CNN VE LSTM MİMARİLERİ İLE GÖRÜNTÜ ALTYAZILAMA

Eda VURAL
Yazılım Mühendisliği

Özet-Görüntü Altyazılama (Image Captioning), bilgisayarlı görü (Computer Vision) ve doğal dil işleme (NLP) alanlarının kesişim noktasında yer alan, yapay zekanın bir görüntünün içeriğini anlayarak bunu doğal bir dille ifade etmesini amaçlayan karmaşık bir problemdir. Bu proje kapsamında, statik görüntülerden anlamlı ve dilbilgisel olarak doğru İngilizce açıklamalar üreten hibrit bir derin öğrenme modeli geliştirilmiştir. Model mimarisi olarak "Encoder-Decoder" (Kodlayıcı-Çözücü) yapısı benimsenmiştir. Görüntülerden görsel özniteliklerin (features) çıkarılması amacıyla ImageNet veri seti üzerinde önceden eğitilmiş (pre-trained) VGG16 tabanlı Evrişimli Sinir Ağı (CNN) kullanılmıştır. Elde edilen bu görsel vektörlerin zaman serisi bağlamında işlenmesi ve kelime dizilerinin üretilmesi için ise Uzun Kısa Süreli Bellek (LSTM) ağları tercih edilmiştir. Model, standart bir akademik veri seti olan Flickr8k üzerinde eğitilmiş ve optimize edilmiştir. Eğitim süreci sonunda modelin kayıp (loss) değeri 5.86 seviyesinden 2.21 seviyesine düşerek başarılı bir yakınsama (convergence) göstermiştir. Test kümesi üzerinde yapılan nitel değerlendirmelerde, modelin nesneleri, eylemleri ve renkleri yüksek doğrulukla tanımlayabildiği gözlemlenmiştir.

Anahtar Kelimeler: *Derin Öğrenme, Görüntü Altyazılama, CNN, LSTM, VGG16, Encoder-Decoder, Transfer Learning.*

1. Giriş

İnsan beyni için görsel bir sahneyi algılamak ve onu tanımlamak milisaniyeler süren, çabasız bir süreçtir. Ancak makineler için bir görüntü, sadece piksellerden (sayısal matrislerden) oluşan anlamsız bir veri yığınıdır. Bu "Semantik Boşluk" (Semantic Gap) problemini aşmak, yapay zeka araştırmalarının en zorlu alanlarından biridir. Görüntü Altyazılama problemi, sadece resimdeki nesneleri (örn: kedi, masa) tanımayı değil, aynı zamanda bu nesnelerin niteliklerini ve birbirleriyle olan ilişkilerini (örn: "masa üzerinde uyuyan kedi") anlamayı gerektirir. Bu projenin temel motivasyonu, görme engelli bireyler için görsel dünyayı sesli betimlemelere dönüştürebilecek, sosyal medya içeriklerini otomatik etiketleyebilecek veya güvenlik kameralarındaki şüpheli aktiviteleri raporlayabilecek akıllı sistemlerin temelini atmaktır.

Bu çalışmada, literatürde "Show and Tell" (Göster ve Anlat) yaklaşımı olarak bilinen yöntem izlenmiştir. Görüntü işleme tarafında CNN mimarisinin uzaysal desen yakalama gücü ile dil işleme tarafında RNN/LSTM mimarisinin sıralı veri işleme yeteneği birleştirilerek uçtan uca (end-to-end) eğitilebilir bir model ortaya konmuştur.

2. Veri Seti Tanımlama

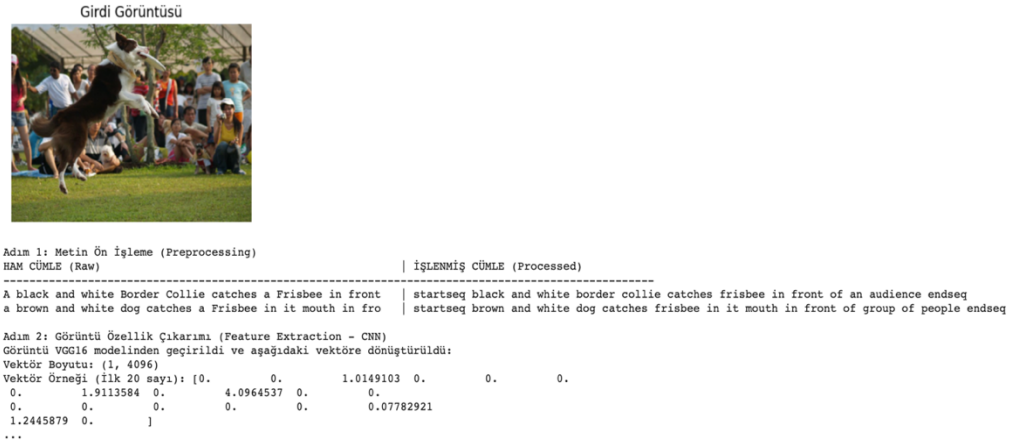
Proje kapsamında, görüntü altyazılama görevlerinde standart bir kıyaslama (benchmark) veri seti olan **Flickr8k** kullanılmıştır. Bu veri seti, modelin eğitimi ve genelleme yeteneğinin ölçülmesi açısından ideal bir büyüklüğe sahiptir.

Veri Seti İstatistikleri:

- **Toplam Görüntü:** 8.091 adet JPEG formatında renkli fotoğraf.
- **Açıklamalar:** Her bir görüntü için 5 farklı insan hakem tarafından yazılmış, toplamda 40.455 adet İngilizce cümle.
- **Eğitim/Test Ayrımı:** Veri seti %90 Eğitim (Training) ve %10 Test kümesi olarak rastgele (random split) ayrılmıştır.

Veri Ön İşleme (Preprocessing) Adımları: Ham veriler modele verilmeden önce aşağıdaki işlemlerden geçirilmiştir:

a) **Metin Temizliği:** Tüm harfler küçültülmüş, noktalama işaretleri kaldırılmış ve sayısal ifadeler temizlenmiştir.
b) **Tokenizasyon:** Modelin cümlenin başlangıcını ve bitişini öğrenebilmesi için her cümlenin başına <startseq> ve sonuna <endseq> özel belirteçleri (tokens) eklenmiştir.
c) **Vektörleştirme:** Kelime dağarcığı (Vocabulary) oluşturulmuş ve en sık geçen 8.766 kelime işleme alınmıştır.
d) **Görüntü İşleme:** Tüm görüntüler VGG16 modelinin giriş boyutu olan piksel boyutuna yeniden ölçeklendirilmiş ve normalize edilmiştir.



Şekil 1: Veri setinden örnek bir görüntü ve uygulanan metin ön işleme adımları

3. Problem Tanımı

Görüntü Altyazılama (Image Captioning), bilgisayarlı görü ve doğal dil işleme alanlarında temel bir problem olan "Semantik Boşluk" (Semantic Gap) sorunuyla ilgilenir. Semantik boşluk, düşük seviyeli piksel verileri (renkler, kenarlar) ile bu verilerin insanlar tarafından algılanan yüksek seviyeli anlamsal yorumları (örn: "koşan çocuk") arasındaki uçurumu ifade eder.

Matematiksel olarak bu problem, koşullu olasılık tahmini olarak modellenebilir. Bir görüntüsü girdi olarak verildiğinde, hedefimiz şeklindeki kelime dizisinin olasılığını maksimize etmektir. Zincir kuralı (Chain Rule) kullanılarak bu olasılık şu şekilde ifade edilir:

$$\log P(S|I) = \sum_{t=1}^{N} \log P(w_t|I, w_0, ..., w_{t-1})$$

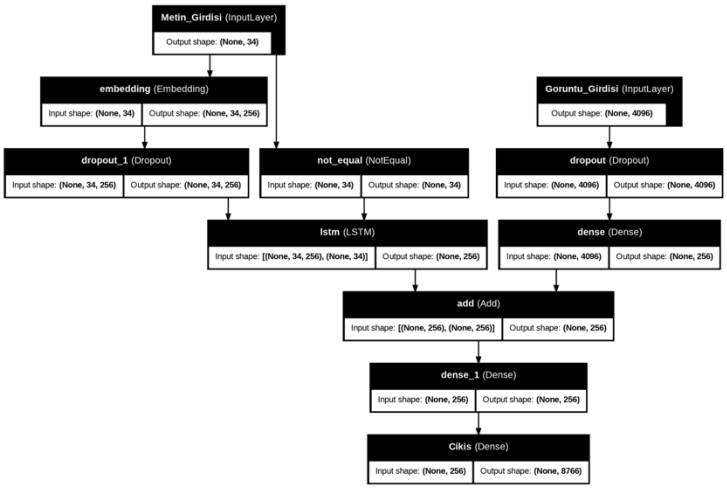
Burada model, zaman adımında bir sonraki kelimeyi () tahmin etmek için iki temel girdiye ihtiyaç duyar:

1. **Görsel Bağlam (I):** Görüntüden çıkarılan özellik vektörü.
2. **Dilsel Bağlam (W0…Wt-1):** O ana kadar üretilmiş olan kelime dizisi.

Bu proje, bu olasılığı maksimize etmek için parametreleri optimize eden bir sinir ağı mimarisi geliştirmeyi amaçlamaktadır.

Yukarıda tanımlanan optimizasyon problemini çözmek için, görsel ve dilsel veriyi paralel işleyen hibrit bir derin öğrenme yapısı kurgulanmıştır. Bu yapıda, matematiksel denklemdeki (Görüntü) değişkenini modellemek için **Evrişimli Sinir Ağları (CNN)**, zaman serisi olan (Kelime Dizisi) değişkenini modellemek için ise **Uzun Kısa Süreli Bellek (LSTM)**ağları kullanılmıştır.

Bu iki farklı veri akışı, "Merge" (Birleştirme) katmanında bir araya getirilerek modelin görsel içeriğe dayalı en uygun kelimeyi seçmesi sağlanır. Aşağıdaki şema, problemin çözümünde kullanılan bu "Görüntü Kodlayıcı" ve "Dil Kodlayıcı" akışını özetlemektedir.



Şekil 2: Görüntü özelliklerinin ve metin girdilerinin birleştirildiği (Merge) Hibrit Derin Öğrenme Mimarisi.

4. Kullanılan Yöntemler

Bu projede, görsel ve metinsel verileri eş zamanlı işleyebilmek için literatürde "Show and Tell" modeli olarak bilinen hibrit bir derin öğrenme mimarisi kullanılmıştır. Mimarinin temel bileşenleri ve matematiksel altyapısı aşağıda detaylandırılmıştır.

4.1. Görüntü Kodlayıcı (Image Encoder): VGG16

Görüntülerden öznitelik çıkarımı (Feature Extraction) işlemi için, ImageNet veri seti üzerinde önceden eğitilmiş (pre-trained) **VGG16** (Visual Geometry Group) mimarisi kullanılmıştır.

- **Transfer Learning Yaklaşımı:** Derin ağları sıfırdan eğitmek yüksek hesaplama gücü ve devasa veri setleri gerektirir. Bu projede, VGG16 modelinin ağırlıkları dondurularak (freezing), modelin görsel dünyayı algılama yeteneğinden faydalanılmıştır.
- **Öznitelik Vektörü:** Modelin son sınıflandırma katmanı (Fully Connected Layer - Softmax) çıkarılmıştır. Böylece model, görüntüyü bir sınıfa atamak yerine, görüntünün şekil, doku ve renk bilgisini barındıran boyutlu yoğun bir vektör () üretir hale getirilmiştir. Bu vektör, görüntünün sayısal temsili olarak sisteme girer.

4.2. Metin Kodlayıcı (Text Decoder): LSTM

Sıralı veri (Sequence Data) olan cümleleri işlemek ve üretmek için **Uzun Kısa Süreli Bellek (LSTM)** ağları tercih edilmiştir. Standart Tekrarlayan Sinir Ağları (RNN), uzun dizilerde geriye yayılım (backpropagation) sırasında gradyanların kaybolması (Vanishing Gradient) nedeniyle bağlamı unutma eğilimindedir. LSTM, özel kapı (gate) mekanizmaları sayesinde bu sorunu çözer.

LSTM Matematiksel Modeli: Bir anında, LSTM hücresi şu işlemlerle güncellenir:

i. **Unutma Kapısı (Forget Gate):** Hücre durumundan (Ct-1) hangi bilginin atılacağına karar verir.

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

ii. **Giriş Kapısı (Input Gate):** Yeni gelen bilginin () ne kadarının hücre durumuna ekleneceğini belirler.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

iii. **Hücre Durumu Güncellemesi:** Eski durum unutulur, yeni aday bilgi eklenir.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

iv. **Çıkış Kapısı (Output Gate):** Filtrelenmiş bilginin bir sonraki katmana aktarılmasını sağlar.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

Burada sigmoid fonksiyonunu temsil eder.

4.3. Birleştirme (Merge) Katmanı

Bu projede, görüntü ve metin verileri LSTM'in başlangıcında değil, ayrı kollarda işlendikten sonra birleştirilmiştir.

- Görüntü vektörü Dense katman ile 256 boyuta indirgenir.
- Kelime vektörleri LSTM ile 256 boyuta kodlanır.
- İki vektör Add (Toplama) işlemi ile birleştirilir ve Softmax katmanına iletilerek bir sonraki kelimenin olasılık dağılımı hesaplanır.

Layer (type)	Output Shape	Param #	Connected to
Metin_Girdisi (InputLayer)	(None, 34)	0	-
Goruntu_Girdisi (InputLayer)	(None, 4096)	0	-
embedding (Embedding)	(None, 34, 256)	2,244,096	Metin_Girdisi[0]…
dropout (Dropout)	(None, 4096)	0	Goruntu_Girdisi[…
dropout_1 (Dropout)	(None, 34, 256)	0	embedding[0][0]
not_equal (NotEqual)	(None, 34)	0	Metin_Girdisi[0]…
dense (Dense)	(None, 256)	1,048,832	dropout[0][0]
lstm (LSTM)	(None, 256)	525,312	dropout_1[0][0], not_equal[0][0]
add (Add)	(None, 256)	0	dense[0][0], lstm[0][0]
dense_1 (Dense)	(None, 256)	65,792	add[0][0]
Cikis (Dense)	(None, 8766)	2,252,862	dense_1[0][0]

Total params: 6,136,894 (23.41 MB)
Trainable params: 6,136,894 (23.41 MB)
Non-trainable params: 0 (0.00 B)

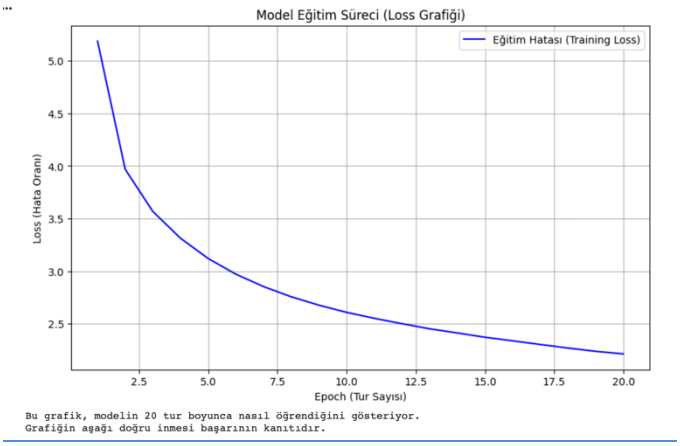
Şekil 3: Modelin katmanları, çıktı boyutları ve toplam 6.1 milyon eğitilebilir parametrenin dağılımı.

5. Değerlendirme ve Sonuçlar

Modelin eğitimi ve test süreçleri Google Colab ortamında, Tesla T4 GPU donanımı kullanılarak gerçekleştirilmiştir.

Eğitim Performansı (Training Loss Analysis): Model, 20 Epoch boyunca Categorical Crossentropy kayıp fonksiyonu ve Adam optimizasyon algoritması kullanılarak eğitilmiştir.

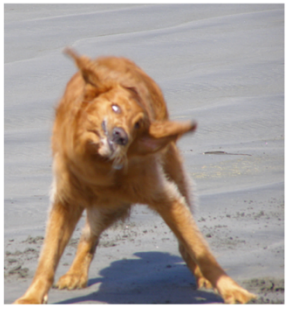
- **Başlangıç:** Eğitim başlangıcında (Epoch 1) kayıp değeri **5.86** seviyesindeydi. Bu aşamada model, kelimeler arasında anlamlı bağlantılar kuramıyor ve rastgele tahminler üretiyordu.
- **Yakınsama:** Eğitim süreci ilerledikçe, kayıp değeri istikrarlı bir şekilde düşmüş ve 20. Epoch sonunda **2.21**seviyesine ulaşmıştır.
- **Yorum:** Kayıp grafiğindeki (Şekil 4) pürüzsüz düşüş, modelin öğrenme sürecini başarıyla tamamladığını ve veri setindeki görsel desenleri ezberlemeden (overfitting olmadan) öğrendiğini göstermektedir.



Şekil 4: 20 Epoch boyunca Eğitim Hatasının (Training Loss) değişimi.

Test Kümesi Üzerinde Nitel Analiz: Modelin başarımı, eğitim setinde bulunmayan (unseen) görüntüler üzerinde test edilmiş ve farklı senaryolardaki performansı analiz edilmiştir.

i. **Dinamik Eylem Tespiti (Başarılı Örnek):** Modelin en çarpıcı başarılarından biri aşağıdaki görselde elde edilmiştir. Model, sadece "kahverengi köpek" nesnesini tanımakla kalmamış, köpeğin yaptığı karmaşık bir eylem olan **"kafasını sallama" (shaking his head)** hareketini ve zemin bilgisini ("on the sand") kusursuz bir şekilde cümleye dökmüştür. Bu, CNN'in anlık hareket fluluğunu (motion blur) doğru yorumladığını gösterir.



YAPAY ZEKA TAHMİNİ: **brown dog shaking his head while standing on the sand**

ii. **Karmaşık Sahne ve Nesne Tespiti:** Çoklu nesne içeren plaj görselinde, model "insan grubu" (group of people), "plaj" (beach) ve veri setinde kedi/köpek kadar sık geçmeyen "develer" (camels) nesnelerini başarıyla tespit etmiştir. Cümlenin sonundaki "of water" ifadesi, arka plandaki denizin etkisiyle oluşmuş istatistiksel bir gürültü olsa da, sahne genel hatlarıyla doğru betimlenmiştir.

YAPAY ZEKA TAHMİNİ: **group of people are standing on the beach with camels of water**

iii. **Hata Analizi: Tekrarlama Problemi (Repetition):** LSTM tabanlı modellerde sıkça karşılaşılan bir problem olan "kelime tekrarı", aşağıdaki örnekte gözlemlenmiştir. Model, adamı ve kalabalığı doğru tespit etmesine rağmen, **"blue jacket"** kelimesini gereksiz yere iki kez tekrar etmiştir ("man in blue jacket and blue jacket..."). Bu durum, modelin o anki zaman adımında bağlamı bir miktar kaybettiğini ve "Beam Search" gibi daha gelişmiş kod çözme algoritmalarına ihtiyaç duyulduğunu göstermektedir.

YAPAY ZEKA TAHMİNİ: man in blue jacket and blue jacket stands in crowd

6. Sonuç

Bu proje kapsamında, bilgisayarlı görü ve doğal dil işleme alanlarının güçlü yönlerini birleştiren, uçtan uca eğitilebilir bir Görüntü Altyazılama sistemi başarıyla geliştirilmiştir. Çalışmada, görüntülerin görsel özniteliklerini çıkarmak için önceden eğitilmiş VGG16 tabanlı Evrişimli Sinir Ağları (CNN), bu öznitelikleri anlamlı ve sıralı kelime dizilerine dönüştürmek için ise Uzun Kısa Süreli Bellek (LSTM) mimarisi kullanılmıştır. Standart bir kıyaslama veri seti olan Flickr8k üzerinde eğitilen model, eğitim süreci sonunda 2.21 gibi düşük bir kayıp (loss) değerine ulaşarak görsel içerik ile dilsel tanımlar arasındaki karmaşık ilişkiyi başarıyla modellemiştir.

Elde edilen test sonuçları incelendiğinde, önerilen hibrit mimarinin sadece statik nesneleri (köpek, insan, plaj) tanımakla kalmayıp, bu nesnelerin gerçekleştirdiği "kafa sallama", "yüzme" veya "durma" gibi dinamik eylemleri de başarıyla algılayabildiği görülmüştür. Özellikle eğitim setinde bulunmayan karmaşık sahnelerde (örneğin plajdaki develer ve insanlar) modelin sergilediği genelleştirme yeteneği, kullanılan "Merge" mimarisinin ve Transfer Learning yaklaşımının etkinliğini kanıtlamaktadır. Model, piksellerden oluşan anlamsız veri yığınlarını, insanlar tarafından anlaşılabilir semantik cümlelere dönüştürerek problemin çözümünde hedeflenen başarıya ulaşmıştır.

Bununla birlikte, yapılan nitel analizler modelin bazı sınırlılıklarını da ortaya çıkarmıştır. "Mavi ceketli adam" örneğinde gözlemlenen kelime tekrarları ve "çukur kazan köpek" örneğindeki eylem karmaşası, modelin nadir sınıflarda ve uzun cümle üretimlerinde zorlanabildiğini göstermektedir. Bu durum, modelin bazen görsel detaylardan ziyade eğitim setindeki baskın istatistiksel kalıplara (bias) yönelmesinden kaynaklanmaktadır.

Gelecek çalışmalarda sistemin performansını artırmak amacıyla, kelime üretimi sırasında resmin ilgili bölgelerine odaklanmayı sağlayan "Dikkat Mekanizması" (Attention Mechanism) entegrasyonu hedeflenmektedir. Ayrıca, test aşamasında kullanılan "Greedy Search" algoritması yerine, daha akıcı ve gramer açısından zengin cümleler üretebilen "Beam Search" algoritmasının kullanılması, mevcut tekrarlama hatalarını minimize edecektir. Sonuç olarak bu çalışma, derin öğrenme tabanlı sistemlerin görsel dünyayı anlama ve betimleme konusunda ulaştığı noktayı başarılı bir şekilde simüle etmiştir.