Derin Sinir Ağlarıyla Osmanlıca Optik Karakter Tanıma

1. Giriş

- Osmanlıca, Arap alfabesiyle yazılan ve 13. yüzyıldan 20. yüzyıla kadar 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. Osmanlı Alfabesi ve Karakteristikleri

- Osmanlıca, Arap, Fars ve Türk dillerinden alınan harflerle oluşturulmuş bir alfabeye sahiptir.
- Harfler bitişik yazılır ve bazı harfler farklı pozisyonlarda farklı şekiller alır.
- OCR için harflerin bitişme özellikleri, nokta sayıları ve konumları gibi ayırt edici özellikler önemlidir.

3. Veri Kümesi

- 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.

4. Derin Öğrenme Mimarisi

- CNN (Evrişimli Sinir Ağları): Görüntüdeki özellikleri çıkarmak için kullanılmıştır.
- RNN (Yinelemeli Sinir Ağları): Özellikle LSTM (Uzun Kısa Süreli Bellek) modelleri, metin dizilerini tanımak için kullanılmıştır.
- CTC (Bağlantısal Zaman Sınıflandırması): Karakter dizilerini etiketlemek için kullanılmıştır.

5. Deneyler ve Karşılaştırmalar

- Karşılaştırılan OCR Araçları: Tesseract (Arapça ve Farsça), Google Docs (Arapça), Abby FineReader (Arapça), Miletos (Osmanlıca).
- Doğruluk Ölçütleri: Karakter, katar ve kelime tanıma doğruluk oranları.

- Sonuçlar:

- Karakter Tanıma: Osmanlica.com Hibrit modeli %88,86 ham, %96,12 normalize ve %97,37 bitişik doğruluk oranlarıyla en iyi performansı göstermiştir.
- Katar Tanıma: Hibrit model %80,48 ham, %91,60 normalize ve %97,37 bitişik doğruluk oranlarıyla diğer araçlardan daha başarılı olmuştur.
- Kelime Tanıma: Hibrit model %44,08 ham ve %66,45 normalize doğruluk oranlarıyla en yüksek performansı sergilemiştir.

6. Harf Türüne Göre Tanıma Doğrulukları

- Arapça harfler, noktalı/noktasız harfler, nokta sayısı ve konumu gibi özelliklere göre tanıma doğrulukları analiz edilmiştir.
- Noktalı harflerde tanıma hataları daha yüksek çıkmıştır.

7. Hiper Parametre Kestirimi

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

8. 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.
- İleride, hemze ve med işaretli harflerin tanınması ve karakter düzeltme adımlarının eklenmesi planlanmaktadır.

9. 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.