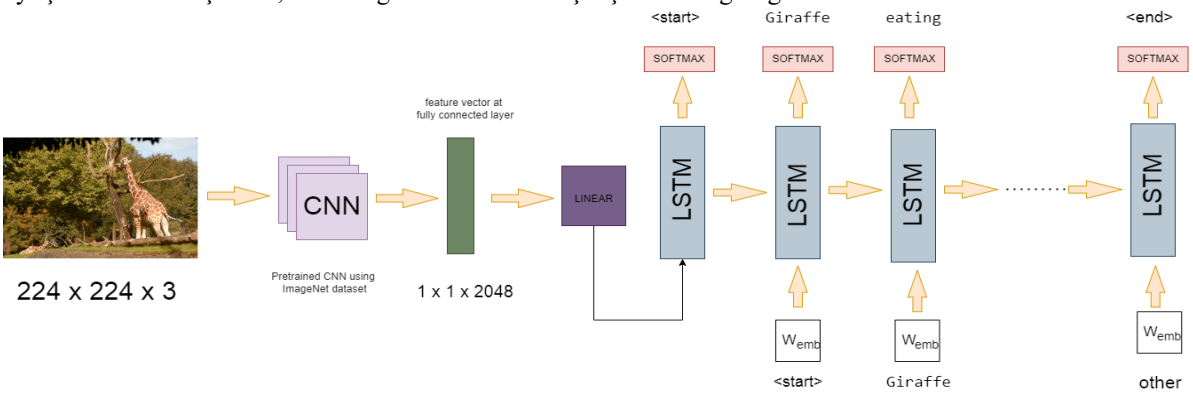
VGG16, görsel nesne tanıma araştırmalarında yaygın olarak kullanılan büyük bir CNN mimarisidir. VGG16'daki 16 sayısı, ağırlıklarla birlikte 16 katmanı olduğu anlamına gelmektedir. Bu, yaklaşık 138 milyon parametreye sahip büyük bir ağdır [23]. Eğitim sırasında, ImageNet veri kümesi sabit boyutlu 224 x 224 RGB görüntü içermektedir. Yani girdi olarak (224,224.3) tensor bulunmaktadır. Eğitim setinde hesaplanan ortalama RGB değerinin her pikselden çıkarılması burada yapılan tek ön işlemedir [24].

Model 13 evrişim katmanına, ardından 5 max-pooling katmanına, üç tam bağlantılı katmana (full connected layers) ve son olarak softmax katmana sahiptir. Ağ, giriş boyutlarına, 3*3 filtre boyutuna ve aynı dolguya sahip 64 kanala sahiptir. Daha sonra max-pooling (2.2) katmanından sonra, 256 filtre boyutuna (3.3) evrişim katmanına sahip iki katman vardır. Önceki katmanda olduğu gibi, max-pooling (2,2) eklenmektedir. Filtre boyutu (3,3) ve 256 filtre ile 2 evrişim katmanı eklenmektedir. Evrişim ve maksimum havuzlama katmanlarından sonra bir (7,7,512) özellik haritası elde edilir. Daha sonra, tam olarak bağlı 3 katmandan sonra, sınıflandırma vektörünü normalleştirmek için softmax katmanına geçilir. Sınıflandırma vektörünün ilk 5 kategorisinin çıktısından sonra, tüm gizli katmanlar aktivasyon fonksiyonu olarak ReLU'yu kullanmaktadırlar. ReLU, daha hızlı öğrenme ile sonuçlandığından ve ayrıca kaybolan gradyan probleminin olasılığını azalttığı için hesaplama açısından daha verimlidir [25]. Şekil 3, kullanılan VGG16 mimarisinin yapısını göstermektedir.



Şekil 3. Kullanılan VGG16 mimarisi

ESA mimarisine dayalı modelde, giriş görüntüsündeki ana nesneler için nesne algılama modeli tarafından etiketler ve sınırlayıcı kutular oluşturulur. Nesne özellikleri, bu sınırlayıcı kutu kullanılarak nesne algılama modelindeki ara katmanlardan çıkarılır ve nesneleri tanımlayan metinsel sözcükleri elde etmek için cümleler ayrıştırılmaktadır. Şekil 4, nesne algılama modelinin çalışma mantığını göstermektedir.



Şekil 4. Nesne algılama modeli

3.3 Kodlayıcı-Kod çözücü

Giriş görüntüsü, VGG16 modeli kullanılarak sisteme tanıtılmaktadır. Görüntü sırasıyla 5 evrişim katmanından ve ardından 2 yoğunluk katmanından (dense layers) geçmektedir. Son olarak görüntü çıktı katmanından geçmeden thought vektörüne gönderilmektedir. Çünkü sonuç olarak bir tahmine değil, bir thought vektörüne ihtiyaç vardır. Kod çözücüye (decoder) verilecek cümlenin başına "başlangıç token", sonuna "bitiş token" eklenir. Thought vektörünü ve resim içeriğini tanımlayan doğru cümle Decodera'ya girdi olarak verilir. Decoder çıktı olarak kelimeleri üretecek ve üretilen bu kelimeler olması gereken kelimelerle karşılaştırılarak sinir ağı oluşturulacaktır.

4. Deneysel sonuçlar ve tartışma

Literatürde metin üretilmesine yönelik çalışmaların başarısını değerlendirmek için kullanılan çeşitli otomatik değerlendirme kriterleri bulunmaktadır. BLUE, ROUGE, METEOR ve CIDer, görüntülerden metin üretilmesi için en çok kullanılan değerlendirme kriterleridir. Bu kriterlerin değerleri 0-1 arasında değişmektedir. 1'e yakın değerler yapay zekanın insan çevirisi kadar başarılı olduğunu, 0'a yakın değerler ise başarı oranının çok düşük olduğunu göstermektedir. [10]'da metin üretme için otomatik metrikler üzerine kapsamlı bir araştırma yapılmış ve

bu metriklerin nasıl yorumlanacağı ayrıntılı olarak açıklanmıştır. Bu çalışmada kullanılan kriterler BLUE-1, BLUE-2, BLUE-3, BLUE-4, METEOR ve ROUGE_L'dir.

BLUE: Makine tarafından çevrilmiş metnin bir dizi yüksek kaliteli referans çevirisine benzerliğini ölçen sıfır ile bir arasında bir metriktir. 0 değeri, makine tarafından çevrilen çıktının referans çevirisiyle eşleşmediği anlamına gelirken, 1 değeri, referans çevirileriyle mükemmel bir eşleşme olduğu anlamına gelmektedir [26].

METEOR: Makine çevirisi değerlendirmesi için kullanılan ve insan muhakemesi ile daha iyi bir korelasyona sahip başka bir ölçüm metriğidir [27].

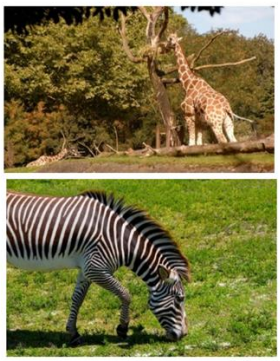
ROUGE: Bu metrik hatırlamaya dayalıdır ve çoğunlukla özet değerlendirme için kullanılmaktadır. Geri çağırmayı hesaplamak için kullanılan özelliğe bağlı olarak ROUGE, ROUGE-N, ROUGE-L, ROUGE-W ve ROUGE-S gibi birçok türde kullanılabilmektedir [28].

Önerilen yaklaşım kullanılarak MSCOCO veri setinde metin üretme için elde edilen performans değerlendirme sonuçları Tablo 2'de gösterilmektedir.

Tablo 2. Değerlendirme sonuçları

Algoritma	Model Performansı
BLUE-1	0,641
BLUE-2	0,420
BLUE-3	0,334
BLUE-4	0,185
METEOR	0,056
ROUGE_L	0,117

Önerilen yaklaşım kullanılarak elde edilen görüntüler için ingilizce altyazı sonuçları Şekil 5 ve Şekil 6'da gösterilmektedir. Literatürde görüntü yakalamaya yönelik 2 cümleli ve 3 cümleli metin üretimi yaygındır. Diğer çalışmalardan farklı olarak bu çalışmada önerilen yaklaşımla hem az nesne içeren görüntülerden hem de çok nesne içeren görüntülerden İngilizce olarak 5 cümle üretilmeye çalışılmıştır.



A giraffe eating food from the top of the tree.
A giraffe standing up nearby a tree.
A giraffe mother with its baby in the forest.
Two giraffes standing in a tree filled area.
A giraffe standing next to a forest filled with trees.

A zebra grazing on lush green grass in a field.
Zebra reaching its head down to ground where grass is.
The zebra is eating grass in the sun.
A lone zebra grazing in some green grass.
A zebra grazing on grass in a green open field.

Şekil 5. Az nesne içeren görüntüler için metin üretme sonuçları

An unusual looking red bus going down a road.
A bus driving down the street next to some buildings and trees.
A red tram carrying a group of people on a tour.
A red double decker vehicle driving down a street.
A big red bus that is driving down the road.

A living room filled with furniture on top of a hard wood floor.
A living room with big couches and a ceiling fan.
A living room filled with couches chairs and a water cooler sitting on a hard wood floor.
Living room with wood floor TV couches and chair.
A very big nice-looking room with a bright window.

A watery glass jar full of blooming flowers.
A vase filled with yellow and white flowers.
A glass jar is filling with white and yellow flowers.
A glass vase is holding a bunch of flowers.
Tall glass mug shaped vase with yellow flowers.

Şekil 6. Çok nesneli görüntüler için metin üretme sonuçları

Önerilen yaklaşımın performans sonuçları Tablo 2'ye göre analiz edildiğinde BLUE-1 sonucunun tatmin edici olduğu ancak METEOR ve ROUGE_L metriklerinin istenilen düzeyde olmadığı görülmektedir. Sonuçlar kendi içinde karşılaştırıldığında BLUE-1 ve BLUE-2'nin diğer metriklere göre daha iyi sonuçlar verdiği görülmektedir. Bunun olası nedenleri, çok nesneli görüntülerde önerilen VGG16 mimarisinin yeterli performans gösterememesi, bu mimarinin az nesneli görüntülerde daha başarılı olmasıdır. Şekil 5 ve Şekil 6 incelendiğinde, az nesneli görüntüler ve çok nesneli daha karmaşık nesneler için önerilen yaklaşım, görüntüdeki nesneleri algılamada kısmen yeterli olsa da nesneler arasında ilişki kurmada zayıf kalmaktadır. Ayrıca sonuçlar incelendiğinde önerilen yaklaşımın dilbilgisi açısından doğru cümleler kurma konusunda bazı eksiklikleri bulunmaktadır. Örneğin, Şekil 6'daki "A living room filled with couches chairs and a water cooler sitting on a hard wood floor" cümlesinde olduğu gibi, kanepeler (sofas), su soğutucusu (water cooler) ve sert ahşap (hard wood) zemin algılanmışken resimdeki kutular (boxes) ve bilgisayar (computer) algılanmamıştır.

5. Sonuç

Bu çalışmada, resim alt yazıları elde etmek için literatürde var olan yaklaşımlar incelenmiş ve ESA+TSA tabanlı bir görüntülerden metin üretilmesi yaklaşımı önerilmiştir. Önerilen yaklaşımda, MSCOCO veri setinden alınan görüntüler girdi olarak modele verilmiş ve çıktı olarak o görüntüyü açıklayan bir cümle elde edilmesi amaçlanmaktadır. Görüntüler VGG16 modeline verilmiş ve model çıktısından sonra elde edilen veriler thought vektörüne yazılmıştır. Düşünce vektörünü ve görüntünün içeriğini açıklayan doğru cümle, kod çözücüye girdi olarak verilir. Çalışmanın sonunda her bir görsel için 5 adet cümle üretilmiştir. Elde edilen cümleler görüntüdeki nesneleri algılamıştır. Önerilen yaklaşımın, görüntülerdeki nesneleri algılamak için tatmin edici cümleler ürettiği görülmüştür. Ancak üretilen cümlelerin performanslarını ölçmek için kullandığımız metriklerden sadece BLUE-1 performans ölçeği 1 değerine yakın çıkmıştır. Diğer performans metriklerinin 1'e yakın çıkmadığı görülmektedir. Bu çalışmamızın dezavantajlarındandır. Bundan sonraki çalışmalarda daha detaylı nesne algılama sürecinin elde edilmesi, görüntü başlığı oluşturulurken arka plan algısının iyileştirilmesi ve görüntüyü açıklayan cümle oluşturulurken daha anlamlı ve uzun cümlelerin elde edilmesi planlanmaktadır.

KAYNAKLAR

[1] C. P. Chaudhari ve S. Devane, "Capturing Semantic Knowledge In Object Localization In Captioning Images", içinde 2021 International Conference on Communication information and Computing Technology (ICCICT), Haz. 2021, ss. 1-4. doi: 10.1109/ICCICT50803.2021.9510175.

[2] A. U. Dey, S. K. Ghosh, E. Valveny, and G. Harit, "Beyond visual semantics: Exploring the role of scene text in image understanding," *Pattern Recognition Letters*, vol. 149, pp. 164–171, Sep. 2021, doi: 10.1016/j.patrec.2021.06.011.

[3] R. A. Davis, Z. Xiao, and X. Qi, "Capturing semantic relationship among images in clusters for efficient content-based image retrieval," in *2012 19th IEEE International Conference on Image Processing*, Sep. 2012, pp. 1953–1956. doi: 10.1109/ICIP.2012.6467269.

[4] C. Bai, A. Zheng, Y. Huang, X. Pan, ve N. Chen, "Boosting convolutional image captioning with semantic content and visual relationship", *Displays*, c. 70, s. 102069, Ara. 2021, doi: 10.1016/j.displa.2021.102069.

[5] C. Wang, Y. Shen, and L. Ji, "Geometry Attention Transformer with position-aware LSTMs for image captioning," *Expert Systems with Applications*, vol. 201, p. 117174, Sep. 2022, doi: 10.1016/j.eswa.2022.117174.

[6] S. Wang *et al.*, "Multi-label semantic feature fusion for remote sensing image captioning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, pp. 1–18, Feb. 2022, doi: 10.1016/j.isprsjprs.2021.11.020.

[7] C. Wu, J. Wu, H. Cao, Y. Wei, and L. Wang, "Dual-View Semantic Inference Network for image-text matching," *Neurocomputing*, vol. 426, pp. 47–57, Feb. 2021, doi:10.1016/j.neucom.2020.09.079.

[8] Y. Wang, Y. Xie, J. Zeng, H. Wang, L. Fan, and Y. Song, "Cross-modal fusion for multi-label image classification with attention mechanism," *Computers and Electrical Engineering*, vol. 101, p. 108002, Jul. 2022, doi: 10.1016/j.compeleceng.2022.108002.

[9] S. Zhao, L. Li, and H. Peng, "Aligned visual semantic scene graph for image captioning," *Displays*, vol. 74, p. 102210, Sep. 2022, doi: 10.1016/j.displa.2022.102210.

[10] E. Battini Sonmez, T. Yildiz, B. D. Yilmaz, and A. E. Demir, "Türkçe dilinde görüntü altyazısı: veritabanı ve model," *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, Jul. 2020, doi: 10.17341/gazimmfd.597089.

[11] Y. Wu, W. Liu, and S. Wan, "Multiple attention encoded cascade R-CNN for scene text detection," *Journal of Visual Communication and Image Representation*, vol. 80, p. 103261, Oct. 2021, doi: 10.1016/j.jvcir.2021.103261.

[12] M. Mustafa, "An energy efficient and improved language translator with cnn based deep encoder and decoder," *Materials Today: Proceedings*, Feb. 2021, doi: 10.1016/j.matpr.2020.12.1204.

[13] J. Chen and H. Zhuge, "Extractive summarization of documents with images based on multi-modal RNN," *Future Generation Computer Systems*, vol. 99, pp. 186–196, Oct. 2019, doi: 10.1016/j.future.2019.04.045.

[14] H. Zhan, S. Lyu, Y. Lu, and U. Pal, "DenseNet-CTC: An end-to-end RNN-free architecture for context-free string recognition," *Computer Vision and Image Understanding*, vol. 204, p. 103168, Mar. 2021, doi: 10.1016/j.cviu.2021.103168.

[15] C. Bai, A. Zheng, Y. Huang, X. Pan, and N. Chen, "Boosting convolutional image captioning with semantic content and visual relationship," *Displays*, vol. 70, p. 102069, Dec. 2021, doi: 10.1016/j.displa.2021.102069.

[16] M. Kılıçkaya, E. Erdem, A. Erdem, N. İ. Cinbiş, and R. Çakıcı, "Data-driven image captioning with meta-class based retrieval," in *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, Apr. 2014, pp. 1922–1925. doi: 10.1109/SIU.2014.6830631.

[17] Y. Lu, C. Guo, X. Dai, and F.-Y. Wang, "Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training," *Neurocomputing*, vol. 490, pp. 163–180, Jun. 2022, doi: 10.1016/j.neucom.2022.01.068.

[18] Z. Yang, P. Wang, T. Chu, and J. Yang, "Human-Centric Image Captioning," *Pattern Recognition*, vol. 126, p. 108545, Jun. 2022, doi: 10.1016/j.patcog.2022.108545.

[19] V. Agrawal, S. Dhekane, N. Tuniya, and V. Vyas, "Image Caption Generator Using Attention Mechanism," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Jul. 2021, pp. 1–6. doi: 10.1109/ICCCNT51525.2021.9579967.

[20] J. Li, N. Xu, W. Nie, ve S. Zhang, "Image Captioning with multi-level similarity-guided semantic matching", *Visual Informatics*, c. 5, sy 4, ss. 41-48, Ara. 2021, doi: 10.1016/j.visinf.2021.11.003.

[21] A. Shokraei Fard, D. C. Reutens, and V. Vegh, "From CNNs to GANs for cross-modality medical image estimation," *Computers in Biology and Medicine*, vol. 146, p. 105556, Jul. 2022, doi: 10.1016/j.compbiomed.2022.105556.

[22] E. Paul and S. R.s., "Modified convolutional neural network with pseudo-CNN for removing nonlinear noise in digital images," *Displays*, vol. 74, p. 102258, Sep. 2022, doi: 10.1016/j.displa.2022.102258.

[23] L.-Y. Ye, X.-Y. Miao, W.-S. Cai, and W.-J. Xu, "Medical image diagnosis of prostate tumor based on PSP-Net+VGG16 deep learning network," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106770, Jun. 2022, doi: 10.1016/j.cmpb.2022.106770.

[24] A. A. Pravitasari, N. Iriawan, U. S. Nuraini, and D. A. Rasyid, "12 - On comparing optimizer of UNet-VGG16 architecture for brain tumor image segmentation," in *Brain Tumor MRI Image Segmentation Using Deep Learning Techniques*, J. Chaki, Ed. Academic Press, 2022, pp. 197–215. doi: 10.1016/B978-0-323-91171-9.00004-1.

[25] J. Brownlee, "A Gentle Introduction to the Rectified Linear Unit (ReLU)," *Machine Learning Mastery*, Jan. 08, 2019. https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/ (accessed Jun. 29, 2022).

[26] 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*, Philadelphia, Pennsylvania, USA, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

[27] Banerjee, S. and Lavie, A. (2005) "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" in Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or

Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005

[28] D. Raj, "Metrics for NLG evaluation," *Explorations in Language and Learning*, Sep. 16, 2017. https://medium.com/explorations-in-language-and-learning/metrics-for-nlg-evaluation-c89b6a781054 (Erişim tarihi: 16 Haziran, 2022).