Derin Sinir Ağlarıyla Osmanlıca Optik Karakter Tanıma

1. Giriş

Osmanlıca, Arap alfabesiyle yazılan ve 13. yüzyıldan 20. yüzyıla kadar Osmanlı İmparatorluğu'nda kullanılan bir yazı dilidir. Günümüzde Osmanlıca metinlerin okunması ve anlaşılması zor olduğundan, bu metinlerin dijitalleştirilmesi için optik karakter tanıma (OCR) teknolojisine ihtiyaç duyulmaktadır. Bu çalışmada, Osmanlıca metinlerin OCR ile dijitalleştirilmesi için derin sinir ağları (CNN+RNN) kullanılarak bir web tabanlı sistem geliştirilmiştir.

2. Veri Kümesi

Proje için kullanılan veri kümeleri şunlardır:

• **Orijinal Veri**: Yaklaşık 1.000 sayfa Osmanlıca metin görüntüsü.

• **Sentetik Veri**: Yaklaşık 23.000 sayfa sentetik olarak üretilmiş metin görüntüsü.

• **Hibrit Veri**: Orijinal ve sentetik verilerin birleşimi.

• **Test Verisi**: 21 sayfalık orijinal Osmanlıca metin görüntüsü.

Sıklıkları ise;

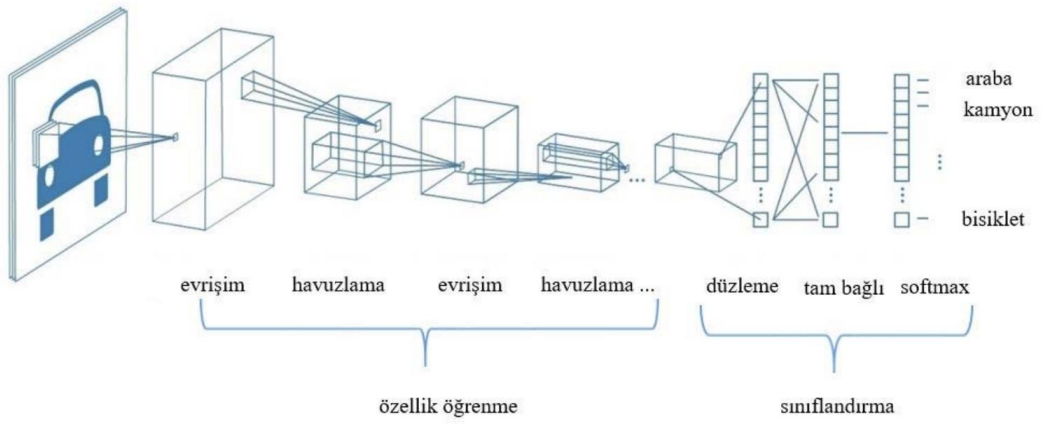
Küme	Sayfa	Satır	Kelime	Karakter
Sentetik	26B	1.3M	263B	78M
Orijinal	1B	18B	35B	252B
Eğitim	27B	1.3M	298B	78M
Test	21	420	3B	23B

3. Derin Öğrenme Mimarisi

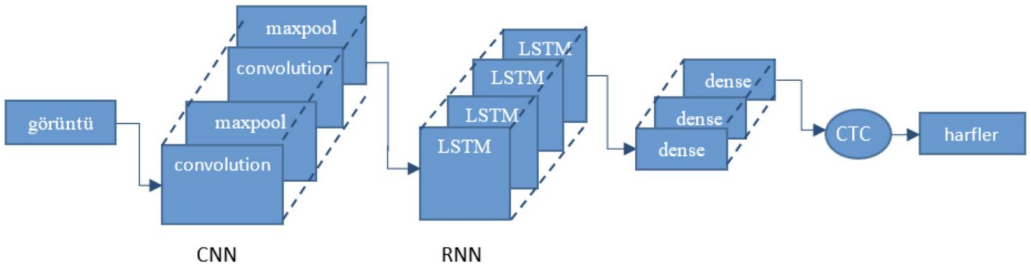
CNN (Evrişimli Sinir Ağları): Görüntüdeki özellikleri çıkarmak için kullanılmıştır. Görüntüdeki görsel örüntüleri tanımak için evrişim katmanları ve havuzlama katmanları kullanılır.

RNN (Yinelemeli Sinir Ağları): Özellikle LSTM (Uzun Kısa Süreli Bellek) modelleri, metin dizilerini tanımak için kullanılmıştır. LSTM, uzun süreli bağımlılıkları öğrenebilir ve hem önceki hem de sonraki karakterler arasındaki bağlamı öğrenerek daha doğru tahminler yapar.

CTC (Bağlantısal Zaman Sınıflandırması): Karakter dizilerini etiketlemek için kullanılır ve girdi ile çıktı arasındaki hizalamayı öğrenir.



Şekil 2. Görüntü tanımada kullanılan standart CNN mimarisi [30] (Conventional CNN architecture used in image recognition)



4. Deneyler ve Karşılaştırmalar

Karşılaştırılan OCR araçları:

• **Tesseract (Arapça ve Farsça)**: Açık kaynaklı ve ücretsiz bir OCR aracı.

• **Abby FineReader**: Ticari bir OCR aracıdır.

• **Google Docs**: Google'ın ücretsiz OCR hizmetidir.

• **Miletos**: Osmanlıca için özel olarak geliştirilmiş bir OCR aracı.

5. Hiper Parametre Kestirimi

Farklı hiper parametreler (filtre boyutu, öğrenme hızı, LSTM boyutu, aktivasyon fonksiyonu) üzerinde deneyler yapılmıştır. Öğrenme hızının artırılması, doğruluk oranlarında iyileşme sağlamıştır.

Deney	Parametre	Ham	Normalize	Bitişik
Orijinal deney	663703	88,86	96,12	97,37
Öğrenme hızı	663783	88,63	95,63	96,80
LSTM boyutu	194919	88,01	95,26	96,43
Aktivasyon fonk.	663783	88,26	95,33	96,51
Filtre boyutu	664039	88,44	95,64	96,89

6. Sonuçlar

Geliştirilen Osmanlica.com Hibrit OCR modeli, diğer OCR araçlarına kıyasla daha yüksek doğruluk oranları elde etmiştir. Osmanlıca metinlerin normalize edilmesi, OCR doğruluğunu artırmada kritik bir rol oynamaktadır. Ayrıca, hemze ve med işaretli harflerin tanınması ve karakter düzeltme adımlarının eklenmesi planlanmaktadır.

7. Gelecek Çalışmalar

• OCR sonrası karakter düzeltme adımlarının eklenmesi.

• Hemze ve med işaretli harflerin tanınmasına yönelik iyileştirmeler.

• Daha büyük ve çeşitli veri kümeleri üzerinde eğitim yapılması.