02210224024 – Emir Taşdemir

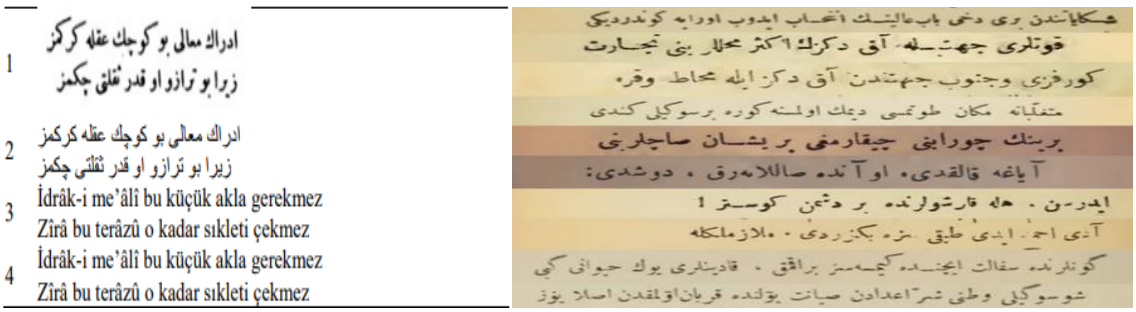
Derin Sinir Ağlarıyla Osmanlıca Optik Karakter Tanıma (OCR)

Giriş

Osmanlıca, 13. ve 20. yüzyıllar arasında Osmanlı İmparatorluğu'nda kullanılan ve Arap alfabesiyle yazılan bir dildir. Osmanlıca belgeler, günümüzde arşivlerde milyonlarca sayfa olarak saklanmaktadır. Ancak, bu belgelerin dijital hale getirilmesi ve okunabilir metne dönüştürülmesi büyük bir zorluk teşkil etmektedir.

Osmanlı alfabesi, Arapça ve Farsça'dan alınmış harflerden oluşmakta olup, kelimeler sağdan sola yazılmaktadır. Harflerin kelime içindeki konumuna bağlı olarak farklı şekillerde yazılması ve harflerin bitişik yazılabilmesi, optik karakter tanıma (OCR) süreçlerinde ciddi zorluklara neden olmaktadır. Osmanlıca metinlerin okunması ve dijital hale getirilmesi, kültürel mirasın korunması ve akademik çalışmalar açısından büyük bir öneme sahiptir.

Optik Karakter Tanıma (OCR) teknolojisi, Osmanlıca belgeleri dijital metne çevirmede kritik bir rol oynayabilir. Ancak, Osmanlı alfabesinin karmaşıklığı, harflerin farklı biçimlerde yazılabilmesi ve karakterlerin birbiriyle karışabilmesi nedeniyle mevcut OCR sistemleri genellikle düşük doğruluk oranlarına sahiptir. Bu çalışmada, derin sinir ağlarıyla Osmanlıca matbu nesih hattı OCR modeli geliştirilmiş ve mevcut OCR araçlarıyla karşılaştırılmıştır.



* Osmanlıca matbu nesih hattı örnekleri

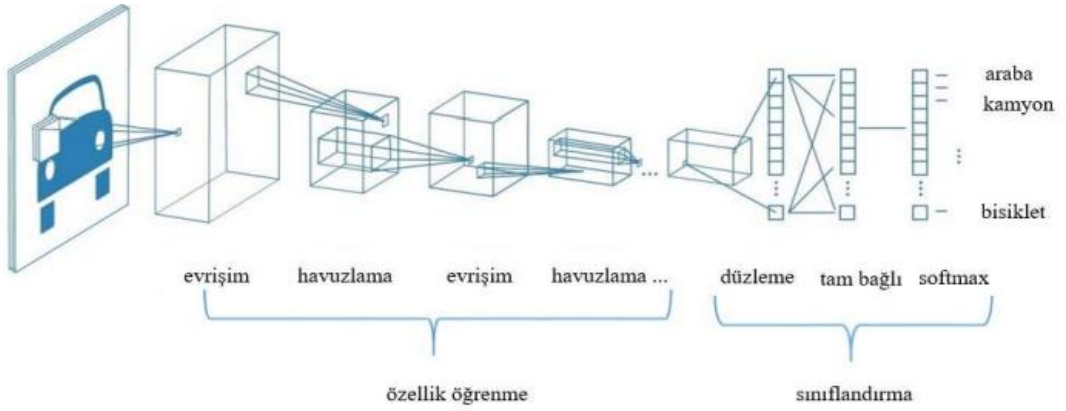
Kullanılan Yöntem ve Teknolojiler

Bu çalışmada, CNN ve RNN tabanlı CRNN (Convolutional Recurrent Neural Network) mimarisi kullanılarak bir OCR modeli geliştirilmiştir. Osmanlıca gibi karmaşık yapıya

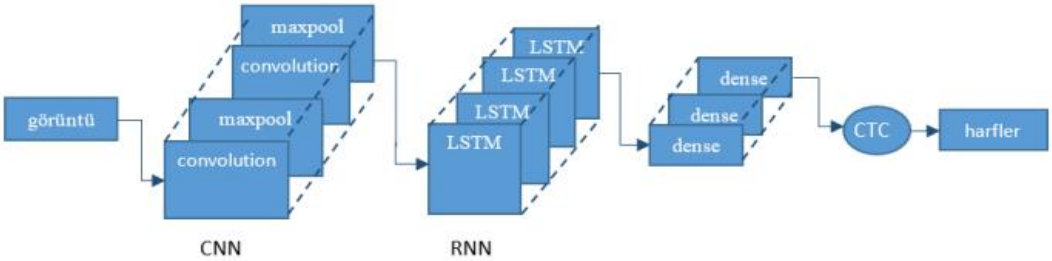
sahip alfabelerin tanınması için geliştirilen bu model, hem görüntü hem de dizilim tabanlı bilgileri öğrenerek daha yüksek doğruluk oranlarına ulaşmayı amaçlamaktadır.

- CNN (Evrişimsel Sinir Ağları): Görüntü tabanlı veri analizinde kullanılarak harflerin karakteristik özelliklerini öğrenir. Osmanlıca harflerin eğimli, yuvarlak ve bazen benzer şekillerde olması nedeniyle, CNN katmanları farklı yazım biçimlerini ayırt edebilmek için derin özellik çıkarımı yapar.
- RNN (Tekrarlayan Sinir Ağları): Osmanlıca kelimeler genellikle birleşik harflerden oluştuğu için, ardışık karakterlerin doğru sırada tanınmasını sağlamak amacıyla RNN katmanları kullanılmıştır. Özellikle LSTM (Long Short-Term Memory) hücreleri ile önceki ve sonraki karakterler arasındaki ilişkiler öğrenilerek tanıma doğruluğu artırılmıştır.
- CTC (Connectionist Temporal Classification): Karakterlerin sıralamasını düzeltmek ve eksik karakterleri tamamlamak için kullanılan bir yöntemdir. Osmanlıca'da kelime içinde harflerin farklı formlarda yazılabilmesi nedeniyle, modelin esneklik kazanması için CTC fonksiyonu kullanılmıştır.

Bu model, Osmanlıca OCR işlemlerinde mevcut yöntemlere göre daha yüksek doğruluk sağlamak için özel olarak tasarlanmıştır.



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 Seti ve Eğitim Süreci

Çalışmada kullanılan veri seti üç ana gruptan oluşmaktadır:

- **Orijinal Veri Seti**: 1000 sayfalık Osmanlıca dokümanlardan oluşturulmuştur. Bu veri seti, Osmanlıca matbu nesih hattı belgelerinden oluşan yüksek kaliteli görüntüler içermektedir.
- **Sentetik Veri Seti**: 70 farklı Arapça fontuyla üretilmiş 23.000 sayfadan oluşmaktadır. Sentetik veri, modelin daha fazla örnekle eğitilmesini ve gerçek dünya senaryolarına daha iyi adapte olmasını sağlamak için kullanılmıştır.
- **Hibrit Veri Seti**: Orijinal ve sentetik veri setlerinin birleşimidir. Bu veri kümesi, hem gerçek Osmanlıca belgelerini hem de yapay olarak üretilmiş belgeleri içererek modelin daha dengeli bir şekilde öğrenmesini sağlar.

Modelin eğitimi sırasında, her veri setinden alınan veriler ile farklı eğitim senaryoları oluşturulmuştur. İlk olarak sadece orijinal veri setiyle eğitim yapılmış, ardından sentetik veri seti eklenerek modelin genelleme kabiliyeti artırılmıştır. Son olarak, hibrit veri seti ile eğitilen model, hem orijinal hem de sentetik verilerde en yüksek doğruluğa ulaşmıştır.

Küme	Sayfa	Satır	Kelime	Karakter
Sentetik	26B	1.3M	263B	78M
Orijinal	1B	18B	35B	252B
Eğitim	27B	1.3M	298B	78M
Test	21	420	3B	23B

*Veri kümesi sıklıkları

Sonuçlar ve Karşılaştırmalar

Geliştirilen model, mevcut OCR araçlarıyla karşılaştırılmıştır. Hybrid OCR modeli, diğer modellere kıyasla en yüksek doğruluğa ulaşmıştır. Karakter, bağlı karakter ve kelime tanıma doğrulukları aşağıdaki gibidir:

Model	Ham Doğruluk (%)	Normalleştirilmiş (%)	Bitişik (%)
Hibrit Model	88.86	96.12	97.37
Orijinal Model	87.73	94.87	96.16
Sentetik Model	73.16	77.64	78.10
Google Docs	83.86	92.02	91.43
Abby FineReader	71.98	80.19	81.05
Tesseract Arabic	76.92	82.37	81.27

Tesseract Persian	75.30	83.85	83.48
Miletos	75.76	86.46	86.88

OCR modelleriyle yapılan karşılaştırmada, hibrit modelin karakter tanıma, bağlı karakter tanıma ve kelime tanıma doğruluklarında en iyi sonuçları verdiği görülmüştür. Bu sonuçlar, derin öğrenme tabanlı OCR modellerinin geleneksel OCR yöntemlerine göre çok daha başarılı olduğunu göstermektedir.

Hiper parametre kestirimi: Derin sinir ağlarında hiper parametre kestirimi, modelin başarısını doğrudan etkiler. Bu parametreler, ağ yapısı, optimizasyon algoritması, aktivasyon fonksiyonu ve öğrenme hızı gibi unsurları içerir. Bu parametrelerin seçimi genellikle sezgisel bir yaklaşımla yapılır ve en iyi değerleri bulmak için deneysel bir süreç gerektirir.

Tablo 23. Hiper parametre kestirim deneyi I: Karakter tanıma
(Hyperparameter estimation experiment I: Character recognition)

Deney	Parametre	Ham	Normalize	Bitişik
Orijinal deney	663703	88,86	96,12	97,37
Öğrenme hızı	663783	88,63	95,63	96,80
LSTM boyutu	194919	88,01	95,26	96,43
Aktivasyon fonk.	663783	88,26	95,33	96,51
Filtre boyutu	664039	88,44	95,64	96,89

Tablo 24. Hiper parametre kestirim deneyi II: Katar tanıma
(Hyperparameter estimation experiment II: Ligature recognition)

Deney	Parametre	Ham	Normalize	Bitişik
Orijinal deney	663703	80,48	91,60	92,14
Öğrenme hızı	663783	80,56	91,53	91,39
LSTM boyutu	194919	78,73	89,86	89,62
Aktivasyon fonk.	663783	79,08	90,03	89,56
Filtre boyutu	664039	79,77	90,77	91,15

Tablo 25. Hiper parametre kestirim deneyi III: Kelime tanıma
(Hyperparameter estimation experiment III: Word recognition)

Deney	Parametre	Ham	Normalize
Orijinal deney	663703	44,08	66,45
Filtre boyu	664039	43,01	65,53
Aktivasyon fonk.	663783	42,04	63,40
LSTM boyutu	194919	40,64	63,05
Öğrenme hızı	663783	42,33	64,47

*Hiper parametre kestirim Örnek Tablolar

Doğruluk Oranı ve Hata Dağılımları

Doğruluk Oranı (Accuracy):
Doğruluk oranı, modelin ne kadar doğru tahminlerde bulunduğunu ölçen bir performans metrikidir. Genellikle sınıflandırma problemlerinde kullanılır ve şu şekilde hesaplanır.

Hata Dağılımları:
Hata dağılımları, modelin tahminlerinde yaptığı hataların nasıl dağıldığını inceleyen bir kavramdır. Bu, modelin hangi durumlarda daha fazla hata yaptığını anlamamıza yardımcı olur.

Tablo 20. Katar tanıma doğruluk oranı ve hata dağılımları (Ligatura recognition accuracy and error distributions) (%)

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 21. Kelime tanıma doğruluk oranı ve hata dağılımları (Word recognition accuracy and error distributions) (%)

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

*Örnek tablolar

Sonuç ve Gelecek Çalışmalar

Bu çalışma, Osmanlıca matbu nesih hattının OCR işlemlerinde derin öğrenme tabanlı modellerin üstünlüğünü göstermektedir. Geliştirilen model, Osmanlıca karakterleri daha yüksek doğrulukla tanıyabilmekte ve dijital arşivleme süreçlerine büyük katkı sağlamaktadır.

Gelecekte yapılabilecek çalışmalar şunlardır:

- Modelin el yazısı Osmanlıca belgeleri için de optimize edilmesi.
- Gerçek zamanlı OCR sistemleri için hız optimizasyonlarının yapılması.
- Farklı yazı stillerine uyarlanarak daha geniş çapta bir model oluşturulması.

Bu tür iyileştirmelerle Osmanlıca belgelerin dijitalleştirilmesi daha hızlı ve hatasız hale getirilebilir.