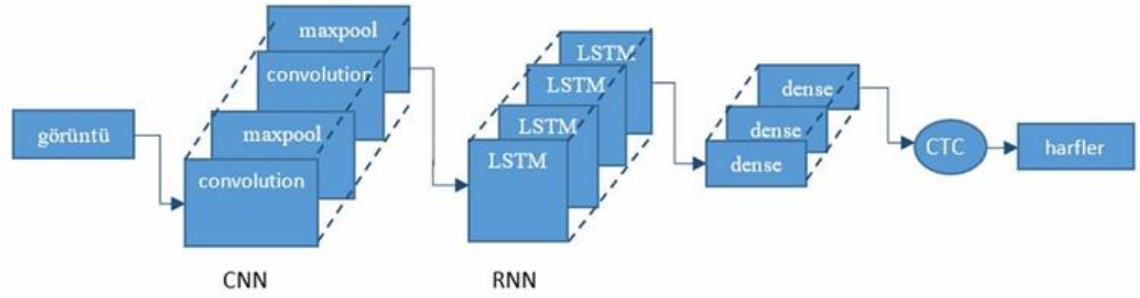
Derin sinir ağlarıyla Osmanlıca optik karakter tanıma

Projenin Amacı ve Gerekçesi

- **Amaç:**
  Osmanlıca matbu nesih belgelerinin dijital metne dönüştürülmesi için, derin öğrenme (OCR) tekniklerini kullanarak hatasız ve yüksek performanslı bir sistem geliştirmek.
- **Gerekçe:**
  Osmanlıca dokümanlar, kültürel mirasın önemli parçalarıdır; ancak eski yazı stili ve karakter özellikleri nedeniyle geleneksel OCR araçları (Tesseract, Google Docs, Abby FineReader, Miletos) yetersiz kalmaktadır.

İlgili Çalışmalar

- **Klasik Yaklaşımlar:**
  SVM, HMM, LDA ve geleneksel yapay sinir ağları kullanılarak yapılan çalışmalar.
- **Derin Öğrenme Yaklaşımları:**
  Son yıllarda CNN, RNN ve CRNN tabanlı yöntemlerle hem Arapça hem de Osmanlıca OCR'de önemli gelişmeler kaydedilmiştir.



Şekil 1: Osmanlıca OCR için CRNN mimarisi (CRNN architecture for Ottoman OCR)

- **Özellikler ve Veri Analizleri:**
  Osmanlıca harflerin, katarların ve kelimelerin frekans dağılımları, karakterlerin ayırt edici özellikleri (nokta, harf gövdesi, bağlantı özellikleri) üzerine analizler yapılmıştır.

Veri Kümeleri

- **Eğitim Verisi:**
  - *Orijinal Veri:* Yaklaşık 1000 sayfa, 18.000 satır, 35.000 kelime, 252.000 karakter
  - *Sentetik Veri:* 26.000 sayfa, 1,3 milyon satır, 263.000 kelime, 78 milyon karakter
  - *Hibrit Veri:* Orijinal ve sentetik verilerin birleşimi

- **Test Verisi:**
  8 farklı eserden seçilen 21 orijinal sayfa; kalite ve yazı tipi çeşitliliğini yansıtacak şekilde belirlenmiş, eğitimde kullanılmamış veriler.

Derin Öğrenme Modeli ve Deneysel Çalışmalar

Model Mimarisi

- **CRNN Yaklaşımı:**
  - *CNN Bölümü:* Görüntüdeki görsel örüntüleri (kenarlar, şekiller) çıkarır; evrişim, havuzlama ve aktivasyon (ReLU, tanh) işlemleriyle çalışır.
  - *RNN Bölümü:* İki yönlü LSTM katmanları kullanılarak, satırdaki karakter dizilerindeki bağlamı ve sırayı öğrenir.
  - *CTC Katmanı:* Karakter dizisi olasılıklarını hesaplayarak, modelin çıktı dizisini oluşturur.

Deneysel Süreç

- **Veri Hazırlama:**
  Belgeler, ImageMagick ve OpenCV yardımıyla satır, kelime ve karakter düzeyinde segmentlere ayrılmıştır.
- **Eğitim ve Test:**
  Model, orijinal, sentetik ve hibrit veri kullanılarak eğitilmiş; eğitim sürecinde hiper parametreler (learning rate, LSTM boyutu, aktivasyon fonksiyonu, filtre boyutu) ayarlanmıştır.
- **Karşılaştırma:**
  Geliştirilen model, Tesseract (Arapça/Farsça), Abby FineReader, Google Docs ve Miletos gibi tanınmış OCR araçlarıyla, ham, normalize ve bitişik metinler üzerinden ölçümlenmiştir.
- **Ölçüm Kriterleri:**
  - Karakter, katar ve kelime tanıma oranları
  - Hata analizinde ekleme, silme ve yer değiştirme oranları (Python difflib SequenceMatcher kullanılarak hesaplanmıştır)

Sonuçlar, Karşılaştırmalar ve Gelecek Perspektifleri

Temel Sonuçlar

- **Karakter Tanıma:**
  Hibrit model, ham metinde %88,86; normalize metinde %96,12; bitişik metinde ise %97,37 doğruluk oranı elde etmiştir.

- **Kelime Tanıma:**
  Kelime tanıma oranları, kelimenin tüm harflerinin doğru tanınmasına bağlı olduğundan daha düşük; ham metinde %15 ila %44, normalize metinde %24 ila %66 arası değerler gözlemlenmiştir. Hibrit model diğer araçlara göre belirgin üstünlük sağlamıştır.
- **Hata Dağılımı:**
  - Karakter hataları, çoğunlukla yanlış tanınan harflerden kaynaklanmakta.
  - Noktalı harflerde hata oranı, noktasızlara göre daha yüksek seyretmektedir.

Hiper Parametre Ayarları ve Deneysel İyileştirmeler

- Yapılan deneylerde; filtre boyutu, LSTM boyutu, aktivasyon fonksiyonu ve öğrenme hızı üzerinde değişiklikler denenmiştir.
- En belirgin iyileşme, öğrenme hızının artırılmasıyla elde edilmiştir.
- Diğer parametrelerde yapılan değişiklikler doğruluk oranında belirgin artış sağlamamıştır.

Gelecek Çalışmalar ve Sonuç Değerlendirmesi

- **Geliştirme Alanları:**
  - OCR sonrası hata düzeltme adımlarının eklenmesi
  - Kelime bölme ve bitişme sorunlarının giderilmesi
  - Daha geniş ve çeşitli veri kümeleri kullanılarak model performansının artırılması
  - Otomatik hiper parametre arama yöntemlerinin (grid search, random search) uygulanması

Model	Ham	Normalize	Bitişik	Değişen	Silinen	Eklenen
Osmanlica Hibrit	88,86	96,12	97,37	1,60	1,93	2,50
Osmanlica Orijinal	87,73	94,87	96,16	2,30	2,50	2,81
Osmanlica Sentetik	73,16	77,64	78,10	14,92	5,77	6,15
Google Docs	83,86	92,02	91,43	4,24	3,19	3,50
Abby FineReader	71,98	80,19	81,05	13,47	8,23	3,45
Tesseract Arabic	76,92	82,37	81,27	12,79	6,15	2,89
Tesseract Persian	75,30	83,85	83,48	11,18	7,14	2,51
Miletos	75,76	86,46	86,88	10,94	6,21	1,57

Tablo 1: Karakter tanıma doğruluk oranı ve normalize metin hata dağılımları (%)

Model	Ham	Normalize	Bitişik	Değişen	Silinen	Eklenen
Osmanlica Hibrit	80,48	91,60	92,14	7,22	0,26	0,21
Osmanlica Orijinal	78,34	89,10	88,75	9,57	0,52	0,39
Osmanlica Sentetik	55,64	61,63	56,59	31,65	3,46	1,61
Google Docs	75,51	83,11	72,63	15,20	0,38	0,41
Abby FineReader	51,52	61,58	57,59	35,57	2,73	1,21
Tesseract Arabic	59,32	65,89	59,05	30,45	1,39	0,99
Tesseract Persian	57,90	66,94	61,47	31,14	0,87	0,90
Miletos	60,56	73,61	69,81	27,63	0,71	0,33

Tablo 2: Katar tanıma doğruluk oranı ve hata dağılımları (%)

Model	Ham	Normalize	Değişen	Silinen	Eklenen
Osmanlica Hibrit	44,08	66,45	31,27	0,56	0,28
Osmanlica Orijinal	40,84	61,13	35,49	0,56	0,64
Osmanlica Sentetik	15,55	24,53	70,86	0,60	2,64
Google Docs	38,64	50,78	44,88	0,47	0,94
Abby FineReader	13,28	24,40	75,01	0,86	0,81
Tesseract Arabic	20,05	26,43	66,95	1,67	6,51
Tesseract Persian	16,59	27,02	69,44	2,09	2,33
Miletos	14,92	31,22	70,80	0,00	1,70

Tablo 3: Kelime tanıma doğruluk oranı ve hata dağılımları (%)

- **Genel Değerlendirme:**
  Geliştirilen Osmanlıca OCR sistemi, diğer mevcut OCR araçlarına kıyasla üstün performans göstererek, hem karakter hem de kelime tanımada önemli başarılar elde etmiştir. Bu sonuçlar, kültürel mirasın dijitalleştirilmesi ve Osmanlıca dokümanların erişilebilirliğinin artırılması açısından büyük önem taşımaktadır.