

Bilgisayarla Görme Projesi Raporu

Proje Konusu: Image Captioning

Dersin Yürütücüsü: Oğuz Altun

Anlatım Linki

Muhammed Kayra Bulut-23501059

Kasım 2023

1 Giriş

1.1 Projenin Amacı

Bu projenin temel amacı, derin öğrenme ve bilgisayarla görme tekniklerini kullanarak otomatik görüntü alt yazılandırma sistemi geliştirmektir. Sistem, verilen bir görüntüyü analiz edecek ve içeriğini açıklayan kısa ve anlamlı bir metin üretecektir. Bu teknolojinin potansiyel uygulama alanları arasında görsel içeriklerin aranabilirliğinin artırılması, görsel engelliler için erişilebilirlik çözümleri ve otomatik içerik üretimi bulunmaktadır.

1.2 Görüntü İşleme (Image Processing) Nedir?

Görüntü işleme, dijital görüntüler üzerinde çeşitli işlemler yaparak bu görüntüleri iyileştirmek, analiz etmek ve bazı sonuçlar çıkarmak için kullanılan bir bilgisayar bilimi dalıdır. Bu alandaki temel işlemler arasında görüntü geliştirme ve restorasyon, gürültünün azaltılması, kenar tespiti ve görüntü sınıflandırma yer alır [7]. Görüntü işleme, çeşitli alanlarda, özellikle tıbbi görüntüleme, uydu görüntüleme, robotik ve güvenlik sistemlerinde geniş bir uygulama yelpazesi sunar [6]. Bilgisayarla görme (computer vision), görüntü işlemenin ötesine geçerek, görüntülerdeki nesneleri, sahneleri ve etkinlikleri anlamaya ve yorumlamaya odaklanır. Görüntü işlemenin en ilgi çekici uygulamalarından biri, otomatik görüntü alt yazma (image captioning) olup, bu alanda yapılan çalışmalar, derin öğrenme tekniklerinin kullanılmasını içerir [8] [1].

1.3 Doğal Dil İşleme (NLP) Nedir?

Doğal Dil İşleme (NLP), insan diliyle etkileşimde bulunabilen bilgisayar sistemlerinin geliştirilmesiyle ilgilenen bir yapay zeka ve dilbilim dalıdır. NLP'nin temel amacı, insan dilinin karmaşık yapısını anlayarak ve işleyerek bu dili kullanışlı bilgilere dönüştürmektir. Temel NLP uygulamaları arasında metin sınıflandırma,

duygu analizi, dil modelleme, otomatik özet çıkarma ve dil çevirisi bulunur [4]. Görüntü alt yazılandırımda NLP, üretilen görsel özellikleri anlamlı ve doğal görünen metin ifadelerine dönüştürmek için hayati bir rol oynar [8] [9]. Bu süreç, derin öğrenme tekniklerinin ve dikkat mekanizmalarının kullanılmasıyla desteklenir [10] [2].

1.4 Görüntü Alt Yazılandırma (Image Captioning) Nedir?

Görüntü alt yazılandırma, bir görüntüyü tanımlayan ve açıklayan metin ifadeleri otomatik olarak üreten bir bilgisayarla görme ve doğal dil işleme görevidir [1]. Bu süreç, genellikle bir "encoder-decoder" modeli kullanılarak gerçekleştirilir [8]. Encoder, görüntüden önemli özellikleri çıkarmak için derin sinir ağlarından faydalanırken [2], decoder bu özellikleri alıp anlamlı bir cümle veya ifadeye dönüştürür [5]. Bu alanın zorlukları arasında, görüntülerdeki çeşitliliği ve karmaşıklığı doğru bir şekilde metne dökme [3] ve görüntüler arasındaki ince detayları yakalayıp yorumlama yeteneği bulunur [7].

2 Proje Tasarımı ve Geliştirme



2.1 Veri Seti ve Ön İşleme

Projemizde kullanılan "Flickr 8k Dataset", yapay zeka ve görüntü işleme alanlarında kritik bir öneme sahiptir. Bu özel veri seti, 8,000 adet çeşitlilik gösteren görüntülerle zenginleştirilmiş olup, her bir görüntüyü beş farklı açıdan açıklayan altyazılarla eşleştirmektedir. Bu şekilde veri kümesi yaklaşık 40.000 adet altyazı içermektedir. Bu özellik, makine öğrenimi modellerinin daha detaylı ve çeşitli veri üzerinde eğitilmesine olanak tanıyarak, algoritmaların gerçek dünya senaryolarına daha iyi adapte olmasını sağlar.

Tablo 1’de sunulan örnekler, veri setinin zenginliğini ve çeşitliliğini göstermektedir. Burada, her görüntü için farklı ifadelerle yazılmış beş adet metin içeriği bulunmaktadır, bu da her bir görüntüyü çok yönlü olarak anlamlandırma imkanı tanımaktadır. Özellikle, bu metinlerin manuel olarak hazırlanması, doğal dil işleme (NLP) sistemlerinin geliştirilmesi için oldukça değerli bir kaynak oluşturmaktadır.

Bu veri setinin seçimi, algoritmaların gerçek dünya koşullarında daha doğru ve etkili sonuçlar vermesine yardımcı olacak şekilde yapılmıştır. Görüntüler, tanınmış kişileri veya yerleri içermemek üzere altı farklı Flickr grubundan titizlikle seçilmiştir. Bu, etik ve gizlilik standartlarının korunmasında önemli bir adımdır ve aynı zamanda veri setinin genel kullanılabilirliğini artırmaktadır.

Table 1: Örnek Veri Tablosu

Resim Dosyası	Metin İçeriği
	A blonde horse and a blonde girl in a black sweatshirt are staring at a fire in a barrel. A girl and her horse stand by a fire. A girl holding a horse’s lead behind a fire. "A man, and girl and two horses are near a contained fire." Two people and two horses watching a fire.
	A couple stands close at the water’s edge. The two people stand by a body of water and in front of bushes in fall. Two people hold each other near a pond. Two people stand by the water. Two people stand together on the edge of the water on the grass.

Veri setindeki görüntülerin ön işleme için özel bir işlev zinciri kullanılmıştır. Her bir görüntü, TensorFlow kütüphanesi kullanılarak okunmuş, JPEG formatında çözülmüş ve [310x310] piksel boyutlarına yeniden boyutlandırılmıştır. Ardından, görüntülerin renk değerleri normalize edilerek 0 ile 1 arasında bir aralığa dönüştürülmüştür.

Altyazılar için uygulanan ön işleme süreci, metinleri küçük harfe çevirme, noktalama işaretlerini ve fazladan boşlukları kaldırma adımlarını içerir. Her bir

altyazı, başına '[start]' ve sonuna '[end]' tokenleri eklenerek standart bir formata dönüştürülmüştür. Bu işlemler, altyazıların model tarafından daha etkili bir şekilde işlenmesini sağlamaktadır.

2.2 Model Mimarisi

2.2.1 Encoder: Görüntü Öznitelikleri Nasıl Çıkarılır

Projenin Encoder kısmı, InceptionV3 modelini temel alır. Bu model, ImageNet ağırlıkları ile önceden eğitilmiş ve en üst katman dahil edilmeden kullanılmıştır. Modelin eğitilebilirliği kapalı tutularak, görsel içerikten öznitelik vektörleri çıkarmak için kullanılır. InceptionV3'ün çıktısı, global ortalama havuzlama (pooling) ile işlenerek yoğun bir vektör haline getirilir ve bu vektör Decoder'a iletilmek üzere hazırlanır.

2.2.2 Transformer Encoder Katmanı

Projede yer alan Transformer Encoder katmanı, çoklu başlık dikkat mekanizması ve yoğun katmanlardan oluşur. Bu katman, katman normalizasyonları ve derinlemesine dikkat mekanizması aracılığıyla girdi verilerini işler. Transformer Encoder, modele karmaşık dikkat ilişkilerini modellemesini sağlayarak, görsel öğeler arasındaki bağlantıları daha etkili bir şekilde anlamasına yardımcı olur.

2.2.3 Decoder: Metin Oluşturma ve Alt Yazılandırma

Decoder kısmı, metin ve pozisyon gömme katmanları ile birlikte çoklu başlık dikkat mekanizmaları ve yoğun katmanları içeren bir Transformer Decoder katmanından oluşur. Kendi kendine dikkat mekanizması ve Encoder-Decoder dikkat mekanizması, girdi metni ve Encoder çıktısını birleştirerek, tahmini metin çıktılarını üretir. Bu süreç, metni çözümleme ve tahmin etme kapasitesini artırarak, görüntü için uygun ve anlamlı altyazıları üretir.

2.2.4 Transformer Modellerin Rolü

Transformer modeller, projede, görsel ve metinsel veriler arasındaki karmaşık ilişkileri öğrenmek ve anlamlandırmak için kritik bir rol oynar. Bu modeller, çoklu başlık dikkat mekanizmalarını kullanarak, modelin hem görsel içeriğin derinlemesine analizini yapmasını hem de bu içeriğe uygun metin üretimini gerçekleştirmesini sağlar. Transformer modeller, paralel işleme yetenekleri ve uzun mesafeli bağımlılıkları etkili bir şekilde yakalama kabiliyetleri sayesinde, görsel alt yazılandırma görevinde yüksek başarı gösterir. Bu yaklaşım, modelin görsel öğeler arasındaki incelikli ilişkileri ve bunların dil ile olan bağlantılarını daha iyi anlamasına imkan tanır.

2.2.5 Modelin Eğitim ve Test Süreçleri

Model, eğitim ve test adımlarında kayıp ve doğruluk metriklerini hesaplar. Eğitim süreci, görsel ve metin verileri üzerinden modeli güncelleyerek, tahmin

edilen altyazıların doğruluğunu ve kalitesini artırmaya yöneliktir. Test süreci ise, modelin genel performansını ve altyazı üretme yeteneğini değerlendirir.

Bu model mimarisi, görüntü ve metin verileri arasındaki karmaşık ilişkileri öğrenmek ve anlamlandırmak için gelişmiş yapay zeka tekniklerini kullanır. Encoder ve Decoder bileşenleri arasındaki etkileşim, modelin veri setindeki görüntüleri ve ilgili altyazıları etkili bir şekilde işlemesini sağlar.

2.3 Performans Metrikleri ve Değerlendirme

Modelin performansı, çeşitli metrikler kullanılarak değerlendirilmiştir. Bu metrikler arasında doğruluk, kayıp oranı ve belki de daha önemlisi, üretilen altyazıların kalitesini ölçen BLEU skoru gibi dil modelleme metrikleri bulunur. Eğitim sürecinde, modelin kaybı ve doğruluğu sürekli olarak izlenmiş ve optimize edilmiştir. Test aşamasında, modelin gerçek dünya verileri üzerindeki performansı, seçilen metriklerle detaylı bir şekilde analiz edilmiştir. Bu değerlendirme, modelin güçlü yönlerini ve geliştirilmesi gereken alanları ortaya koyarak, gelecekteki iyileştirmeler için yön gösterici olmuştur.

3 Sonuçlar

3.1 Modelin Performans Analizi








Modelin eğitimi 20 epoch boyunca gerçekleştirilmiştir. İlk epoch sonunda, modelin eğitim kaybı 5.4380 ve doğruluk oranı %20.67 olarak kaydedilmiştir. İlerleyen epoch'larla birlikte modelin performansında belirgin bir iyileşme gözlemlenmiştir. Özellikle, 20. epoch sonunda modelin eğitim kaybı 2.5666'ya düşmüş ve doğruluk oranı %44.58'e yükselmiştir. Doğrulama (validation) kaybı ve doğruluk oranları da benzer bir iyileşme eğilimi göstermiş, 20. epoch sonunda sırasıyla 3.0596 ve %40.21 olarak kaydedilmiştir. Bu sonuçlar, modelin eğitim sürecinde sürekli olarak geliştiğini ve veri seti üzerindeki genel performansının arttığını göstermektedir.

Tablo 2'ye bakıldığında eğitim verilerinin bağlamından pek ayrılınmadığında modelin başarılı sonuçlar ürettiği gözükmemektedir. Örneğin model kardaki çocuklar ve köpeklerin olduğu fotoğraf için "a group of children are playing in a snow" altyazısının üretmiştir. Bu üretilen çıktı fotoğrafa bakıldığında akla gelen ilk cümlelerden biridir. Ya da model yine kamp ateşi olan fotoğraf için "a group of people are sitting in a campfire at night" altyazısını üretmiştir. Bu üretilen çıktı incelendiğinde yine benzer şekilde fotoğrafa bakıldığında akla ilk gelecek cümlelerden birini model başarılı şekilde üretmiştir.

3.2 Karşılaşılan Zorluklar ve Çözüm Yolları

Projede karşılaşılan başlıca zorluklar arasında, modelin karmaşık yapıları öğrenmede yavaş ilerlemesi ve overfitting (aşırı öğrenme) riski bulunmaktadır. Bu zorlukların üstesinden gelmek için çeşitli optimizasyon teknikleri ve düzenleme yöntemleri uygulanmıştır. Ayrıca, modelin daha hızlı ve etkili öğrenmesi için hiperparametre ayarlamaları yapılmıştır.

Table 2: Örnek Veri Tablosu

Resim Dosyası	Modelin Tahmini
	a brown dog is running through the snow
	a girl in a pink dress is standing on a sidewalk
	a group of young girls are standing in front of a crowd
	a small dog is jumping in the air to catch a frisbee
	a small dog runs through the grass
	a group of children are playing in a snow
	a group of people are sitting in a campfire at night

3.3 Gelecek Çalışmalar

Modelin eğitim ve doğrulama sonuçları, görüntü alt yazılandırma görevinde umut verici bir performans sergilediğini göstermektedir. Gelecek çalışmalarda, modelin doğruluğunu ve genel performansını daha da artırmak için ek veri set-

leri ve gelişmiş eğitim teknikleri kullanılabilir. Ayrıca, modelin farklı dillerde ve çeşitli görsel bağlamlarda nasıl performans gösterdiğini araştırmak da değerli olacaktır. Bu çalışma, görüntü alt yazılandırma alanında ileriye dönük araştırmalara yol gösterici olabilir.

3.4 Projenin Genel Değerlendirmesi

Bu proje, görüntü alt yazılandırma alanında önemli bir adım olarak değerlendirilebilir. Model, eğitim süresince gösterdiği sürekli iyileşme ile veri setindeki karmaşık görsel ve metinsel ilişkileri etkili bir şekilde öğrenebilme kapasitesini kanıtlamıştır. Elde edilen sonuçlar, modelin hem görsel özellikleri başarılı bir şekilde çıkarabildiğini hem de bu özellikleri anlamlı metin ifadelerine dönüştürebildiğini göstermektedir.

References

- [1] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5561–5570, 2018.
- [2] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020.
- [3] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134, 2019.
- [4] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.
- [5] Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos. Automatic image captioning. In *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, volume 3, pages 1987–1990. IEEE, 2004.
- [6] Himanshu Sharma, Manmohan Agrahari, Sujeet Kumar Singh, Mohd Firoj, and Ravi Kumar Mishra. Image captioning: a comprehensive survey. In *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, pages 325–328. IEEE, 2020.
- [7] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559, 2022.
- [8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016.
- [9] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE international conference on computer vision*, pages 4894–4902, 2017.
- [10] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.