Derin sinir ağlarıyla Osmanlıca Optik Karakter Tanıma

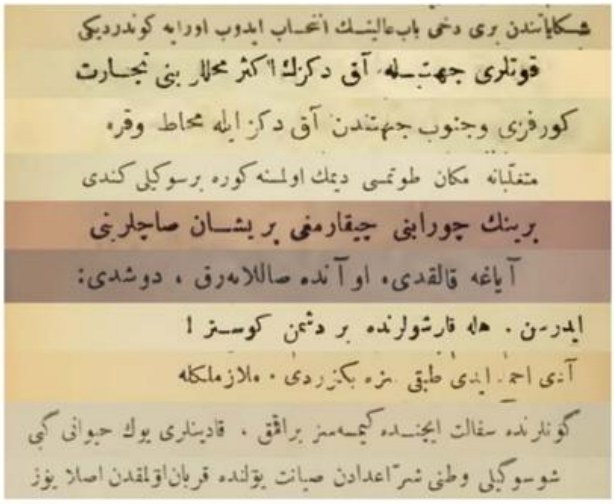
Osmanlıca ve Önemi:
Osmanlıca, **13. yüzyıldan 20. yüzyıla kadar Osmanlı İmparatorluğu'nda kullanılan** ve **Arap alfabesiyle yazılan bir yazı dili**dir. Arapça ve Farsça kelimelerin yoğun şekilde kullanıldığı Osmanlıca, günümüzde Latin alfabesine geçiş yapılması ve kelimelerin büyük kısmının kullanım dışı kalması nedeniyle **okuması ve anlaması zor bir dil** haline gelmiştir.

Osmanlıca Arşivlerinin Önemi ve Erişim Zorluğu:
Osmanlı dönemine ait **kitaplar, dergiler, gazeteler, defterler, kayıtlar ve belgeler** büyük bir kültürel, sanatsal ve tarihi miras barındırmaktadır. Ancak, bu **kaynaklardaki bilgilere hızlı ve etkin şekilde ulaşmak** günümüz insanı için oldukça zor olduğundan, **teknolojik çözümler** gereklidir.

Osmanlıca Belgelerin Bilgisayarlı Aktarımı:
Proje kapsamında **Osmanlıca dokümanların tamamen otomatik olarak Türkçeye aktarımı** dört aşamada gerçekleştirilir:

1. **Doküman-görüntü dönüşümü:** Osmanlıca belgeler taranarak dijital hale getirilir.
2. **Görüntü-metin dönüşümü (OCR - Optik Karakter Tanıma):** Görüntü halindeki Osmanlıca belgeler, metne çevrilir.
3. **Osmanlıca-Türkçe alfabe çevirisi (harfçevrim):** Arap harfleriyle yazılmış Osmanlıca metin, Latin harflerine çevrilir.
4. **Osmanlıca-Türkçe dil çevirisi:** Harf çevirimi tamamlanan metin, anlam kaybı olmadan Türkçeye çevrilir.

Nesih Yazı Hattı:
Şekil 1'de gösterildiği gibi, **nesih hattı** Osmanlı'nın son dönemlerinde **kitap, gazete ve dergilerde yaygın olarak kullanılan bir yazı tipidir**. Projede, Osmanlıca metinlerin OCR ile tanınması ve Latin harflerine aktarılması sırasında özellikle **nesih hattı karakterlerinin doğru bir şekilde tespit edilmesi** kritik bir aşamadır.



Şekil 1. Osmanlıca matbu nesih hattı örnekleri
(Ottoman documents in printed naksh font)

Derin Öğrenme Mimarisi (Deep Learning Architecture)

Bu bölümde, OCR (Optik Karakter Tanıma) için kullanılan derin sinir ağları tabanlı bir derin öğrenme modeli olan **CRNN (Convolutional Recurrent Neural Network)** mimarisi tanıtılmaktadır. CRNN, **CNN (Convolutional Neural Network)** ve **RNN (Recurrent Neural Network)** mimarilerinin birleşiminden oluşan

bir modeldir. Bu mimari, görüntüdeki metinleri tanımak ve karakter dizilerini doğru bir şekilde tahmin etmek için kullanılır.

CNN (Convolutional Neural Network) Mimarisi

CNN, görüntüdeki görsel örüntüleri (örneğin kenarlar, dokular, harf şekilleri) tanımak için kullanılır. Girdi görüntüsü, CNN katmanlarından geçerek işlenir. İlk katman olan **evrişim (convolutional) katmanı**, görüntüdeki özellikleri (örneğin kenarlar, köşeler) belirler. Bu katmanda, doğrusal olmayan aktivasyon fonksiyonları (örneğin ReLU) kullanılarak özellik haritaları oluşturulur. Ardından, **havuzlama (pooling) katmanı** ile özellik haritalarının boyutu küçültülür ve özellik sayısı azaltılır. Bu işlem, hesaplama karmaşıklığını düşürür ve modelin genelleme yeteneğini artırır. Daha sonra, iki boyutlu özellik haritaları, **düzleme (flattening) katmanı** ile tek boyutlu bir vektöre dönüştürülür. Bu vektör, sonraki katmanlarda kullanılmak üzere hazırlanır. Son olarak, **tam bağlı (fully-connected) katmanlar** ile özellikler sınıflandırılır. Son katmanda genellikle **softmax** fonksiyonu kullanılarak sınıflandırma tamamlanır.

CNN katmanları, girdi verisini giderek daha soyut ve üst düzey özelliklere dönüştürür. Örneğin, ilk katmanda kenarlar, ikinci katmanda harf parçaları, üçüncü katmanda harfler tanınır. Hangi özelliklerin seçileceği ve hangi özelliklerin hangi katmanda temsil edileceği, eğitim sırasında belirlenir. Bu sayede, özellik belirleme ve özellik seçimi adımları derin öğrenme içinde çözümlenir.

RNN (Recurrent Neural Network) Mimarisi

RNN, zaman serisi verilerini işlemek için kullanılır. OCR'de, metin karakterlerinin bir dizisi olarak modellenmesi gerektiğinden, RNN bu tür problemlerde etkilidir. RNN'in bir türü olan **LSTM (Long Short-Term Memory)**, uzun süreli bağımlılıkları öğrenebilir. Özellikle **iki yönlü LSTM (Bidirectional LSTM)**, hem önceki hem de sonraki karakterler arasındaki bağlamı öğrenerek daha doğru tahminler yapabilir. LSTM'in **CTC (Connectionist Temporal Classification)** ile birlikte kullanımı, birçok NLP (Doğal Dil İşleme) probleminde başarılı sonuçlar vermektedir. CTC katmanı, LSTM'in çıktılarını yorumlar ve karakter dizilerini etiketler. CTC, girdi ve çıktı dizileri arasındaki hizalamayı öğrenir ve her bir karakterin olasılığını hesaplar.

CRNN Mimarisi

CRNN mimarisi, önce CNN katmanları ile görüntüdeki özellikleri çıkarır, ardından RNN (özellikle LSTM) ile karakter dizilerini tahmin eder. Son olarak, CTC katmanı ile karakterler etiketlenir. Mimarinin **CNN bölümü**, 16 filtreli 3x3 evrişim katmanı ve 3x3'lük max-pooling katmanı içerir. **RNN bölümü** ise 4 katmanlıdır. İlk katman 64 çıktılı bir LSTM, ikinci ve üçüncü katmanlar 128 çıktılı iki yönlü LSTM, dördüncü katman ise 256 çıktılı ileri yönlü bir LSTM içerir. RNN'den sonra, sınıflandırmanın yapıldığı bir **yoğun (dense) katman** ve bu katman çıktısını yorumlayan bir **CTC fonksiyonu** kullanılır.

Eğitim Süreci

Eğitim sırasında, dokümanlar satırlara bölünür. Satır görüntüleri CRNN'e girdi olarak verilir, satırlardaki metin ise çıktı olarak kullanılır. Eğitim hiper parametreleri şu şekildedir: **Learning Rate: 0.002**, **Momentum: 0.5**, **Epochs: 3.000.000**. Görüntüler, satır, kelime ve karakterlere bölünür. Bu işlem için **ImageMagick** ve **OpenCV** kütüphaneleri kullanılır. Algoritmik olarak üretilen **sentetik veri**, bölütleme işlemini kolaylaştırır ve eğitim sürecini hızlandırır.

CRNN mimarisi, özellik belirleme ve seçimini otomatik olarak gerçekleştirir. Veri, katmanlar boyunca sıkıştırılarak ve soyutlanarak işlenir. Derin öğrenmedeki "derinlik" kavramı, verinin dönüşüme uğradığı ara katmanları ifade eder. Bu katmanlar, verinin daha üst düzey özelliklerini temsil eder. CNN ve RNN'in birleşimi olan CRNN, OCR gibi karmaşık problemlerde başarılı sonuçlar vermektedir. Özellikle LSTM ve CTC'nin birlikte kullanımı, karakter dizilerinin doğru bir şekilde tanınmasını sağlar.

Veri Kümesi (Data Set)

Bu bölümde, Osmanlıca metinlerin OCR (Optik Karakter Tanıma) modelini eğitmek için kullanılan **eğitim veri kümesi** ve modelin performansını değerlendirmek için kullanılan **test veri kümesi** tanıtılmaktadır. Veri setleri, sayfa, satır, kelime ve karakter sıklıkları açısından detaylandırılmıştır.

Eğitim Veri Kümesi (Training Data Set)

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

1. **Orijinal Veri**:
    - **Kaynak**: Değişik Osmanlıca eserlerden yaklaşık 1000 sayfa görüntü toplanmıştır.
    - **İşlem**: Bu görüntüler yarı otomatik yöntemlerle metin dosyalarına dönüştürülmüştür.
    - **İçerik**:
        - 18 bin satır,
        - 35 bin kelime,
        - 252 bin karakter.
    - **Görüntü Özellikleri**:
        - Sayfa boyutu: 1400 x 2000 piksel,
        - Çözünürlük: 300 dpi,
        - Ortalama satır sayısı: 20,
        - Font boyutu: 12 nokta,
        - Satır yüksekliği: 48 nokta.
2. **Sentetik Veri**:
    - **Oluşturma Yöntemi**: Orijinal veri hazırlamak uzun ve zahmetli olduğundan, metin-görüntü dönüşüm araçları kullanılarak sentetik veri üretilmiştir.
    - **İçerik**:
        - 26 bin sayfa,
        - 1.3 milyon satır,
        - 263 bin kelime,
        - 78 milyon karakter.
    - **Görüntü Özellikleri**:
        - Sayfa boyutu: 2500 x 4800 piksel,
        - Çözünürlük: 300 dpi,
        - Font boyutu: 12 nokta,
        - Ortalama satır sayısı: 42,
        - Satır yüksekliği: 48 nokta.
    - **Font Çeşitliliği**: 70 farklı Arapça fontu kullanılmıştır.
3. **Hibrit Veri**:
    - **Oluşum**: Orijinal ve sentetik veri kümelerinin birleşimidir.

Test Veri Kümesi (Test Data Set)

- **Kaynak**: 8 farklı Osmanlıca eserden 21 orijinal sayfa görüntüsü kullanılmıştır.
- **Seçim Kriterleri**: Test kümesi, eğitimde kullanılmayan orijinal sayfalardan oluşur. Kalitesi düşük, harfleri silik ve farklı kâğıt renklerine sahip örnekler özellikle seçilmiştir.
- **İçerik**:
    - 21 sayfa,
    - 420 satır,
    - 3 bin kelime,
    - 23 bin karakter.
- **Ortalama Özellikler**:
    - Her sayfada ortalama 20 satır,
    - Her satırda ortalama 7 kelime ve 55 karakter.
- **Paylaşım**: Test kümesi, **osmanlica.com/test** adresinde paylaşılmıştır. Bu paylaşım şunları içerir:
    - Görüntü dosyaları,
    - Altışar adet OCR test çıktısı,
    - Doğru metin dosyaları,
    - **difflib** kütüphanesi ile doğruluk oranı hesaplayan bir Python dosyası.

Veri Kümesi Sıklıkları (Dataset Frequencies)

- **Sentetik Veri**:
  - Sayfa: 26 bin,
  - Satır: 1.3 milyon,
  - Kelime: 263 bin,
  - Karakter: 78 milyon.
- **Orijinal Veri**:
  - Sayfa: 1 bin,
  - Satır: 18 bin,
  - Kelime: 35 bin,
  - Karakter: 252 bin.
- **Eğitim Verisi (Toplam)**:
  - Sayfa: 27 bin,
  - Satır: 1.3 milyon,
  - Kelime: 298 bin,
  - Karakter: 78 milyon.
- **Test Verisi**:
  - Sayfa: 21,
  - Satır: 420,
  - Kelime: 3 bin,
  - Karakter: 23 bin.

CRNN Hiper Parametreleri (CRNN Hyperparameters)

- **Conv2D (3x3, 16 filtre + tanh)**: 160 parametre.
- **MaxPool2D (3x3)**: 0 parametre.
- **LSTM1 (64 çıktı, y ekseninde)**: 20,736 parametre.
- **LSTM2 (128 çıktı, x ekseninde ileri)**: 98,816 parametre.
- **LSTM3 (128 çıktı, x ekseninde geri)**: 131,584 parametre.
- **LSTM4 (256 çıktı, x ekseninde)**: 394,240 parametre.
- **Yoğun Katman (71 düğüm)**: 18,247 parametre.
- **Toplam Parametre Sayısı**: 663,703.

Deney, Karşılaştırma ve Sonuçlar (Experiment, Comparison, and Results)

Karşılaştırmada Kullanılan OCR Araçları (OCR Tools Used in Comparison)

- **Tesseract Arapça ve Farsça**: Açık kaynaklı ve ücretsiz bir OCR aracıdır. Arapça ve Farsça için eğitilmiş modeller kullanılmıştır.
- **Abby FineReader**: Ticari bir OCR aracıdır. Arapça ve Farsça dahil birçok dilde yüksek performans sunar.
- **Google Docs**: Google'ın ücretsiz OCR hizmetidir. Online olarak kullanılır ve yaygın olarak tercih edilir.
- **Miletos**: Osmanlıca için özel olarak geliştirilmiş ticari bir OCR aracıdır.

6.2. Doğruluk Oranı için Kullanılan Metin Türleri (Text Types Used for Recognition Accuracy)

- **Ham Metin**: Herhangi bir işlemden geçmemiş metin. Hatalar nedeniyle doğruluk oranları yanıltıcı olabilir.

- **Normalize Metin**: Boşluk, BIDI (yazı yönü) ve karakter normalizasyonu işlemlerinden geçirilmiş metin.
- **Bitişik Metin**: Boşlukların kaldırıldığı ve kelimelerin bitiştirildiği metin. Kelime bölümleme hatalarını önlemek için kullanılır.

6.3. Karakter Tanıma Oranları (Character Recognition Rates)

- **Ham Metin**: Doğruluk oranları %73 ile %89 arasında değişmektedir. En yüksek performans Osmanlıca Hibrit modelindedir (%89).
- **Normalize Metin**: Doğruluk oranları %78 ile %96 arasındadır. Osmanlıca Hibrit modeli %96 doğruluk oranıyla en iyi sonucu vermiştir.
- **Bitişik Metin**: Doğruluk oranları %80 ile %97 arasındadır. Osmanlıca Hibrit modeli %97 doğruluk oranıyla liderdir.

6.4. Katar Tanıma Oranları (Ligature Recognition Accuracy)

- **Ham Metin**: Doğruluk oranları %51 ile %80 arasındadır. Osmanlıca Hibrit modeli %80 ile en iyi performansı gösterir.
- **Normalize Metin**: Doğruluk oranları %61 ile %91 arasındadır. Osmanlıca Hibrit modeli %91 doğruluk oranıyla öne çıkar.

6.5. Kelime Tanıma Oranları (Word Recognition Accuracy)

- **Ham Metin**: Doğruluk oranları %15 ile %44 arasındadır. Osmanlıca Hibrit modeli %44 ile en iyi sonucu verir.
- **Normalize Metin**: Doğruluk oranları %24 ile %66 arasındadır. Osmanlıca Hibrit modeli %66 doğruluk oranıyla liderdir.

6.6. Harf Türüne Göre Karakter Tanıma Oranları (Character Recognition Accuracy by Letter Type)

- **Arapça Harfler**: En yüksek hata oranları bu grupta görülmektedir.
- **Osmanlıca Harfler**: Frekansları düşük olmasına rağmen, hata oranları yüksektir.
- **Noktalı Harfler**: Noktasız harflere göre daha yüksek hata oranlarına sahiptir.
- **Nokta Sayısı**: Nokta sayısı arttıkça hata oranları artmaktadır.

6.7. Hiper Parametre Kestirimi (Hyper-Parameter Tuning)

- **Hiper Parametreler**: Filtre boyutu, aktivasyon fonksiyonu, LSTM katman boyutu ve öğrenme hızı gibi parametreler test edilmiştir.
- **Sonuçlar**: Hiper parametre değişiklikleri, doğruluk oranlarında önemli bir artış sağlamamıştır. En iyi sonuçlar orijinal hiper parametrelerle elde edilmiştir.

Genel Sonuçlar

- **Osmanlıca Hibrit Modeli**: Hem karakter hem de kelime tanıma oranlarında en yüksek performansı gösterdi (%97 bitişik metin doğruluk oranı).
- **Diğer Araçlar**: Tesseract, Abby FineReader ve Google Docs, Osmanlıca metinlerde daha düşük performans sergiledi.

- **Hata Analizi**: Hataların büyük çoğunluğu karakter tanıma hatalarından kaynaklanmaktadır. Kelime bölümleme ve bitişme hataları da önemli bir sorun teşkil etmektedir.

Sonuçlar (Conclusions)

Bu çalışma, Osmanlıca OCR (Optik Karakter Tanıma) alanında önemli bir adım olarak değerlendirilebilir. Osmanlıca metinlerin dijital ortama aktarılması, hem tarihsel belgelerin korunması hem de araştırmacılar için erişilebilir hale getirilmesi açısından büyük bir öneme sahiptir. Bu çalışmada, Osmanlıca OCR için geliştirilen bir model, piyasadaki diğer yaygın OCR araçlarıyla karşılaştırılmış ve önemli sonuçlar elde edilmiştir.

Karşılaştırma ve Analiz

Çalışmada, Osmanlıca için özel olarak geliştirilmiş **Miletos** OCR aracı, **Google Docs**, **Tesseract** (Arapça ve Farsça modelleri) ve **Abby FineReader** gibi yaygın OCR araçlarıyla karşılaştırılmıştır. Bu karşılaştırma, Osmanlıca metinlerin tanınmasında hangi araçların daha etkili olduğunu ortaya koymuştur. Özellikle **Osmanlıca Hibrit modeli**, hem karakter hem de kelime tanıma oranlarında diğer araçlardan belirgin şekilde daha yüksek performans göstermiştir.

Normalizasyonun Önemi

Osmanlıca metinlerin normalizasyonu (boşluk, yazı yönü ve karakter normalizasyonu) bu çalışmada vurgulanan önemli bir konudur. Normalizasyon işlemi, metinlerin standart bir forma dönüştürülmesini sağlayarak OCR performansını artırmaktadır. Bu çalışmada, metinleri normalize eden bir Python programı paylaşıma açılmıştır.

Kapsamlı Performans Analizi

Bu çalışma, sadece karakter tanıma oranlarını değil, aynı zamanda **katar** ve **kelime tanıma oranlarını** da ham, normalize ve bitişik metin üzerinde hesaplayarak kapsamlı bir analiz sunmuştur. Özellikle bitişik metin üzerinde yapılan analizler, kelime bölümleme hatalarının etkisini minimize etmiştir.

Geliştirilen OCR Sistemi

Bu çalışmada, Osmanlıca matbu nesih dokümanlarını metne dönüştüren web tabanlı bir OCR sistemi geliştirilmiştir. Sistem, **sentetik** ve **orijinal verilerle** eğitilmiş ve **21 sayfalık bir test veri kümesi** üzerinde değerlendirilmiştir. Test sonuçları, sistemin yüksek doğruluk oranlarına sahip olduğunu göstermiştir.

Hiper Parametre Kestirimi

Hiper parametreler üzerinde yapılan deneylerde, filtre boyutu, öğrenme hızı, LSTM katman boyutu ve aktivasyon fonksiyonu gibi parametreler test edilmiştir. Ancak, bu değişiklikler doğruluk oranlarında önemli bir artış sağlamamıştır. En iyi sonuçlar, orijinal hiper parametrelerle elde edilmiştir.

Çalışmanın Kısıtları ve Gelecek Çalışmalar

Çalışmanın bazı kısıtları bulunmaktadır:

- Sistem sadece **matbu nesih hattı** için tasarlanmıştır,**Hemze** ve **med işareti** taşıyan harfler tanınamamaktadır.,OCR sonrası karakter düzeltme adımı bulunmamaktadır.Bu kısıtların giderilmesine yönelik gelecek çalışmalar planlanmaktadır. Ayrıca, Osmanlıca-Türkçe uçtan uca aktarım sürecinde OCR adımının başarısı, sonraki adımların doğruluğu için kritik öneme sahiptir. Bu nedenle, OCR performansının daha da artırılması hedeflenmektedir.