**Osmanlıca Optik Karakter Tanıma (OCR)**

**Derin Öğrenme ile Osmanlıca OCR Modeli**

**Giriş**

* Osmanlıca, 13. yüzyıldan 20. yüzyıla kadar Osmanlı İmparatorluğu'nda kullanılan bir yazı dili olup, Arap alfabesiyle yazılmıştır.
* Günümüzde bu belgelerin dijital ortama aktarılması ve işlenmesi zor bir süreçtir. Bu nedenle optik karakter tanıma teknolojileri büyük önem taşımaktadır.
* Çalışmanın amacı, Osmanlıca nesih hattı ile basılmış dokümanları otomatik olarak tanıyabilen ve metne dönüştürebilen bir OCR modeli geliştirmektir.
* Günümüzde kullanılan OCR sistemleri Osmanlıca metinleri tanımada düşük başarı göstermektedir. Bu yüzden bu çalışmada derin öğrenme tabanlı bir yaklaşım benimsenmiştir.

**Kullanılan Derin Öğrenme Modeli**

* Çalışmada CNN (Convolutional Neural Network) ve LSTM (Long Short-Term Memory) mimarilerinin birleşimi kullanılmıştır.
* CNN, görsel verilerden özellik çıkarımı yaparken, LSTM, karakter dizilerinin sıralı doğasını öğrenerek daha doğru tahminler yapmayı sağlar.

**Modelin eğitimi için üç farklı veri kümesi oluşturulmuştur:**

* **Orijinal veri kümesi:** 1000 Osmanlıca sayfadan oluşmaktadır.
* **Sentetik veri kümesi:** Yapay olarak oluşturulmuş 23.000 sayfadan oluşmaktadır.
* **Hibrit veri kümesi:** Hem orijinal hem de sentetik verileri içeren karışık bir veri kümesidir.

Modelin performansı, popüler OCR araçları olan Google Docs, Abby FineReader, Miletos ve Tesseract (Arapça ve Farsça versiyonları) ile karşılaştırılmıştır.

OCR çıktılarının doğruluğunu artırmak için bir normalizasyon algoritması kullanılmıştır. Bu algoritma, hataları düzelterek karakter, kelime ve bileşen seviyesinde analiz yapmaya olanak tanımaktadır.

**Sonuçlar**

Yapılan deneylerde hibrit modelin en iyi sonucu verdiği görülmüştür:

**Karakter tanıma doğruluğu:**

* **Ham metin:** %88.86
* **Normalleştirilmiş metin:** %96.12
* **Bitişik metin:** %97.37

**Bağlı harf bileşeni tanıma doğruluğu:**

* **Ham metin:** %80.48
* **Normalleştirilmiş metin:** %91.60
* **Bitişik metin:** %97.37

**Kelime tanıma doğruluğu:**

* **Ham metin:** %44.08
* **Normalleştirilmiş metin:** %66.45

Hibrit model, diğer OCR modellerine göre açık bir üstünlük göstermektedir. Google Docs, Abby FineReader ve Tesseract gibi araçlarla yapılan karşılaştırmalarda daha yüksek doğruluk oranlarına ulaşılmıştır.

Model, osmanlica.com platformunda kullanılabilir hale getirilmiştir ve geniş bir arşiv için Osmanlıca belgelerin dijitalleştirilmesi sürecinde kullanılabilir.