Derin sinir ağlarıyla Osmanlıca optik karakter tanıma

Giriş

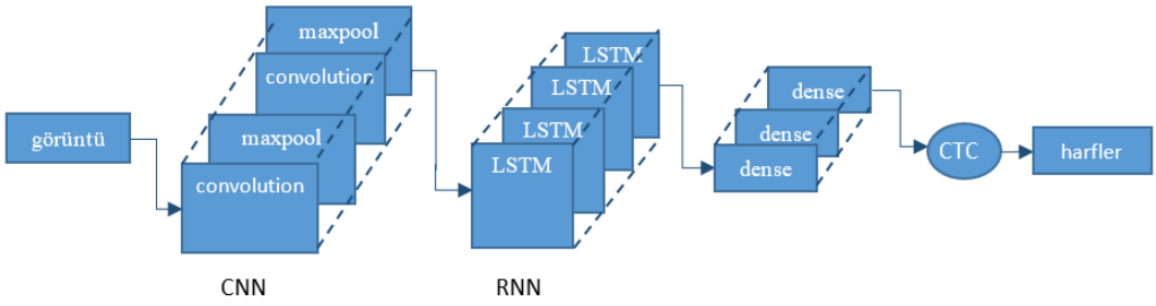
Osmanlıca**,** 13. yüzyıldan 20. yüzyıla kadar Osmanlı İmparatorluğu'nda kullanılan ve Arap alfabesiyle yazılan bir dildir. Günümüzde Latin alfabesine geçişle birlikte Osmanlıca metinlerin okunması zorlaşmış, ancak bu belgeler kültürel ve tarihî açıdan büyük önem taşımaktadır. Bu çalışmada, Osmanlıca metinleri dijitalleştirmek için derin öğrenme tabanlı bir OCR modeli geliştirilmiş ve mevcut sistemlerle karşılaştırılmıştır. Özellikle Osmanlıca matbu nesih hattındaki yazıları tanımaya yönelik deneysel sonuçlar sunulmaktadır.

Kullanılan Derin Öğrenme Mimarisi

Makale, Convolutional Recurrent Neural Network (CRNN) adı verilen bir **CNN + RNN tabanlı hibrit model** kullanmaktadır.

1. **CNN (Evrişimli Sinir Ağı)**:
   - Görüntüdeki **harf ve kelime desenlerini** tanımak için kullanılmıştır.
   - Evrişim katmanları ile harflerin şekilleri ve özellikleri çıkarılmıştır.
   - ReLU (Rectified Linear Unit) aktivasyon fonksiyonu ile doğrusal olmayan öğrenme sağlanmıştır.
2. **RNN (Tekrarlayan Sinir Ağı) - LSTM (Uzun Kısa Süreli Bellek)**:
   - Metinlerin zamansal bağımlılıklarını öğrenmek için kullanılmıştır.
   - İki yönlü LSTM (Bidirectional LSTM) mimarisi, karakterleri bağlamsal olarak anlamlandırmıştır.
   - CTC (Connectionist Temporal Classification) katmanı, harfleri belirli bir sırayla tanımak için uygulanmıştır.
3. **CTC (Connectionist Temporal Classification) Katmanı**:
   - Modelin, kelimelerin bölünme noktalarını anlamasına yardımcı olmuştur.
   - **Kelime tahmini ve düzeltilmiş karakter dizisi üretme** görevini üstlenmiştir.

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



Şekil 3. Osmanlıca OCR için CRNN mimarisi (CRNN architecture for Ottoman OCR)

Veri Kümesi ve Eğitim Süreci

- **3 farklı veri seti** kullanılarak model eğitilmiştir:
    1. **Orijinal Veri Seti**: 1.000 sayfalık Osmanlıca doküman içermektedir.
    2. **Sentetik Veri Seti**: 23.000 sayfa üretilmiş Osmanlıca metinden oluşmaktadır.
    3. **Hibrit Veri Seti**: Orijinal ve sentetik verilerin birleşimi olarak oluşturulmuştur.
- **Eğitim Parametreleri**:
    - **Öğrenme oranı (Learning Rate):** 0.002
    - **Momentum:** 0.5
    - **Eğitim Epoch Sayısı:** 3.000.000 iterasyon

Deneysel Sonuçlar ve Başarı Oranları

- **Modelin doğruluk oranları**:
    - **Karakter tanıma**: %88,86 (ham), %96,12 (normalize), %97,37 (bitişik)
    - **Bağlı karakter tanıma**: %80,48 (ham), %91,60 (normalize), %97,37 (bitişik)
    - **Kelime tanıma**: %44,08 (ham), %66,45 (normalize)
- **Mevcut OCR araçlarıyla kıyaslama**:
    - Model, Tesseract (Arapça ve Farsça), Google Docs OCR ve Abby FineReader gibi mevcut sistemlere göre daha iyi performans göstermiştir.

Sonuç ve Katkılar

- CNN ve RNN'in birleşimi olan **CRNN modeli**, Osmanlıca OCR konusunda mevcut sistemlerden daha başarılı olmuştur.
- Bağlamsal ve dil özellikleri dikkate alınarak bir hata düzeltme mekanizması geliştirilmiştir.
- Çalışma sonucu elde edilen model, Osmanlıca karakter tanıma için en yüksek doğruluk oranına ulaşan OCR sistemlerinden biri olmuştur.

Gelecek çalışmalar için modelin el yazısı Osmanlıca metinlere uyarlanması planlanmaktadır.