Derin Sinir Ağları ile Osmanlı Optik Karakter Tanıma

Giriş

Osmanlı arşiv ve kütüphanelerindeki belgeler, yüzlerce yıllık bir kültürel, sanatsal ve tarihsel mirası temsil eder. Bu belgelerin dijital ortama aktarılması, yalnızca tarih araştırmaları için değil, aynı zamanda kültürel mirasın koruma ve erişilebilirliğini artırmak adına da büyük önem taşımaktadır. Ancak, bu belgelerin çoğu Osmanlıca yazılmıştır ve mevcut OCR teknolojileri bu dili tanıma konusunda yeterli başarıyı gösterememektedir. Bu nedenle, etkili bir Osmanlıca OCR sistemi geliştirmek adına yapılan bu çalışma kritik bir ihtiyaçtan doğmaktadır.

Araştırmanın Amacı

Bu araştırmanın temel amacı, Osmanlıca metinlerin doğru ve etkili bir şekilde dijitalleştirilmesi için bir OCR sistemi geliştirmektir. Proje, Osmanlıca'dan günümüz Türkçesine metin aktarımını sağlamak amacıyla derin öğrenme yöntemlerini ve yapay zeka uygulamalarını kullanarak gerçekleştirilmiştir. Özellikle, mevcut sistemlerin sınırlamalarını aşmak, karışık yazım yapılarını tanımak ve metinlerin doğruluğunu artırmak hedeflenmektedir.

Yöntem Bilgisi

1. Derin Öğrenme Yaklaşımlarının Kullanımı

Çalışmada, görsel tanıma süreçleri için derin sinir ağı mimarileri kullanılmıştır. Modeller arasında en çok tercih edilenler;

- **CNN (Konvolüsyonel Sinir Ağı)**: Görüntü sınıflandırma ve analizinde etkili bir mimaridir.
- **RNN (Tekrarlayan Sinir Ağı)**: Özellikle sıralı veri analizi ve dil işleme alanlarında kullanılır.

Bu çalışmada, CNN ve RNN kombinasyonu, karakter tanıma işlemlerinin daha etkili bir şekilde gerçekleştirilmesi amacıyla uygulanmıştır. Derin öğrenme mimarisi, veriyi yüksek katmanlarda işleyerek daha az hata ile karakterlerin tanınmasını sağlar.

2. Veri Setleri

Araştırma için üç farklı veri seti oluşturulmuştur:

- **Orijinal Veri Seti:** Osmanlıca belgelerden oluşan yaklaşık 1000 sayfalık bir koleksiyon. Bu veri seti, tarihsel belgelerin gerçek örneklerini içerir ve modelin gerçek hayatta karşılaşabileceği zorlukları simüle eder.
- **Sentetik Veri Seti:** Yaklaşık 23,000 sayfa içeren yapay olarak oluşturulmuş belge koleksiyonu. Bu set, modelin genel geçerliliğini artıracak türden geniş bir veri tabanı sağlar ve çeşitli karakter kombinasyonlarını içerir.

- **Hibrit Veri Seti:** Orijinal ve sentetik verilerin birleştirilmesiyle oluşturulmuş bir veri setidir. Bu, modelin daha geniş bir kapsama ve farklı yazı stillerine adaptasyon yeteneğini artırır.

3. Performans Değerlendirmesi

Geliştirilen OCR sistemi, mevcut OCR araçlarıyla (Tesseract, Google Docs, Abby FineReader ve Miletos) karşılaştırılmıştır. Karşılaştırma için kullanılan kriterler:

- **Karakter Tanıma Doğruluğu**
- **Bağlı Karakter Katarı Tanıma**
- **Kelime Tanıma Doğruluğu**

Hibrit modelin başarı oranları, çeşitli testler sonucunda aşağıda belirtilmiştir:

- **Karakter Tanıma:** %88.86 ham, %96.12 normalize ve %97.37 bitişik metin.
- **Bağlı Karakter Katarı Tanıma:** %80.48 ham, %91.60 normalize ve %97.37 bitişik.
- **Kelime Tanıma:** %44.08 ham, %66.45 normalize.

Hata Analizi ve İstatistiksel Çalışmalar

Araştırmada, yanlış tanıma oranlarını azaltmaya yönelik bir hata analizi yapılmıştır. Bu analizde, karakterlerin farklı yazımı, karakterler arasındaki yakınlık ve benzerliklerin tanınması gibi faktörler incelenmiştir. Ayrıca Osmanlıca karakterlerin, kelimelerin ve cümlelerin sıklık analizi yapılarak, metin içinde yapısal özellikler üzerine önemli bilgiler sunulmuştur. Bu veriler, ilerideki çalışmalarda yapılacak olan daha karmaşık dil modellerinin gelişiminde yararlı olacaktır.

Çıktılar ve Uygulama

Geliştirilen OCR aracı, çevrimiçi bir platformda kullanıma sunulmuştur ve bu sistem, profesyonel tarihçiler, araştırmacılar ve akademisyenler tarafından Osmanlıca belgelerin dijitalleştirilmesi için kullanılabilir. Proje, TÜBİTAK destekli Osmanlıcadan Günümüz Türkçesine Yapay Zeka Destekli Uçtan Uca Aktarım Projesi çerçevesinde geliştirilmiştir.

Sonuç

Bu çalışma, Osmanlıca metinlerin dijital ortamlara aktarılması konusunda önemli bir gelişme sunmaktadır. Derin öğrenme teknikleri ile OCR sistemleri birleştirilerek, tarihsel belgelerin korunması, dijitalleştirilmesi ve günümüz Türkçesine aktarımı sağlanmış olmuştur. Ayrıca, bu çalışma, Osmanlıca karakter tanıma alanında yenilikçi bir yöntem sunduğu için, gelecekteki dil işleme çalışmaları için de önemli bir temel oluşturmaktadır. İleriye dönük olarak, bu alandaki kısıtlamaların aşılması ve daha fazla veri seti ile sistemin daha da geliştirilmesi planlanmaktadır. Ayrıca, OCR sürecinde karşılaşılan hataların azaltılması için yeni algoritmalar ve teknikler üzerinde çalışmalar yapılması önerilmektedir.

Bu ayrıntılı çalışma, sadece Osmanlı kültür ve tarihi açısından değil, genel olarak dil işleme teknolojileri açısından da büyük bir katkı sağlamaktadır.