Derin sinir ağlarıyla Osmanlıca optik karakter tanıma

Bu çalışma kapsamında, Osmanlıca (Arap alfabesi temelli) belgelerin dijital ortama aktarılmasını sağlayan web tabanlı bir **Derin Sinir Ağları (CNN)** destekli **Optik Karakter Tanıma (OCR)** sistemi geliştirilmiştir.

Üç farklı veri seti hazırlanmıştır:Orijinal veri seti,sentetik veriseti,hibrit veriseti

Hazırlanan bu veri setleri ile **üç farklı OCR modeli** eğitilmiştir.

Önerilen sistemin performansı, piyasadaki diğer araçlarla (**Arapça Tesseract OCR ve Miletos**) 21 sayfalık test seti üzerinden karşılaştırılmıştır. **Karakter, Katar ve Kelime** tanıma başarımı ölçülmüştür. Önerilen hibrit model diğer sistemlere göre belirgin şekilde yüksek doğruluklar sağlamıştır: Kelime tanıma ;ham olarak %44.08 Normalize olarak : %66.45

OCR Süreci Aşamaları (Bütüncül Yaklaşım):

1. Dokümanın Görüntü Olarak Alınması
2. OCR ile Metin Tanıma
3. Osmanlıca Harf Çevirimi (Arapça → Latin)
4. Osmanlıca-Türkçe Dil Çevirisi

Genel Olarak Osmanlıca Alfabe ve Harf/Katar/Kelime Sıklıkları

Başarılı bir OCR sistemi geliştirmek için Osmanlıca harflerin karakteristiklerini iyi anlamak önemlidir. Osmanlıca alfabesi toplam 35 harften oluşur; harfler genellikle sağdan sola bitişik olarak yazılır ve yazım şekilleri harfin kelime içindeki konumuna göre değişir. Harfler; gövde yapısı, noktaların varlığı ve konumu, bitişme durumu gibi özelliklerine göre gruplandırılarak OCR performansı artırılabilir.

Osmanlıca metinlerde:

- **En sık kullanılan 20 harf**, metinlerin %94'ünü oluşturur.
- Katarlar genelde 1-4 harf uzunluğundadır ve sıklıkla sesli harflerle (ا, ر, و, د) sonlanırlar.
- Kelimelerin çoğu 4-5 harften oluşmaktadır ve sık kullanılan kelimeler genellikle bağlaç, edat, zamir, sıfat ve fiillerdir.

Bu özelliklerin analizi, OCR sürecinde karakter tanımayı kolaylaştırarak daha yüksek doğruluk oranları sağlamaktadır. Ayrıca Osmanlıca OCR'da kavislilik, benzer harf şekilleri, noktalama işaretleri ve farklı yazım biçimleri gibi zorluklar mevcuttur. Bu özelliklerin detaylı analizi OCR başarımını artırmakta ve sistemin güvenilirliğini sağlamaktadır.

Benzer Çalışmalar

1. Osmanlıca OCR Çalışmaları:

- Osmanlıca OCR çalışmaları 2000'li yılların başlarında başlamıştır.
- Başlangıçta OCR'ın zorluğu nedeniyle görüntü eşleştirme, içerik tabanlı arama gibi alternatif yöntemler geliştirilmiştir.
- 2000'li yıllarda klasik makine öğrenme yöntemleri yaygın olarak kullanılmıştır:
    - **Support Vector Machines (SVM)**
    - **Hidden Markov Modelleri (HMM)**
    - **Linear Discriminant Analysis (LDA)**
    - **Yapay Sinir Ağları (ANN)**
    - **Çizge-tabanlı algoritmalar**
- Çalışmalar çoğunlukla karakter tanıma ve karakter-katar bölütleme gibi alt problemlere yoğunlaşmıştır.

2. Arap-Tabanlı Alfabelerde Derin Öğrenme ile OCR:

- Son yıllarda derin öğrenme (özellikle CNN ve RNN) modellerinin OCR uygulamalarında yaygınlaşmasıyla yüksek doğruluklar elde edilmiştir.
- Arap alfabesi tabanlı OCR uygulamaları için:
    - CNN+RNN mimarisi kullanılarak Kur'an harf tanıma .
    - Çocukların Arapça el yazısını tanıma için CNN modeli (%88 doğruluk) .
    - Parça interpolasyonu (Segment Interpolation) yöntemiyle çizgi-parçası temelli tanıma .

3. Alternatif Yaklaşımlar:

- Bazı çalışmalar OCR yerine doğrudan Osmanlıca harflerin Türkçe harflere dönüşümüne odaklanmıştır .

Derin Öğrenme Mimarisi

OCR işlemi için bu çalışmada **CNN ve RNN** temelli birleşik (**CRNN**) bir derin sinir ağı modeli kullanılmıştır. Bu mimari, özellikle görüntüden metne dönüştürme için yüksek doğruluk sağlayan bir yapıya sahiptir.

CNN Mimarisi:

- CNN katmanları, satır görüntülerini soldan sağa tarayarak harflerin görsel özelliklerini çıkarır.
- İlk aşamada **evrişim (convolutional)** ve ardından **havuzlama (pooling)** katmanları kullanılarak önemli görsel öznitelikler elde edilir.
- Bu katmanlar sayesinde görüntü verisi soyut ve üst düzey özelliklere dönüştürülür. Bu özellikler eğitim sırasında otomatik olarak belirlenir.

RNN (İki Yönlü LSTM) Mimarisi:

- OCR'da, metin satırlarındaki harflerin dizisel özelliğini yakalamak için **çift yönlü LSTM (Bidirectional LSTM)** katmanları kullanılmıştır.
- Bu katman, harflerin dizideki önceki ve sonraki bağlam bilgisini etkin biçimde öğrenerek doğruluğu artırmaktadır.

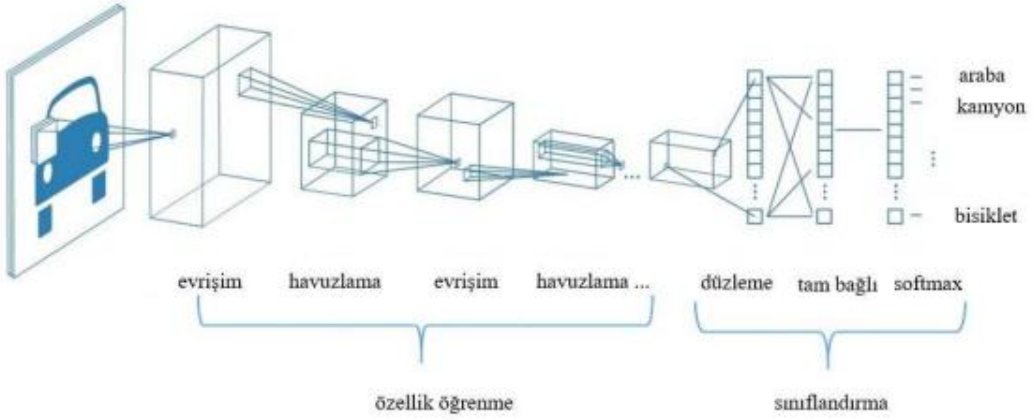
Kullanılan CRNN Mimarisi Detayları:

- Mimari 4 ana katmandan oluşmaktadır:
    1. **CNN katmanı**: Görüntülerden öznitelik çıkarımı
    2. **RNN katmanı** (4 LSTM katmanlı, iki yönlü): Bağlamsal dizisel bilgiler çıkarır.
    3. **Tam bağlantılı katman**: Çıkan öznitelikleri hazırlar.
    4. **CTC (Connectionist Temporal Classification)** katmanı: Harf dizilerinin doğru bir şekilde etiketlenmesini sağlar.

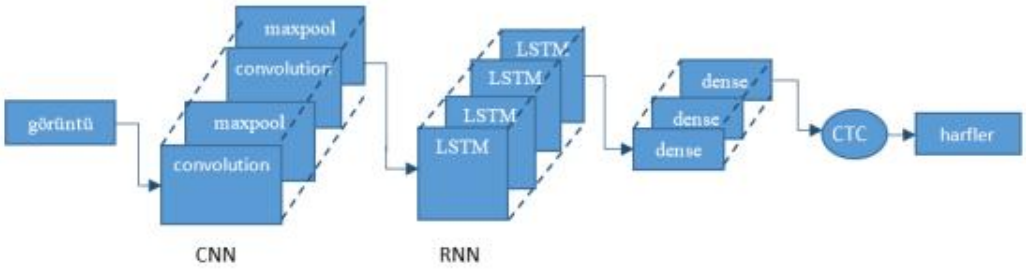
Veri İşleme ve Eğitim Süreci:

- Dokümanlar **satırlara, karakterlere ve katarlara** bölünerek etiketlenir.
- Bölütleme işlemi için **ImageMagick** ve **OpenCV** kütüphaneleri kullanılmıştır.
- Eğitim sırasında satır görüntüleri girdi, metin satırları ise çıktı olarak CRNN modeline verilir.
- Eğitim parametreleri:
    - Öğrenme oranı (**learning rate**): 0.002

- Momentumlu optimizasyon (**SGDM**) kullanılmıştır.
- Epoch sayısı: ~250 iterasyonda eğitim tamamlanmıştır.



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 (Data SET)

Osmanlica.com OCR sisteminin eğitimi için üç farklı veri seti kullanılmıştır:

Eğitim verisi orijinal, sentetik ve hibrit olmak üzere üç farklı kümeden oluşmaktadır:

Orijinal Veri Kümesi: Osmanlıca basılı belgelerden **~1000 sayfa** toplanmıştır. Yarı otomatik yöntemlerle etiketlenmiştir.

Sentetik Veri Kümesi: Orijinal veri toplamak zaman alıcı olduğundan sentetik olarak oluşturulmuştur.Farklı kaynaklardan elde edilen metinler, **70 farklı Arapça font** kullanılarak görüntüye dönüştürülmüştür.

Hibrit Veri Kümesi: Yukarıdaki iki veri setinin birleşimidir (**Orijinal + Sentetik**).Eğitimde çeşitlilik sağlayarak model performansını artırmıştır.

Test Veri Kümesi

Test kümesi, 8 farklı Osmanlıca eserden seçilmiş 21 orijinal sayfa görüntüsünden oluşturulmuştur.

Test verileri, eğitim aşamasında hiç kullanılmayan, tamamen bağımsız örneklerden oluşmaktadır.

Kalitesi düşük, harfleri silik ve farklı kâğıt renklerinde sayfalar seçilerek gerçek kullanım şartlarına yakın test koşulları sağlanmıştır.

Hazırlanan test veri seti, herkesin erişimine açık şekilde osmanlica.com/test adresinde paylaşılmıştır.

Deney, Karşılaştırma ve Sonuçlar

Bu bölümde, Osmanlica.com projesinde geliştirilen OCR modelinin performansı, ticari, ücretsiz ve açık kaynaklı diğer OCR araçları (Tesseract Arapça/Farsça, Abby FineReader, Google Docs ve Miletos) ile karşılaştırılmıştır. Karşılaştırma, model eğitiminde kullanılmayan bağımsız bir test kümesi üzerinde gerçekleştirilmiş ve doğruluk oranları, Python'un difflib kütüphanesi kullanılarak hesaplanmıştır. Doğruluk ölçümü, OCR çıktısı ile referans metin arasındaki eşleşen karakter, katar ve kelime birimleri üzerinden yapılmış; ham metin, normalize metin ve bitişik metin olmak üzere üç farklı metin türü kullanılmıştır.

Deneysel sonuçlarda, ham metinde karakter tanıma doğrulukları %73 ile %89 arasında değişirken, normalize edilmiş metinde bu oran %78'den %96'ya yükselmiştir. Bitişik metin üzerinden hesaplanan doğruluk oranlarında ise Tesseract Farsça (%80), FineReader (%81), Tesseract Arapça (%81), Miletos (%87), Google Docs (%93) ve Osmanlica.com (%97) gibi sıralamalar elde edilmiştir. Osmanlica.com'un %88,64 (ham), %95,92 (normalize) ve %97,18 (bitişik) doğruluk oranları, rakiplerine göre yaklaşık %4 daha iyi performans göstermiştir. Ayrıca, karakter, katar ve kelime tanıma hata analizleri, hataların çoğunlukla belirli pozisyonlarda yoğunlaştığını ve özellikle noktalı harflerde hata oranının noktasızlara göre iki kat daha yüksek olduğunu ortaya koymuştur.

Bölümün son kısmında, modelin performansını etkileyen hiper parametrelerin (örneğin CNN filtre boyutu, aktivasyon fonksiyonu, LSTM yapılandırması ve öğrenme hızı) manuel arama yöntemiyle incelendiği, ancak elde edilen sonuçlarda belirgin bir doğruluk artışı sağlanamadığı; gelecekte sistematik hiper parametre arama yöntemleriyle daha yüksek performanslı modellerin geliştirilebileceği vurgulanmıştır.

Sonuç Olarak

Bu çalışmada, matbu nesih hattı kullanılarak oluşturulan Osmanlıca OCR sistemi, mevcut en gelişmiş ticari ve açık kaynaklı OCR araçlarıyla karşılaştırılmıştır. Şimdiye kadar, Google Docs, Tesseract ve Abby FineReader gibi güçlü araçlarla Osmanlıca OCR performansını doğrudan karşılaştıran tek bir çalışma bulunmamaktadır. Çalışmada, öncelikle Osmanlıca metinlerin normalize edilmesinin gerekliliği vurgulanmış; normalize

edilmiş metinler üzerinden karakter, katar ve kelime tanıma oranları hesaplanmış ve bu kapsamda ilk kez sınırlı da olsa Osmanlıca dil modelleri sunulmuştur.

Geliştirilen Osmanlica.com Hibrit OCR modeli, ham metinde %88,86, normalize metinde %96,12 ve bitişik metinde %97,37 doğruluk oranıyla diğer araçlardan yaklaşık %4 daha üstün performans sergilemiştir. Benzer şekilde, katar tanıma oranlarında da hibrit model, hamda normalize metinlerde diğer araçlardan daha başarılı sonuçlar vermiştir. Kelime tanıma oranlarında ise hibrit model, %44,08 (ham) ve %66,45 (normalize) doğruluk oranlarıyla belirgin üstünlük göstermiştir.

Ayrıca, hiper parametre kestirimi çalışmaları kapsamında filtre boyutu, öğrenme hızı, LSTM boyutu ve aktivasyon fonksiyonu gibi parametreler incelenmiş; ancak yalnızca öğrenme hızının değiştirilmesi anlamlı bir doğruluk artışı sağlamıştır. Çalışmanın sınırlamaları arasında yalnızca matbu nesih hattının hedeflenmesi, hemze ve med işareti taşıyan harflerin tanınmasında eksiklikler ve OCR sonrası karakter düzeltme adımının bulunmaması yer almaktadır. İlerleyen çalışmalarda bu kısıtların giderilerek, Osmanlıca-Türkçe uçtan uca aktarım sürecindeki kritik Osmanlıca OCR adımının başarısının artırılması hedeflenmektedir.