Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, 34, S. 321-327, Mart 2022

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 34, pp. 321-327, March 2022

Research Article

Yeni Bir Türkçe Sesli Kitap Veri Seti Üzerinde Convolutional RNN+CTC, LSTM+CTC ve GRU+CTC Modellerinin Karşılaştırılması

Halil İbrahim Yalman [1*], Zekeriya Tüfekci [2]

[1*] Çukurova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Adana, Türkiye, (ORCID: 0000-0003-0841-1309), halilyalman@hotmail.com.tr
[2] Çukurova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Adana, Türkiye, (ORCID: 0000-0001-7835-2741), ztufekci@cu.edu.tr

(2nd International Conference on Applied Engineering and Natural Sciences ICAENS 2022, March 10-13, 2022)

(**DOI:** 10.31590/ejosat.1082109)

ATIF/REFERENCE: Yalman, H. İ. & Tüfekci, Z. (2022). Yeni Bir Türkçe Sesli Kitap Veri Seti Üzerinde Convolutional RNN+CTC, LSTM+CTC ve GRU+CTC Modellerinin Karşılaştırılması. *Avrupa Bilim ve Teknoloji Dergisi*, (34), 321-327.

Öz

Konuşma tanıma insanların çıkardığı ses dalgalarının yazıya dönüştürülmesi işlemidir. Geçmişten günümüze birçok konuşma tanıma modeli ve veri seti üretilmekle beraber ülkemizde bu konuda bir eksiklik olduğu yadsınamaz bir gerçektir. Bu çalışmada, Türkçe konuşma tanıma sistemleri için sesli kitaplardan oluşan özgün bir veri seti geliştirilmiştir. Bu veri seti halihazırda oluşturulmuş olan sesli kitapların bölümlenmesi yoluyla hazırlanmıştır. Bu veri seti üzerinde Evrişimli Sinir Ağları (CNN) ve Bağlantıcı Zamansal Sınıflandırma (CTC) ile birlikte Yinelemeli Sinir Ağı (RNN), Uzun Kısa Süreli Hafıza (LSTM), Geçitli Tekrarlayan Birimler (GRU) modellerinin performansı incelenmiş ve karşılaştırması yapılmıştır. Bu çalışmanın sonuçlarına göre performansı en yüksek olan model LSTM olması ile birlikte daha az parametre kullanan GRU modelinin konuşma tanıma oranı LSTM modelinin konuşma tanıma oranına yakın çıkmıştır.

Anahtar Kelimeler: Konuşma Tanıma, Derin Öğrenme, Evrişimli Sinir Ağları, Uzun Kısa Süreli Bellek, Basit Tekrarlayan Ağlar, Kapılı Tekrarlayan Hücreler, Bağlantıcı Zamansal Sınıflandırma, Türkçe Sesli Kitap Veriseti.

Comparison of Convolutional RNN+CTC, LSTM+CTC and GRU+CTC Models on A New Turkish Audiobook Dataset

Abstract

Speech recognition is the process of converting sound waves produced by humans into text. Although many Speech recognition models and data sets have been produced from the past to the present, it is an undeniable fact that there is a deficiency in this regard in our country. In this study, a unique data set consisting of audio books was developed for Turkish speech recognition systems. This dataset has been prepared by partitioning the audiobooks that have already been prepared. On this dataset, Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Gated Recurrent Units (GRU) models were examined and compared together with Convolutional Neural Networks (CNN) and Connectionist Temporal Classification (CTC). According to the results of this study, although the model with the highest performance is LSTM, the speech recognition rate of the GRU model, which uses fewer parameters, is close to the speech recognition rate of the LSTM model.

Keywords: Speech Recognition, Deep Learning, Convolutional Neural Networks, Long Short Term Memory, Simple Recurrent Networks, Gated Recurrent Units, Connectionist temporal classification, Turkish Audiobook Dataset.

[*] Sorumlu Yazar: halilyalman@hotmail.com.tr

1. Giriş

İnsanların iletişiminde konuşmanın yeri son derecede önemlidir. Günlük hayatımızda kullandığımız makineler komutlarla çalıştığından dolayı konuşmalarımızı metne dönüştürme ihtiyacı duyulmaktadır. Bunun sebebi teknolojinin her adımda hayatımızı daha kolaylaştırmayı hedeflemesinden dolayıdır.

Kullandığımız elektronik sistemlerde sesle çalışmak tuşlu sistemlerle çalışmaktan daha kolaydır. Bununla birlikte farklı sebeplerden dolayı sesle komut verilebilen sistemlere ihtiyaç doğmaktadır. Bu ihtiyaç konuşma tanıma sistemlerinin gelişmesine sebep olmuştur.

Türkçede yapılan konuşma tanıma çalışmaları veri seti azlığı ve var olan veri setlerinin boyutlarının yeterli olmamasından dolayı bu çalışmayı yapma ihtiyacı hissedilmektedir. Bir diğer amaç birçok sistemdeki Türkçe dil desteği eksikliği ciddi vaziyette hissedilmektedir.

Hayatımızın birçok alanında konuşma tanıma sistemlerine ihtiyaç duymaktayız. Örneğin bir telefona daha hızlı komut verirken, otonom bir aracı kullanırken, akıllı saatler ve bilekliklere komut verirken kullanmaktayız ve önümüzdeki yıllarda daha da çok alanda kullanacağımız tahmin edilmektedir. Çünkü teknolojik gelişmeler bu noktaya doğru evrilmektedir.

Dünyada kullanılan birçok dil için konuşma tanıma amaçlı birçok çalışma olmasına rağmen Türkçe için yeterince çalışma bulunmamaktadır. Bunun bir nedeni de Türkçe için veri seti eksikliğidir. Gelecek yıllarda kullanacağımız birçok makinenin Türkçe dil desteği sunmasını istiyorsak, daha büyük ve daha iyi Türkçe veri setlerine ihtiyacımız olacaktır.

Konuşma tanıma konusundaki ilk çalışmalar 1950'lerde Bell Lab da yapıldığı bilinmekle birlikte 1961 yılında IBM tarafından geliştirilen "Shoebox" bu alanda önemli bir yere sahiptir. Bu alandaki gelişmeler belli bir süre durağan seyretmesinin ardından bilgisayar teknolojilerindeki gelişmelerle birlikte bu alandaki çalışmalar hızlanmıştır. Ayrıca bilgisayar sistemlerindeki paralel ve güçlü işlem yapabilme kapasitesi bu alanın önünü açmıştır.

Hidden Markov Modelin (HMM) Konuşma tanıma performansının diğer yöntemlere göre daha yüksek olmasından dolayı 2010 lu yıllara kadar araştırmacılar tarafından tercih edilen yöntem oldu. Renals ve diğerleri (1994) neural networkün (multilayer perceptron) HMM ile birlikte kullanıldığında konuşma tanıma performansını artırabileceğini gösterdi. Dahl ve diğerleri (2012) 2012 yılında Deep Neural Networkün (DNN) HMM ile birlikte kullanıldığında HMM'e göre daha iyi sonuç verdiğini göstermesiyle birlikte konuşma tanıma için DNN kullanımı yaygınlaştı.

Convolutional Neural Networks (CNN) ile konuşma tanıma çalışmalarında, Abdel-Hamid ve diğerleri (2014) tabanlı yapılan çalışmaların DNN tabanlı yapılan çalışmalardan daha başarılı sonuçlar elde ettiğini gösteren CNN modelini önermiştir. Model bu başarıyı daha az parametre kullanarak sağlamıştır.

Recurrent Neural Networks (RNN) ile konuşma tanıma çalışmalarında, Graves ve diğerleri (2013) Long Short Term Memory (LSTM) sistemlerinin Connectionist Temporal Classification (CTC) ile birlikte kullanımının HMM-DNN hibrid modeline göre daha iyi sonuç verdiğini gösterdi.

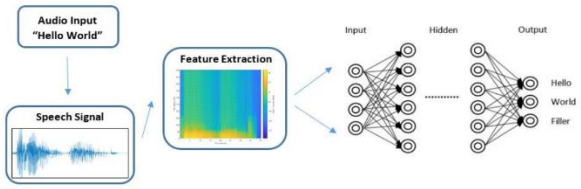
RNN'in bir alt dalı da LSTM benzeri Gated Recurrent Unittir (GRU). Ravanelli ve diğerleri (2018) göre çalışmasında GRU modelinin LSTM'den daha başarılı çalıştığını göstermiştir.

Bu çalışmada konuşma tanıma için yeni bir Türkçe veri seti oluşturmaya ek olarak RNN, LSTM ve GRU gibi son zamanlarda önerilen ve konuşma tanıma oranları yüksek olan akustik modellerin, oluşturulan Türkçe veri seti üzerindeki performansları karşılaştırıldı.

İkinci bölümde bu çalışmada kullanılan RNN, LSTM, GRU, CNN, CTC ve oluşturulan veri seti kısaca açıklandı. Üçüncü bölümde deneysel kurulum anlatılıp deneysel sonuçlar tartışılmıştır. Dördüncü bölümde ise bu çalışmada elde edilen sonuçlar açıklandı.

2. Materyal ve Metot

Konuşma tanıma sistemleri kullanılırken ses verilerine ve o ses verilerine ait metin verilerine ihtiyaç duymaktayız. Öznitelik çıkarımı (Feature Extraction) ile sesten ürettiğimiz matrisleri Sinir Ağları aracılığı ile metin karşılığını öğrenen bir sistem tasarlıyoruz. Bu sistemi başka ses dosyaları ile test ediyoruz ve çıkan sonuçta başarım oranlarını hesaplıyoruz. Bu yaptığımız işlem Şekil 1'de gösterilmiştir.



Şekil 1. Temel konuşma tanıma sisteminin blok diagramı.

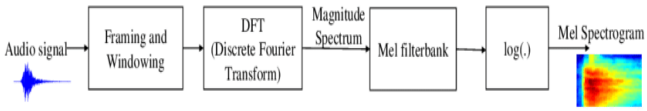
2.1. Konuşma Tanıma için Türkçe Veri Seti

Sesli kitap veri seti 21 kadın ve 21 erkek olmak üzere toplam 42 kişinin konuşmalarından oluşmaktadır. Veri seti toplam 15000 cümleden oluşan 19 saat 4 dakika 14 saniyelik ses verisi içermektedir. Veri setinde bir kadın ve bir erkekten bin beş yüzer diğer kişilerden ise üç yüzer cümle bulunmaktadır. Her bir cümlenin ses verisi bir ses dosyasına kaydedilmiştir. Her bir ses dosyası için bir metin dosyası (label) oluşturulmuştur. Oluşturulan ses dosyalarının uzunluğu 10 saniye ile sınırlandırılmıştır.

Tüm ses verisi kullanılarak oluşturulan veri seti Veri-Seti-3 olarak adlandırılmıştır ve yaklaşık 20 saat ses verisi içermektedir. Ayrıca tüm ses verisinin yarısı kullanılarak oluşturulan 7500 ses dosyasından oluşan 9 saat 21 dakika 41 saniyelik Veri-Seti-2 olarak adlandırılan veri seti oluşturulmuştur. Benzer şekilde tüm ses verisinin dörtte biri kullanılarak oluşturulan 3750 ses dosyasından oluşan 4 saat 50 dakika 51 saniyelik Veri-Seti-1 olarak adlandırılan veri seti oluşturulmuştur. Burada farklı büyüklükte veri seti oluşturmanın amacı eğitim verisinin büyüklüğünün konuşma tanıma performansına etkisini incelemektir. Deneylerimiz bu üç veri seti kullanılarak yapılmıştır.

2.2. Öznitelik Çıkarımı

Konuşma tanımada son yıllarda ses verisinin öznitelik çıkarımı yapılmadan kullanıldığı gibi öznitelik çıkarımı yapılan birçok teknik kullanılmaktadır. Başlıca ana teknikler Linear Predictive Analysis (LPC), Perceptual Linear Predictive (PLP), Relative Spectra Filtering of Log Domain Coefficients (RASTA), Mel-Frequency Cepstral Coefficients (MFCC) ve Mel Spectrogram (Benba vd.2015; Tak vd. 2017; Tanveer vd. 2021). Bu tez çalışmasında Mel Spectrogram kullanılacaktır. Mel Spectrogram özellik çıkarımının block diyagramı Şekil 2. de sunulmuştur.



Şekil 2. Mel Spectrogram öznitelik çıkarımının blok diyagramı (Tak vd., 2017)

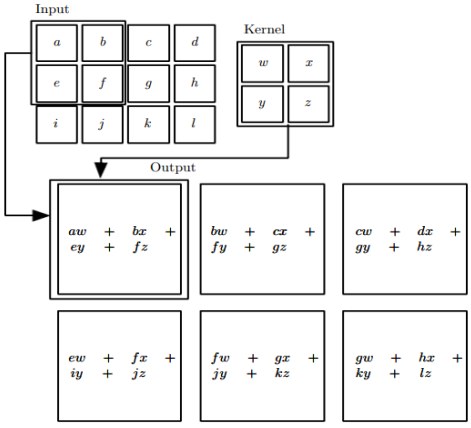
2.3. Sinir Ağları

Sinir Ağları veriler arasında ilişkisel bağ kurarak gelecek verilerdeki ilişkileri tahmin etmek için kullanılan birden fazla katmana sahip bir makine öğrenmesi alt dalıdır. Konuşma tanımada nihai hedef ses verilerini ve ait olan metin verileri ile eğitilen modele başka ses verilerini verildiğinde ona ait metin verilerini hatasız elde etmektir.

Sinir Ağları alt dalı olan 3 model üzerinde çalışmalar yapılacaktır. Bu modeller aşağıda açıklanmıştır.

2.3.1. Convolution Neural Network

Evrişimsel sinir ağları veya CNN olarak da bilinen Evrişimsel ağlar, ızgara benzeri bilinen bir topolojiye sahip veriyi işlemek için kullanılan özel bir sinir ağıdır (Goodfellow vd., 2018).



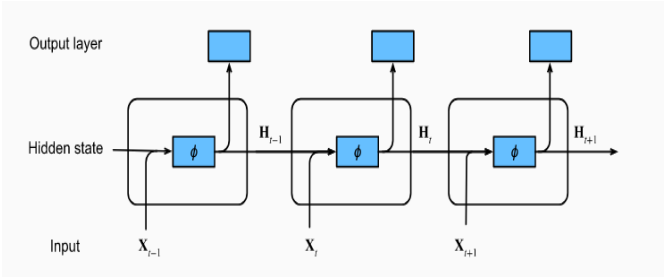
Şekil 3. 2 boyutlu bir CNN örneği (Goodfellow vd., 2018)

Evrişimli sinir ağında kernel giriş verisindeki birimle çarpılır. Kernel her adımda bir atlama sayısı(stride) kadar yana kaydırılarak çarpım işlemi yapılır. Her CNN işlemi sonucunda matris boyutu küçülür. İstenirse matris boyutunun kenarlarına veri eklenerek doldurma(padding) işlemi yapılır.

2.3.2. Recurrent Neural Network

Yinelemeli Sinir Ağları (Recurrent Neural Networks, RNN'ler) sıralı verileri işlemek için özelleştirilmiş sinir ağları ailesidir (Rumelhart vd., 1986). Mevcut durumla birlikte bir önceki durumun bilgilerinin sisteme dahil edilmesi mantığına dayanır.

2.3.2.1. Simple Recurrent Network



Şekil 4. SRN Yapısı

Basit Yinelemeli Ağ (Simple Recurrent Network) RNN modelinin alt türüdür. SRN Modeli Şekil 4'te görüldüğü üzere mevcut durumdaki giriş(X_t) durumunun gizli katman verisi(H_t) bir sonraki andaki duruma girmektedir. Modelin hesaplamalarında aşağıdaki denklemler kullanılır.

$$h_t = \tanh(W_{xh} x_t + W_{hh} h_{t-1} + b_h) \quad (1)$$

$$y_t = \tanh(W_{hy} h_t + b_y) \quad (2)$$

Denklemlerde şu andaki durumdaki giriş x_t, bir önceki andan gelen gizli katman bilgisi h_{t-1} ve bu düğümde kullanılıp diğer düğüme de gönderilecek gizli katman bilgisi ht olarak gösterilmiştir. W_{xh}, W_{hh} ve W_{hy} ağırlık matrisidir. b_h ve b_y bias vektörleridir. Aktivasyon fonksiyonu tanh Hiperbolik Tanjant Fonksiyondur.

SRN modeli RNN modeli olarak bilindiğinden dolayı bu çalışmada RNN isimlendirmesi kullanılacaktır.

2.3.2.2. Long Short Term Memory

Uzun kısa süreli bellek (Long Short Term Memory, LSTM) RNN mimarisinin bir alt türüdür. LSTM'in blok diagramı Şekil 5'te verildiği gibidir. LSTM'in çıkış ve gizli katman parametreleri aşağıdaki formüller kullanılarak hesaplanır.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (4)$$

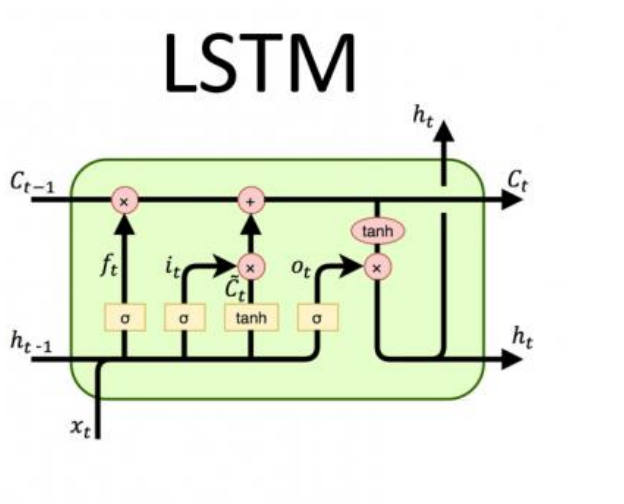
$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (6)$$

$$h_t = o_t tanh(c_t) \quad (7)$$

Giriş kapısı(i), unutma kapısı(f), çıkış kapısı(o) ve hücre durumundan(c) oluşan yapının denklemi Denklem 3 ve 7 arasında belirtilmiştir (Öztürk ve Özkaya, 2021). t ifadesi o anki durumu t-1 ifadesi bir önceki durumu belirtmektedir. Denklem 3'te W_{xi}, W_{hi}, W_{ci} ağırlık matrisini ve bi bias vektörünü ifade eder. Denklem 4'te W_{xf}, W_{hf}, W_{cf} ağırlık matrisini ve b_f bias vektörünü

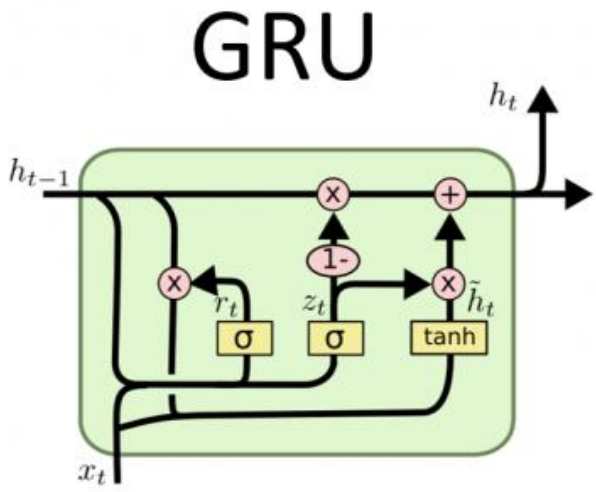
ifade eder. Denklem 5'te W_{xc}, W_{hc} ağırlık matrisini ve b_c bias vektörünü ifade eder. Denklem 6'te W_{xo}, W_{ho}, W_{co} ağırlık matrisini ve b_o bias vektörünü ifade eder. Denklemlerde σ sigmoid fonksiyonudur. Denklem 7'de tanh Hiperbolik Tanjant Fonksiyondur.



Şekil 5. LSTM Yapısı

2.3.2.3. Gated Recurrent Units

Geçitli Tekrarlayan Birimler (Gated Recurrent Units, GRU) LSTM sisteminin sadeleştirilmiş hali olan bir RNN alt türüdür.



Şekil 6. GRU Yapısı

GRU yapısı incelendiğinde yöntemin LSTM'den farkı, tek bir geçit biriminin aynı anda hem unutma faktörünü hem de durum biriminin güncelleme kararını kontrol etmesidir (Goodfellow vd., 2018). Şekil 6'da gösterildiği üzere GRU yapısı Unutma ve durum birimi olarak kullanılan güncelleme birimi(z_t) ve silme biriminden(r_t) oluşmaktadır.

$$r_t = \sigma\left(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}\right) \quad (8)$$

$$z_t = \sigma\left(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}\right) \quad (9)$$

$$\tilde{h}_t = \tanh\left(W_{in}x_t + b_{in} + r_t * \left(W_{hn}h_{(t-1)} + b_{hn}\right)\right) \quad (10)$$

$$h_t = \left(1 - z_t\right) * \tilde{h}_t + z_t * h_{(t-1)} \quad (11)$$

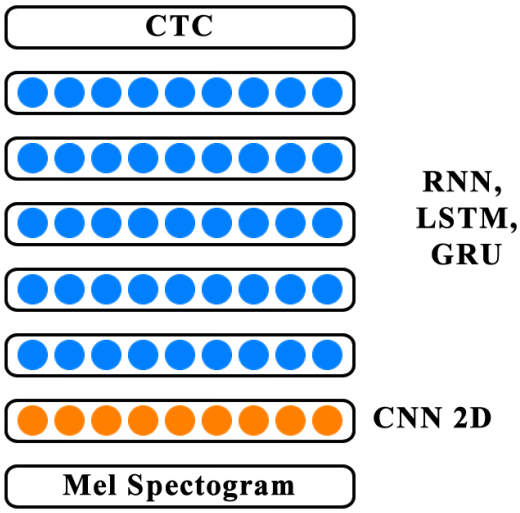
GRU mimarinin denklemleri Denklem (8) ile (11) arasında verilmiştir (Cho vd., 2014). Denklemlerde görüldüğü üzere o andaki zamanı t, bir önceki andaki zamanı (t-1), gizli katman ht, o andaki girişi xt ifade etmektedir. W_{ir}, W_{hr}, W_{iz}, W_{hz}, W_{in}, W_{hn}, ağırlık matrisini ifade etmektedir. b_{ir}, b_{hr}, b_{iz}, b_{hz}, b_{in}, b_{hn} bias vektörleridir. σ sigmoid fonksiyonu, tanh Hiperbolik Tanjant Fonksiyondur.

2.4. Connectionist Temporal Classification

Bağlantıcı Zamansal Sınıflandırma (Connectionist Temporal Classification, CTC) Yenilemeli Sinir Ağlarında sınıflandırma için kullanılan bir katmandır. Akustik modelin çıktısı CTC'ye girdi olarak verilmektedir. CTC'nin çıktısı ise herbir girdi için Türkçe alfabedeki 29 harf, boşluk, " **x** ", " **w** ", " **q** " ," ' " , " . " olmak üzere 35 karakterden biri olabilir. Dolayısıyla CTC girişindeki akustik modelin oluşturduğu diziyi karakter (35 karakter) dizisine dönüştüren bir sınıflandırma yöntemidir.

3. Deneysel Kurulum Ve Sonuçlar

Bu çalışmada kullanılan konuşma tanıma sistemi Şekil-7'de gösterildiği gibidir. Şekil 7'den de anlaşılabileceği gibi bu çalışmada akustik model olarak kullanılan RNN, LSTM ve GRU'nun performansları karşılaştırılmıştır. Bu üç akustik modelin karşılaştırılması için tüm konuşma modellerinde aynı CNN ve CTC yapısı kullanılmıştır.



Şekil 7. Konuşma Tanıma Modeli

3.1. Model Hiperparametreleri

Kullandığımız Modeller Python programlama dilinde PyTorch kütüphanesi kullanılarak oluşturulmuştur. Kullandığımız modellerde Mel Spectrogram öznitelikleri Şekil 2'de gösterildiği gibi PyTorch kütüphanesi kullanılarak elde edilmiştir. Şekil 7'de görüldüğü gibi Mel Spectrogram öznitelikleri tek katmanlı iki boyutlu CNN' girdi olarak verilmekte, CNN'in çıkışı ise beş katmanlı RNN, LSTM veya GRU modellerinden birine girmektedir. Modelin en sonunda ise sınıflandırıcı olarak CTC katmanı kullanılmaktadır.

Kullandığımız modellerde birçok parametreler kullanmaktadır. Bu hiperparametrelerin değerleri modellerin performansını değiştirebilmektedir. Modellerde kullanılan hiperparametreler aşağıda açıklanmıştır:

- **Öznitelik Çıkarımı:** Konuşma sinyali PyTorch kütüphanesi kullanarak her 10 milisaniyede bir 20 milisaniyelik (%50 örtüşme ile) hanning window kullanarak çerçevelere ayrılmış ve her bir çerçeveden 128 adet Mel Spectrogram özniteliği çıkarılmıştır.
- **Katman Sayısı:** Tek 2D-CNN katmanı kullanılmıştır. 2D-CNN için 32 filtre kullanılmıştır. Her bir filtrenin kernel boyutu (3,3) olup atlama(stride) boyutu (2,2)'dir. RNN, LSTM ve GRU modellerinde ise beşer katman kullanılmıştır.
- **Düğüm Sayısı (Node)**: RNN, LSTM ve GRU modellerinde her katmanda 512 düğüm kullanılmıştır.
- **Eğitim Döngüsü (Epoch):** Her çalışma 300 döngü çalıştırılmıştır.
- **Öğrenme oranı**: Deneylerimizde 0.0005 öğrenme adımı kullanmıştır.
- **Eğitim ve Test Oranı:** Her bir veri setin %97'si eğitim ve %3'ü test için ayrılmıştır.
- **Grup Boyutu (Batch Size):** Her denemede grup boyutu 4 olarak alınmıştır.
- **Aktivasyon Fonksiyonu:** CNN aktivasyon fonksiyonu olarak Gelu kullanılmıştır.
- Her katmanın ardından normalizasyon işlemi ve hattan düşme (Dropout) işlemi uygulanmıştır.
- **Hata oranı (LER):** CTC çıktısı olan karakter dizisinin hata oranının bulmak için kullanıyoruz. Hedef ve tahmin arasındaki sonuçlar arasındaki Edit Distance oranı ile bulunan sonuçlardır. LER değeri aşağıda verilen formül ile hesaplanmaktadır.

$$LER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \quad (12)$$

Denklem 12'de LER iki ifade arasında dönüşüm yapmak için eklenen harf sayısı(I), silinen harf sayısı(D), değiştirilen harf sayısı(S) ve doğru olarak bilinen harf sayısı(C) ile iki dizi arasındaki benzerlik ölçümüdür. N ise iki dizi arasındaki referans ifadenin boyutunu gösterir.

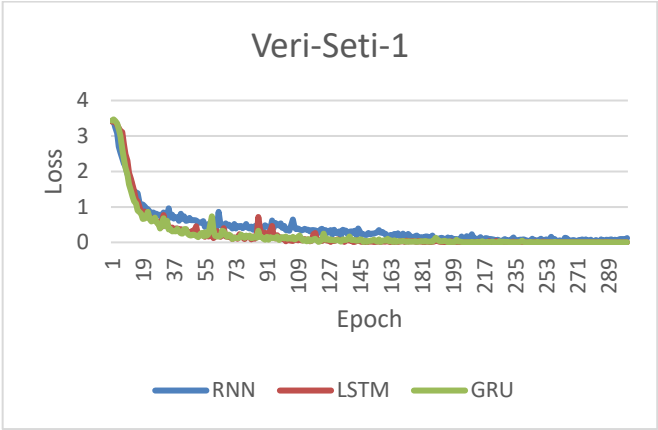
Karşılaştırması yapılan Sinir Ağları modellerinden RNN, LSTM ve GRU modellerinin katman sayısı ve düğüm sayılarının aynı olmasına rağmen her bir modelin iç yapısının farklı olmasından dolayı her bir modelin toplam parametre sayısı birbirinden farklıdır. Akustik modellerin toplam parametre sayıları Tablo 1'de gösterilmiştir. Tablo 1'den görüldüğü gibi RNN modelinin toplam parametre sayısı en azdır. En fazla toplam parametre sayısına sahip model ise LSTM'dir.

Tablo 1. Toplam parametre sayıları

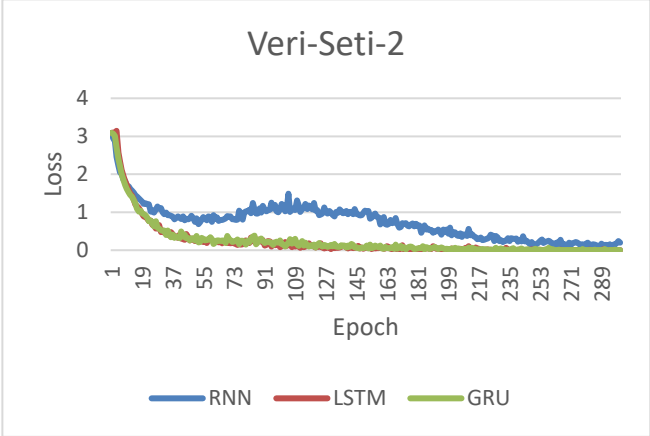
RNN	1197923
LSTM	4547939
GRU	3431267

3.2. Deneysel Sonuçlar ve Tartışma

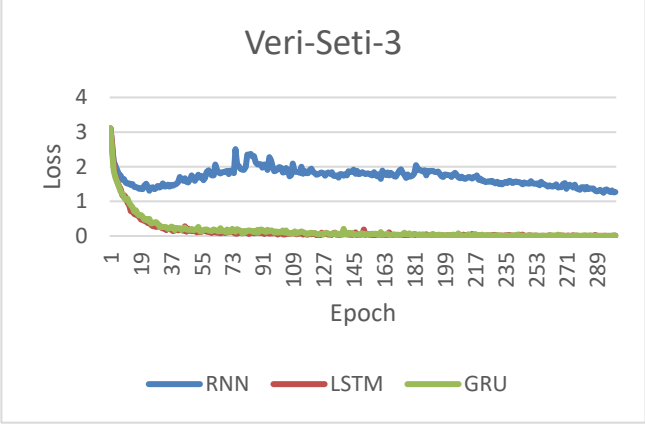
Şekil 8,9 ve 10 sırasıyla Veri-Seti-1, Veri-Seti-2, ve Veri-Seti-3 için loss'un epoch değerine göre değişimini göstermektedir. Üç şekilden de görüldüğü gibi LSTM ve GRU akustik modeli için epoch değeri arttıkça lossunda düzenli olarak düştüğü görülmektedir. RNN için ise loss değerinin aynı kararlılıkla düşmediği görülmektedir.



Şekil 8. Veri-Seti-1 Kayıp Fonksiyonu



Şekil 9. Veri-Seti-2 Kayıp Fonksiyonu



Şekil 10. Veri-Seti-3 Kayıp Fonksiyonu

Tablo 2, Tablo 3 ve Tablo 4 sırasıyla Veri-Seti-1, Veri-Seti-2, ve Veri-Seti-3 için tüm modellerin LER, eğitim süresi ve toplam loss değerlerini göstermektedir. İlgili tablolardan görüleceği üzere her bir veri setinde RNN en yüksek LER değerini ve LSTM'in en düşük LER değerini verdiği görülmektedir. Buradan LSTM'in tartışmasız olarak tüm veri setleri için en iyi LER değeri verdiği sonucuna ulaşabiliriz. Bununla beraber Tablolardan GRU'nun

LER değerlerinin her bir veri seti için LSTM'in LER değerlerine çok yakın olduğu görülmektedir.

Modellerin eğitim sürelerini karşılaştırdığımızda Tablo 2, Tablo 3 ve Tablo 4'den görüleceği üzere her bir veri seti için RNN'in en kısa eğitim süresine sahip olduğu, LSTM'in ise en uzun eğitim süresine sahip olduğu görülebilir.

RNN modelinin LER değerlerinin diğer modellere göre çok yüksek olmasından dolayı RNN modelinin konuşma tanıma için kullanılmasının uygun olmadığı sonucuna varabiliriz.

LSTM ve GRU akustik modellerinin LER değerlerini karşılaştırdığımızda Tablo 2, Tablo 3 ve Tablo 4'den görüleceği üzere her bir veri seti için LSTM'in yaklaşık %3 civarında GRU'ya göre daha düşük LER değerine sahip olduğu görülmektedir. Dolayısıyla bizim için LER çok önemliyse akustik model olarak LSTM kullanmanın daha avantajlı olduğu görülmektedir.

Eğitim süresi ve parametre sayısını göz önüne aldığımızda Tablo 2, Tablo 3 ve Tablo 4'den görüleceği üzere her bir veri seti için LSTM modelinin eğitim süresinin GRU modeline göre yaklaşık olarak %30 daha fazla olduğu görülmektedir. Ayrıca Tablo 1'den görüleceği üzere LSTM'in parametre sayısı GRU'nun parametre sayısından yaklaşık olarak %32 fazladır. Eğitim süresi ve parametre sayılarını da dikkate aldığımızda akustik model olarak GRU kullanmak LSTM'e göre daha avantajlı olabilir. LSTM'in daha iyi LER sonucu vermesine rağmen GRU daha hızlı eğitilebilmekte ve LSTM'e göre daha az parametreye sahiptir. Bundan dolayı GRU LSTM'e göre %3 daha kötü sonuç vermesine rağmen akustik model olarak LSTM yerine GRU tercih edilebilir.

Tablo 2, Tablo 3 ve Tablo 4'den görülen diğer bir sonuç ise veri seti büyüklüğü artıkça tüm modellerin LER değerlerinin düşmesidir.

Tablo 2. Veri-Seti-1 Model Karşılaştırması

Veri-Seti-1(Yaklaşık 5 Saat)

MODELS	LER	EĞİTİM SÜRESİ(dk)	TOPLAM LOSS
CNN-RNN	0.4176	**237**	119.2079
CNN-LSTM	**0.3284**	503	71.9418
CNN-GRU	0.3398	367	**68.0445**

Tablo 3. Veri-Seti-2 Model Karşılaştırması

Veri-Seti-2(Yaklaşık 10 Saat)

MODELS	LER	EĞİTİM SÜRESİ(dk)	LOSS
CNN-RNN	0.3592	**479**	222.9458
CNN-LSTM	**0.2825**	1024	**70.0768**
CNN-GRU	0.2920	734	73.8935

Tablo 4. Veri-Seti-3 Model Karşılaştırması

Veri-Seti-3(Yaklaşık 20 Saat)

MODELS	LER	EĞİTİM SÜRESİ(dk)	LOSS
CNN-RNN	0.4561	1993	509.7223
CNN-LSTM	**0.2329**	2022	**41.2186**
CNN-GRU	0.2485	**1476**	47.4938

4. Sonuç

Bu çalışmada veri setinin büyüklüğünün konuşma tanıma oranına etkisini incelemek amacıyla 5, 10 ve 20 saatlik olmak üzere üç adet veri seti hazırlanmıştır. Deneysel sonuçlar incelendiğinde her bir akustik model için kullanılan veri setinin büyüklüğü artığında ilgili akustik model için LER değerinin düştüğü görülmüştür. Bu sonuç bize eğitim için büyük veri seti kullanmanın önemini göstermektedir.

Veri setleri üzerinde yapılan deneylerde LSTM modelinin en yüksek konuşma tanıma oranını verdiği, RNN'in ise diğer iki modele göre çok düşük tanıma oranı verdiği gösterilmiştir. Dolayısıyla akustik model olarak RNN kullanmanın herhangi bir avantajı bulunmamaktadır. LSTM'in parametre sayısının GRU'nun parametre sayısından yaklaşık olarak %32 fazla olmasına rağmen GRU ile LSTM'in konuşma oranları arasında fazla fark bulunmadığı gösterilmiştir. Ayrıca GRU modelinin eğitim süresinin LSTM'e göre daha kısa olduğu gösterilmiştir. Dolayısıyla akustik model olarak LSTM yerine GRU kullanmak parametre sayısının az olmasından ve eğitim süresinin daha kısa olmasından dolayı daha avantajlı olabilir.

Gelecekteki çalışmalarda daha fazla parametreli ve katmanlı modellerin denemesi planlanmaktadır.

Kaynakça

Abdel-Hamid O., Mohamed A., Jiang H., Deng L., Penn G. and Yu D., (2014) "Convolutional neural networks for speech recognition" IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533-1545, doi: 10.1109/TASLP.2014.2339736.

Benba A., Jilbab A. and Hammouch A., (2015) "Detecting patients with parkinson's disease using mel frequency cepstral coefficients and support vector machines", International Journal on Electrical Engineering and Informatics- Volume 7, Number 2.

Cho K., Van Merriënboer B., Bahdanau D., Bengio Y., (2014) "On the properties of neural machine translation: encoder-decoder approaches." arXiv preprint arXiv:1409.1259.

Dahl G. E., Yu D., Deng L., A. Acero (2012) "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." Ieee Transactıons On Audıo, Speech, And Language Processıng, Vol. 20, No. 1.

Goodfellow I., Bengio Y. and Courville A., (2018) *Derin Öğrenme*, Ankara: Buzdağı Yayınevi

Graves A., Mohamed A., Hinton G. (2013) "Speech recognition with deep recurrent neural networks." ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing- Proceedings. 38. 10.1109/ICASSP.2013.6638947.

Graves A., Fernández S., Gomez F., and Schmidhuber J., (2006) "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." In Proceedings of the 23rd international conference on Machine learning (pp. 369-376).

Öztürk, Ş., Özkaya, U. (2021) "Residual LSTM layered CNN for classification of gastrointestinal tract diseases". Journal of Biomedical Informatics, 113, 103638.

Ravanelli M., Brakel P., Omologo M. and Bengio Y., (2018) "Light gated recurrent units for speech recognition*." in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 2, no. 2, pp.* 92-102, doi: 10.1109/TETCI.2017.2762739.

Renals S. and Bourlard H. (1994) "Connectionist probability estimators in hmm speech recognition." Ieee Transactions On Speech And Audio Processing, Vol.2, No. 1, Part 11.

Rumelhart D., Hinton G. and Williams, R. (1986) "Learning representations by back-propagating errors." Nature 323, 533–536

Tanveer M. H., Zhu H., Ahmed W., Thomas A., Imran B. M. and Salman M., (2021) "Mel-spectrogram and deep cnn based representation learning from bio-sonar ımplementation on uavs", 2021 International Conference on Computer, Control and Robotics.

Tak R. N., Agrawal D. M., and Patil H. A., (2017) "Novel phase encoded mel filterbank energies for environmental soundclassification." In International Conference on Pattern Recognition and Machine Intelligence, pages 317–325.Springer.