SETSCI conference proceedings

SETSCI Conference Proceedings
13 , 53-58, 2022

5th International Symposium on Innovative Approaches in Smart Technologies

28-29 May 2022, Turkey

ISAS 2022

https://doi.org/10.36287/setsci.5.1.011
eISSN: 2687-5527/

Videolardaki Çevresel Sesleri Tanımak İçin Derin Öğrenme Tabanlı Bir Model Geliştirme

Bedirhan KARAKAYA[1*], Emre Beray BOZTEPE[1+] ve Bahadır KARASULU[1]

[1]*Çanakkale Onsekiz Mart Üniversitesi, Bilgisayar Mühendisliği Bölümü, Çanakkale, Türkiye*
[*]*Sorumlu Yazar: bedirhankrkya@gmail.com*
[+]*Sunucu: berayboztepe@outlook.com*

Özet **—** Günümüzde çevremizdeki çeşitli seslerin ayrıştırılarak ortam tanıma işlemi popülerlik kazanmıştır. Videolarda arka plandaki çeşitli sesler, makine öğrenmesi ve derin öğrenme teknikleriyle oldukça yüksek başarıyla sınıflandırılabilmektedir. Böylece anlamsal olarak zenginleştirilmiş video sahneleri betimlenebilmektedir. Bu çalışma, çevresel seslerin tanımlanması için uygun bir derin öğrenme sinir ağı modelinin geliştirilmesi sürecini içermektedir. Geliştirilen modelin eğitilmesinde çeşitli veriler içeren veri kümesinden on tane temel kategori seçilerek modelin tahminleme gücü deneylerde sınanmıştır. Elde bulunan sınırlı veriden öncelikle spektrogramlar elde edilip daha sonra bu spektrogramlar veri arttırma teknikleri kullanılarak zenginleştirilmiştir. Ayrıca, modelin eğitilmesinde üç farklı tasarımsal yaklaşım ile kaynak kodlar yazılmıştır. Bu kodlar Evrişimsel Sinir Ağları, Uzun Kısa Süreli Bellek, Kapılı Tekrarlayan Birim gibi derin öğrenme ağ modeli tabanlı yöntemler kullanılarak oluşturulmuştur. Tasarlanan yedi farklı sinir ağı modeli deneylerde eğitilmiş ve testler ile başarımı kanıtlanmıştır. Oluşturulan modellerin en yüksek başarıma sahip olanı ile yaklaşık %87 oranında doğruluk oranı elde edilmiştir. Elde edilen deneysel sonuçlara ve bilimsel değerlendirmeye çalışmamızda yer verilmektedir.

Anahtar Kelimeler --- Evrişimsel Sinir Ağları, Tekrarlayan Sinir Ağları, Çevresel Ses Tanıma

Development of a Deep Learning Based Model for Recognizing the Environmental Sounds in Videos

Abstract **—** Nowadays, decomposition of various environmental sounds for environment recognition has gained popularity. Various background sounds in videos could be classified with high success with deep learning and machine learning techniques. In this way, semantically enriched video scenes can be depicted. This work contains the process of developing a convenient deep learning neural network model for environmental sounds recognition. In training the developed model, ten main categories have been chosen from a dataset that has various data to test the model's prediction power by experiment. From the limited data available, first, spectrograms have been produced and then, these spectrograms have been enriched by the help of data augmentation techniques. In this way, attribute diversity that was gained from data has been increased. Also, with three different design approaches for training the model, source codes have been written. These codes have been created by using deep learning network model-based methods such as Convolutional Neural Networks, Long Short Term Memory, Gated Recurrent Unit. Seven different designed neural network models have been trained by experiments and achievement has been proved by tests. With the highest accuracy obtained from one of the generated models, approximately %87 of accuracy has been obtained. This work contains obtained experimental results and scientific evaluation.

Keywords --- Convolutional Neural Networks, Recurrent Neural Networks, Environmental Sound Recognition

I. GİRİŞ

Teknolojinin gelişmesiyle beraber ortaya çıkan Ses Tanıma (Sound Recognition), ses sinyalinin analiz edilmesi yöntemine dayanan bir teknolojidir. Bu işlem, başka birçok yöntem gibi Derin Öğrenme (Deep Learning) yöntemleri kullanılarak gerçekleştirilebilmektedir. Bu yöntemlere; spektrogram oluşturarak Evrişimsel Sinir Ağları (Convolutional Neural Network, ESA) oluşturma, Zaman Serisi (Time Series) kullanarak Tekrarlayan Sinir Ağları (Recurrent Neural Network, TSA) kurma gibi yöntemler

örnek olarak verilebilmektedir. Ses tanıma işlemleri, farklı alanlarda kullanılabilmektedir. Çoklu ortam sistemleri için ses sahne olaylarının tespiti [1], [2], [3], [4], dinlenen sesin metne dönüştürülmesi (Speech to Text) bu alanlara örnek verilebilir. Ancak ses tanıma yöntemi için kullanılacak veri kümesi genellikle sınırlı olmaktadır. Bu yüzden, veri arttırma yöntemlerinin, modeli eğitmeden önce çok verimli bir şekilde gerçekleştirilmesi beklenir. Veri kümesinde bulunan veriler arttıkça eğitilen modelden beklenen başarım artacağı için modelin verimliliği de artmaktadır. Video parçaları içerdikleri görsel ve işitsel öğeler nedeniyle, çevresel sesler bakımından oldukça zengin bir ortam olarak çalışmalarda kullanılmaktadır. Çevresel Ses Tanıma (Environmental Sound Recognition) ise günlük yaşantıda, çevreden duyulabilecek her türlü sesi kapsayan geniş bir alandır ve çeşitli ses özniteliklerini barındırmaktadır [5], [6], [7], [8]. Bu seslere, bebek ağlaması, dalga sesi gibi birçok ses örnek verilebilir.

Bu çalışmada, ikinci bölümde literatür taramasına, üçüncü bölümde veri kümesine, dördüncü bölümden on ikinci bölüme kadar çalışmamızda oluşturulan modellere yer verilmiştir. Modellerimizde çevresel seslerin öncelikle spektrogramlara dönüştürülmesi işlemi yapılarak elde bulunan öznitelik kümesiyle oluşturulan verinin arttırılması için farklı tekniklerin kullanılması sayesinde farklı derin öğrenme modelleri oluşturularak elde bulunan verilerle bu modeller eğitilmiştir. On ikinci bölümde deneysel sonuçlara, on üçüncü bölümde sonuçlar ve tartışmalara yer verilmiştir.

II. LİTERATÜR TARAMASI

Derin Öğrenme yönteminin gelişmesi ve Bilgisayarlı Görü (Computer Vision) gibi alanlarda etkili kullanılması ile literatürde bu tarz problemlere çözüm üretmek için birçok çalışma gerçekleştirilmiştir. Derin öğrenme kapsamında çok kullanılan bir mimari olan ESA ile oluşturulan mimari modeller ve bu modellerin sınıflandırma problemlerinde etkili bir sonuç vermeye başlanması ile ses tanıma gibi problemlerin çözümü için ESA sıkça kullanılmaya başlanmıştır.

Ses ile ilgili sınıflandırma problemi çözümüne dair Piczak [2] tarafından önerilen model ile parçalı spektrogramlar üzerinden derin sinir ağı eğitilmiştir. Her ses çerçevesine (incelenen ses penceresine) ait log-mel özellikleri ses özniteliği olarak elde edilmiş ve bu spektrogramlar ESA modeline verilmiştir. ESC-10 veri kümesini kullanılarak bu eğitim sonucunda %81,5 oranında bir doğruluk oranı elde edilmiştir.

Shaobo ve ark. çalışmasında [3] önerilen RawNet modeli ile girdi olarak log-mel özelliğinden ziyade ham dalga biçimi (raw waveform) ile oluşturulmuş spektrogramları kullanarak ESA modeli eğitilmiştir. ESC-10 veri kümesini kullanarak yapılan eğitim sonucunda %85,2 oranında bir doğruluk oranı elde edilmiştir.

Khamparia ve ark. çalışmasında [4] önerilen Khamparia modeli ile log-mel özellikli spektrogramlar elde edilerek ESA ve Tensör Derin Yığınlama Ağları (Tensor Deep Stacking Network, TDYA) kullanılarak iki farklı model oluşturulmuştur. Bu modellerden ESA altyapısı kullanılarak oluşturulan model ile ESC-10 veri kümesinde %77 oranında doğruluk oranı elde edilmiştir. Diğer model olan TDYA kullanılarak oluşturulan model ile ESC-10 veri kümesinde %56 oranında doğruluk oranı elde edilmiştir.

Krizhevsky ve ark. çalışmasında [5] önerilen AlexNet modeli ile log-mel özellikli spektrogramlar elde edilmiştir. Mel ölçeğindeki katsayılardan oluşan öznitelik cinsi olarak MFCC (Mel-frequency cepstral coefficient) [6] tipinde öznitelik elde etme yöntemi ve çapraz tekrarlamaları içeren CRP (Cross Recurrence Plot) [9] yöntemi kullanılarak öznitelik çeşitliliği artırılmıştır. Daha sonra bu üç yöntemden elde edilen spektrogramlar tek bir spektrogram altında tek renk olarak birleştirilmiştir. Bu şekilde elde edilen veri kümesi ile Evrişimli Tekrarlayan Sinir Ağları (Convolutional Recurrent Neural Network, ETSA) kullanılarak model oluşturulmuştur. Bu model, ESC-10 veri kümesi kullanılarak yapılan derin sinir ağı eğitimi sonucunda %86 oranında doğruluk oranına ulaşmıştır.

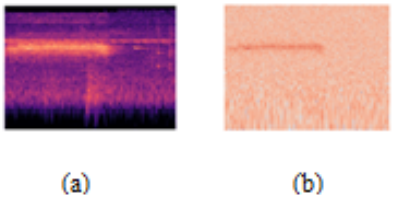
III. VERİ KÜMESİ

Çalışmamızda tasarlanan yapı ve modelleri oluşturmak için çevresel sesleri içeren ve içinde beş adet ana konu kategorisinden oluşan toplamda elli ses sınıfı bulunan veri kümesinden on tane ses kategorisi kullanılmıştır [10]. Bu kategoriler: Hayvanlar (Animals), Doğal ses manzaraları ve su sesleri (Natural soundscapes & water sounds), İnsan, konuşma olmayan sesler (Human, non-speech sounds), İç/Yerel mekân sesleri (Interior/domestic sounds), Dış/Kentsel mekân sesleri (Exterior/urban noises) kategorileridir. Bu kategorilerin her birinde toplam 10'ar tane ses sınıfı bulunmaktadır. Bu kategorilerdeki sesler: "*Chirping birds*", "*Thunderstorm*", "*Breathing*", "*Brushing teeth*", "*Can opening*", "*Clock tick*", "*Chainsaw*", "*Car horn*", "*Church bells*", "*Airplane*" sesleridir. Her bir ses için ise birbirinden farklı kırkar tane ses örneği vardır. Bu ses kayıtlarının uzunlukları yaklaşık beş saniyedir. Veri kümesinin arttırılması için elde bulunan seslere arka plan sesi eklenmiştir. Bu arka plan sesi bir dakikalık beyaz gürültü içeren bir ses parçasıdır. Elde bulunan beş saniyelik seslere eklenmesi için bu arka plan sesinden de 5000 ms'lik bir kesit alınmıştır. Daha sonra bu kesit, seslere entegre edilmiştir. Bu yolla veri kümesindeki veri büyüklüğü iki katına çıkarılmış olmaktadır. Burada elde edilen modifiye edilmiş veri kümesi, modelin temel veri kümesi olarak kullanılmıştır. Ayrıca kurulacak modele göre bu temel veri kümesinin üzerine MFCC tipinde öznitelik elde etme yöntemi, Chroma [7] tipinde öznitelik elde etme yöntemi, Çoklu Maskeleme (Multiple Masking) [11] gibi yöntemler uygulanarak veri kümesinden elde edilecek öznitelik çeşitliliğinin artırılması sağlanmıştır. Genel olarak

çalışmamızdaki veri kümeleri model eğitimi için; %80 eğitim, %10 doğrulama, %10 test olarak ayrılmıştır.

IV. KULLANILAN KÜTÜPHANELER VE GELİŞTİRİLEN ORTAM

Derin öğrenme modellerinin hepsinin oluşturulması için Keras [12] kütüphanesi kullanılmıştır. Ayrıca Keras kütüphanesinin içinde hazır olarak bulunan ve verilen modelin taban modeli olarak kullanılan önceden eğitilmiş olarak gelen ResNet50 [13], MobileNetV2 [14], VGG16 ve VGG19 [15] modelleri kullanılmıştır. Modelin oluşturulması temel olarak, elde bulunan ses kaydının hem Librosa [8] hem de Tensorflow [16] kütüphaneleri yardımıyla spektrogram haline getirilmesine dayanmaktadır. Bu spektrogramlar log-mel özellikleri alınarak oluşturulmuştur. Daha sonra ise bu kütüphanelerin yardımıyla elde bulunan veri kümesi için veri artırılması (data augmentation) programatik olarak sağlanmıştır. Verilen girdi videosu üzerinden seslerin doğrudan ayrıştırılmasını sağlayan SoundNet [17], [18] kütüphanesi de test edilmiştir. Matplotlib [19] kütüphanesinin yardımıyla ise oluşturulan spektrogramlar bir grafik haline getirilmiştir. Ayrıca Numpy [19], Pandas [19], OS [19] vb. kütüphanelerden de yararlanılmıştır. Yapılan eğitimler ve testler sırasında, biri 16 GB RAM bellekli, 2.20 GHz ile çalışan Intel Core i7-8750H işlemci, Nvidia GeForce GTX 1060 ekran kartı donanım özelliklerine sahip ve diğeri 8 GB RAM bellekli, 2.6 GHz ile çalışan Intel Core i7-9750H işlemci ve Nvidia GeForce GTX 1650 ekran kartı donanım özelliklerine sahip iki farklı dizüstü bilgisayar kullanılmıştır. Şekil 1'de çalışmamızda kullanılan bazı spektrogram örnekleri görülmektedir.



Şekil 1. "Kuş Ötmesi (Chirping birds)" kategorisinden bir ses örneği için (a) ve bu ses örneğine arka plan sesi eklenerek (b) oluşturulmuş spektrogram örnekleri.

V. YALNIZCA LIBROSA SPEKTROGRAMLARI KULLANILARAK OLUŞTURULAN MODEL

Veri kümesi, Librosa [8] kütüphanesinde bulunan ve spektrogram oluşturmaya yarayan yöntem ile spektrogram haline getirilmiştir. Daha sonra bu model için veri kümesi istisnai olarak %70'i eğitim, %20'si doğrulama ve %10'u test olmak üzere ayrılmıştır. Tasarım şu şekildedir, önceden eğitilmiş (pretrained) "MobileNetV2" [14] sinir ağı modelinin katman yapısının üzerine iki boyutlu yapıdaki *GenelOrtalamaBiriktirme2B* (GlobalAveragePooling2D) katmanı eklenmiştir. Sonrasında 512 birimlik Gizli Katman (Hidden Layer) oluşturulmuştur. İletim Sönümü (Dropout) işlemi gerçekleştirilip on birimlik Çıkış Katmanı (Output Layer) oluşturulmuştur. Çalışmamızda Tablo 1'de görülen bu modele "Tasarım1" ismi verilmiştir.

Tablo 1. Oluşturulan ilk tasarımın mimari yapısı.

MobileNetV2 ve Evrişimsel Yapı
MobileNetV2
GenelOrtalamaBiriktirme2B
512 Yoğun Katman-ReLU
İletim Sönümü 0.3
10 Yoğun Katman-Softmax

VI. SOUNDNET KULLANILARAK ELDE EDİLEN SONUÇ

SoundNet modeli ve buna ait kütüphanesi [17], [18], bu çalışmamızda da kullanılan ESC-50 (Environmental Sound Classification) [10] veri kümesi ve DCASE (Detection and Classification of Acoustic Scenes and Events) [20] veri kümesinden veriler kullanılarak oluşturulmuş literatürdeki hazır bir ses tanıma modelidir. Bu modele verilmesi için içerisinde çevresel seslerin bulunduğu bir bilgisayar oyununun oynanış videosundan [21] üç dakikalık bir kesit alınıp bu modele verilmiştir. Ancak elde edilen tahmin tatmin edici olmamıştır. Örneğin, videonun bir kısmında: %43 olasılıkla televizyon stüdyosunda olunduğuna, %12 olasılıkla akvaryumda olunduğuna, %24 olasılıkla kayak yapıldığına dair tahminler yapılmıştır ancak, o kısımda bu olayların hiçbiri gerçekleşmemiştir. Sadece belirli kısımlardaki doğru çevresel sahneleri betimleyen tahminlemeler yapılabilmiştir. Bu nedenle kendi geliştirdiğimiz modelin daha sağlıklı tahminleme sonuçları elde edebilmesi için burada bahsedildiği gibi SoundNet altyapısı da incelenmiş ve deneysel olarak değerlendirmiştir.

VII. MFCC TİPİNDE ÖZNİTELİK ELDE EDİLMESİYLE ÖZNİTELİK KÜMESİNİN ARTTIRILMASI

Elde bulunan veri kümesi sınırlı olduğu için öznitelik kümesinin arttırılması modelin daha iyi deneysel sonuçlar vermesini sağlayacaktır. Bu nedenle öznitelik kümesinin arttırılması için eldeki spektrogram şeklindeki görseller biçemindeki ses segmentlerine Librosa [8] kütüphanesi yardımıyla MFCC tipinde ses öznitelikleri elde etme yöntemi uygulanmıştır ve öznitelik kümesi arttırılmıştır. Daha sonra aynı tasarım ve MobileNetV2 [14] sinir ağı modeli kullanılmıştır. Çalışmamızda bu yapıya "Tasarım2" ismi verilmiştir.

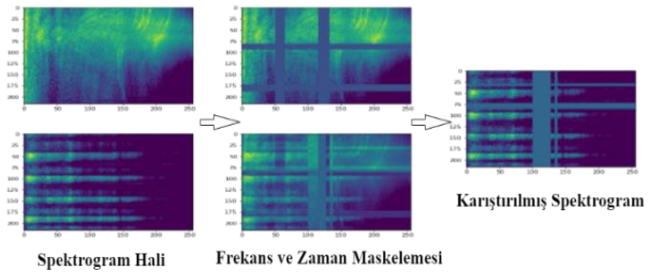
VIII. SES ÖZNİTELİKLERİNİN BİRLİKTE KULLANIMIYLA ÖZNİTELİK KÜMESİNİN ARTTIRILMASI

Veri kümesindeki öznitelik çeşitlerinin arttırılması için bir başka yöntem olan veri kümesindeki seslere Librosa [8] kütüphanesi yardımıyla Chroma tipinde öznitelik elde etme yöntemi uygulanmıştır ve daha önceden MFCC tipinde öznitelik elde etme yöntemi ile oluşturulan mevcut verilerin üzerine eklenmiştir. Arttırılan verilerle beraber yeni model

eğitilmiştir. Çalışmamızda bu yapıya "Tasarım3" ismi verilmiştir.

IX. ÇOKLU MASKELEME İLE ELDE EDİLEN MODEL

Veri kümesindeki verileri arttırmak amacı ile yine bir başka yöntem olan Tensorflow [16] kütüphanesi yardımıyla spektrogram alınıp Çoklu Maskeleme [11] uygulanmıştır. Çoklu Maskeleme işlemi [22], veri arttırmak ve karıştırmak için bir spektrogramın verilen parametrelere göre rastgele kısımlardan kesme düzlemi çizerek eğitim dağılımını genişletmektedir. Frekans maskesi parametresi 36, zaman maskesi parametresi 24 alındığında en uygun maskeler oluşturulduğu için bu değerler seçilmiştir. Verilen her spektrogramda 2 tane frekans, 2 tane zaman maskesi oluşturulmuştur. Seslerden öncelikle bir kısım kesilmiş ve ortak kategoride olan sesler birbirine karıştırılarak yeni veriler elde edilmiştir. Örneğin, "*İnsan-Konuşma İçermeyen Sesler*" kategorisinde bulunan "*Nefes Alma*" ve "*Diş Fırçalama*" sesleri birleştirilip oluşan yeni verilerle veri kümesi arttırılmıştır. Bu model oluşturulurken ise daha önceden bu yöntem kullanılarak oluşturulan modelin kullanıldığı literatürde bulunan bir çalışmadan faydalanılmıştır [23]. Çoklu Maskeleme işlemi, MFCC [6] ve Chroma [7] tipinde öznitelik elde etme yöntemi ile çoğaltılan veri olmadan yalnızca en başta elde edilen spektrogramların üzerine eklenerek veri kümesi oluşturulmuş ve model eğitilmiştir. Çalışmamızda Şekil 2'de örnekleri görülen bu yapıya "Tasarım4" ismi verilmiştir.



Şekil 2. "Uçak (Airplane)" ve "Testere (Chainsaw)" kategorilerinde bulunan spektrogram örnekleri ile çoklu maskeleme işleminin anlatılması.

X. PROGRAMATİK OLUŞTURULAN MODELE VERİLERİN VERİLMESİ

Literatürdeki bir çalışmadaki model [24] temel alınarak geliştirilen tasarımsal modelin Python dilinde kodlanması sayesinde programatik olarak gerçeklenimi şu şekilde çalışmaktadır. Öncelikle dört Evrişimsel Katmanlı bir ESA modeli oluşturulmaktadır. Her Evrişimsel Katmanı oluşturulduktan sonra Toplu Normalleştirme (Batch Normalization) yapılmaktadır. Her iki Evrişimsel Katmanda bir Biriktirme (Pooling) işlemi gerçekleşmektedir. Çalışmamızdaki modellerde, Doğrultulmuş Doğrusal Birim (Rectified Linear Unit-ReLU) kullanılarak ilgili aktivasyon işlemleri de gerçekleştirilmektedir. En sonda yapılan Biriktirme işleminden sonra ise iki tane Kapılı Tekrarlayan Birim (Gated Recurrent Units-KTB) için hiperbolik tanjant (tanh) aktivasyonu ile model oluşturulmaktadır. İletim sönümü ardından "Softmax" işlemi yapılıp sınıflandırma sonucu elde edilmektedir.

Bu programatik yapıya verilen veri kümesi ise bir önceki kısımdaki bahsi geçen veri kümesinin aynısıdır. Çalışmamızda Tablo 2'de görülen bu modele "Tasarım5" ismi verilmiştir.

Tablo 2. Dört Evrişimsel Katman kullanılarak oluşturulan ESA ve TSA tasarımsal mimarisi

Evrişimsel TSA
3x5x32 Evrişim1-Toplu Normalleşme-ReLU
3x5x32 Evrişim2-Toplu Normalleşme-ReLU
4x3 ToplamBiriktirme
3x1x64 Evrişim3-Toplu Normalleşme-ReLU
3x1x64 Evrişim4-Toplu Normalleşme-ReLU
4x1 ToplamBiriktirme
256 KTB1-tanh-İletim Sönümü 0.5
256 KTB2-tanh-İletim Sönümü 0.5
10 Tam Bağlı Katman-Softmax

XI. OLUŞTURULAN YENİ TASARIMDAKİ EVRİŞİMSEL KATMAN SAYISININ ARTTIRILMASI

Programatik tasarımda dört katmanlı kod düzenlenerek Evrişimsel Katman sayısı sekize çıkarılmıştır. Model elli eğitim yineleme sayısı (Epoch) ile eğitilmiştir. KTB katmanları eklendiğinde daha kötü sonuç alındığı için KTB katmanlarının kullanılmamasına karar verilmiştir. Çalışmamızda Tablo 3'te görülen bu modele "Tasarım6" ismi verilmiştir.

Tablo 3. Sekiz Evrişimsel Katman kullanılarak oluşturulan ESA ve TSA ana tasarımsal mimarisi (KTB katmanları çıkarılmadan önceki hali)

Evrişimsel TSA
3x5x32 Evrişim1-Toplu Normalleşme-ReLU
3x5x32 Evrişim2-Toplu Normalleşme-ReLU
4x3 ToplamBiriktirme
3x1x64 Evrişim3-Toplu Normalleşme-ReLU
3x1x64 Evrişim4-Toplu Normalleşme-ReLU
4x1 ToplamBiriktirme
1x5x128 Evrişim5-Toplu Normalleşme-ReLU
1x5x128 Evrişim6-Toplu Normalleşme-ReLU
1x3 ToplamBiriktirme
3x3x256 Evrişim7-Toplu Normalleşme-ReLU
3x3x256 Evrişim8-Toplu Normalleşme-ReLU
2x2 ToplamBiriktirme
256 KTB1-tanh-İletim Sönümü 0.5
256 KTB2-tanh-İletim Sönümü 0.5
10 Tam Bağlı Katman-Softmax

XII. VGG19 VE TSA KULLANILARAK ELDE EDİLEN MODEL

Literatürdeki bir çalışmada [23] açıklanan "*Dikkat Bloğu*" olayı bu çalışmamıza uyarlanmıştır. Bu olayın amacı, gelen girdideki farklı noktalara odaklanmaktır. Yeni yazılan bu üçüncü programatik tasarımın temeli şu şekildedir: Öncelikle iki boyutlu Evrişimsel Katmandan geçen girdi verileri daha sonra Çift Yönlü Uzun Kısa Süreli Bellek (Long Short Term Memory-UKSB) katmanlarına sahip TSA'ya verilmektedir. Daha sonra dikkat işlemi uygulanmaktadır.

Tasarımın kodlama mimarisinde bulunan Karıştırma (Permute) işlemi ve 128 birimli Yoğun Katman (Dense) oluşturulması kısmı bu dikkat olayı çerçevesinde oluşturulmuştur. En sonunda "Softmax" fonksiyonu ile oluşturulmuş Yoğun Katmanlarına uğrayan girdilerin sonucu Tam Bağlı Katman'da çıkmaktadır. Bu mantıkla oluşturulan programatik tasarımın gerçeklenimine verilen veriler ile model eğitilmiştir. Çalışmamızda Tablo 4'te görülen bu modele "Tasarım7" ismi verilmiştir.

Tablo 4. VGG19 ile oluşturulan modelin üzerine TSA modeli oluşturulması ve bu modele gelen girdilerin farklı noktalarının dikkat olayı ile incelenmesi.

VGG19 ve TSA
VGG19
128x384 Girdi
256 Çift Yönlü UKSB
2x1 Karıştırma
128 Yoğun Katman-ReLU
Düzleştirme
512 Yoğun Katman-ReLU
İletim Sönümü 0.5
10 Tam Bağlı Katman-Softmax

XIII. DENEYSEL SONUÇLAR

Deneylerdeki sonuçlar incelendiğinde elde edilen en yüksek test doğruluk değeri ve en düşük kayıp değeri elde edilen model, sekiz Evrişimsel Katman kullanılıp KTB kullanılmadan oluşturulan modeldir. Tablo 5'ten görüleceği gibi, en iyi model olan "Tasarım6" ile test sonuçlarında yüzdelik oran olarak %87 oranında bir doğruluk oranı değeri ve 0.88 değerinde bir kayıp elde edilmiştir. Bu da modelin kabul edilebilir bir başarıya sahip olduğunu göstermektedir. Oluşturulan her model, bir önceki modellerde elde edilen sonuçların iyileştirilmesi için oluşturulmuş veya önceki model üzerinde yapılmış olan değişikliktir.

Tablo 5. Oluşturulan modellerde alınan doğruluk ve eğitim kayıp değerleri

Model	**Deneylerde Elde Edilen Sonuçlar**	
	Doğruluk Oranı	***Eğitim Kayıp Değeri***
Tasarım1	0,60	1,37
Tasarım2	0,74	0,73
Tasarım3	0,59	1,08
Tasarım4	0,71	0,81
Tasarım5	0,73	1,24
Tasarım6	0,87	0,88
Tasarım7	0,42	1,82

Doğruluk (accuracy) oranı, sınıflandırıcının bir deneydeki veri kümesi üzerindeki sınıf ayrıştırıcılığı yeteneğini belirlemek amacıyla yaygın olarak kullanılan nesnel ölçütlerden birisidir. Literatürdeki çapraz tahmin (confusion matrix) tablosuna göre; doğru pozitif (*DP*), yanlış pozitif (*YP*), doğru negatif (*DN*) ve yanlış negatif (*YN*) ölçümleriyle değerlendirilerek aşağıdaki Denklem (1) ile verilmektedir. Bu değer çalışmalarda hem ondalık hem de genişletilerek yüzdelik değer olarak da gösterilebilmektedir [25].

$$Doğruluk = (DP + DN) / (DP + YP + DN + YN) \quad (1)$$

Deneylerde elde edilen kayıp değeri eğitim süreci boyunca ilgili derin sinir ağı modelinin amaç fonksiyonunun (objective function) eğitim yineleme esnasında gösterdiği değişimlerin takip edilmesinde ve eğitim sonunda eğitimin kalitesinin anlaşılmasında yararlı olmaktadır. Tablo 6'da literatürdeki bu alandaki benzer çalışmalar ile bu çalışmamızdaki "Tasarım6" isimli modelin deneysel sonuçlarının karşılaştırılmasına yer verilmektedir.

Tablo 6. Literatürdeki çalışmalar ile başarım açısından karşılaştırma

Model	**Kullanılan Özellik ve Yapı**	**Doğruluk Oranı**
Khamparia [4]	log-mel, TDYA	%56
Khamparia [4]	log-mel, ESA	%77
Piczak [2]	log-mel, ESA	%81,5
RawNet [3]	ham dalga, ESA	%85,2
AlexNet [5]	log-mel, MFCC, CRP, ETSA	%86
Çalışmamızdaki Tasarım6 modeli	Multiple Masking, log-mel, ESA	%87

Tablo 6'dan görüleceği üzere, bu çalışmadaki modelde olduğu gibi yalnızca spektrogramların log-mel özellikleri alınarak oluşturulan ESA modelleri arasından en iyi sonucun bu çalışmadaki modelde olduğu görülmektedir. ESA kullanılması yerine TDYA kullanılması, bu veri kümesinin eğitilmesinde çok daha kötü bir performansa yol açmaktadır. MFCC öznitelikleri kullanım yanı sıra CRP gibi veri arttırma yöntemlerinin kullanılması ve ESA yerine ETSA modelinin oluşturulması da modelin doğruluk oranının artırılmasında etkili olduğu görülmektedir.

XIV. SONUÇLAR VE TARTIŞMA

Bu çalışmada veriden elde edilen öznitelik çeşitliliği arttırılarak hem veri zenginleştirmede hem de sinir ağı eğitiminde düşük seviyeden özniteliklerle yüksek seviyeden öznitelikler arasındaki anlamsal boşluğun azaltılması sağlanmıştır. Çalışmamızda gerçekçi bir video deneyimi için yukarıda SoundNet [17, 18] deneyinde bahsi geçen video oyununun video parçası [21] üzerinden bu çalışmamızda yapılan ses sahnesi tahminlemesinde elde edilen sonuçlar göz önüne alındığında, geliştirdiğimiz modelle bu videonun bir kısmındaki sahnede: %30 olasılıkla doğru bir şekilde araba kornası sesi olduğuna, %1 olasılıkla uçak sesi olan sahne olarak tahminleme yapıldığında ancak bu sahnede aslında araba sesi olduğuna dair tahminlemeler yapılmıştır.

Çalışmamızda tasarlanan yeni tasarımsal modellerin gerçeklenimlerinin çeşitli tipteki ses sahnelerinin ve olaylarının sınıflandırılması ve tahminlenmesinde başarıyla kullanılabileceği anlaşılmaktadır. Bu açıdan çalışmamızda literatüre ana katkı olarak tarafımızca tasarımı ve gerçeklenimi yapılmış bu yeni derin sinir ağı modelleri önerilmektedir. İleriki çalışmalarımızda bu modeller kullanılarak daha geniş çaptaki ses sahneleri ve ses olaylarını tanıma problemlerine çözüm üretmek hedeflenmektedir.

KAYNAKLAR

[1] B., Karasulu. “Çoklu Ortam Sistemleri İçin Siber Güvenlik Kapsamında Derin Öğrenme Kullanarak Ses Sahne ve Olaylarının Tespiti” ACTA INFOLOGICA, 3(2):60-82, 2019. doi: 10.26650/acin.590690

[2] K. J. Piczak, “Environmental sound classification with convolutional neural networks”, 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA pp. 1-6. 2015. doi: 10.1109/MLSP.2015.7324337

[3] L., Shaobo, Y., Yao, J., Hu, G., Liu, X. Yao, and J., Hu. “An ensemble stacked convolutional neural network model for environmental event sound recognition”, Applied Sciences, vol. 8, no. 7 (2018): 1152. 2018. doi: 10.3390/app8071152

[4] A., Khamparia, D., Gupta, N.G., Nguyen, A., Khanna, B., Pandey and P., Tiwari, “Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network”, IEEE Access, vol. 7, pp. 7717-7727, 2019. doi: 10.1109/ACCESS.2018.2888882

[5] A. Krizhevsky, I. Sutskever, G.E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”, Advances in Neural Information Processing Systems, Editörler: F. Pereira and C.J. Burges and L. Bottou and K.Q. Weinberger, Curran Associates, Inc., Vol. 25, 2012. [Online]: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[6] MFCC (Mel-frequency cepstral coefficient) dokümantasyonu, 2022. https://librosa.org/doc/main/generated/librosa.feature.mfcc.html

[7] Chroma ses özniteliği dokümantasyonu, 2022. [Online]. https://librosa.org/doc/main/generated/librosa.feature.chroma_stft.html

[8] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg and O. Nieto. "librosa: Audio and Music Signal Analysis in Python", In Proceedings of the 14th Python in Science Conference, volume 8, 2015. doi: 10.25080/Majora-7b98e3ed-003

[9] N., Marwan, M., Thiel, N.R. and Nowaczyk. “Cross Recurrence Plot Based Synchronization of Time Series”. In: Nonlinear Processes in Geophysics 9 (2002), 325-331. 2002.

[10] ESC-50 veri kümesi, 2022. [Online]. https://github.com/karolpiczak/ESC-50

[11] Çoklu Maskeleme işlemi dokümantasyonu, 2022. [Online]. https://www.tensorflow.org/io/tutorials/audio

[12] Keras Kütüphanesi dokümantasyonu, 2022. [Online]. https://keras.io

[13] ResNet50 sinir ağı modeli dokümantasyonu, 2022. [Online]. https://keras.io/api/applications/resnet/#resnet50-function

[14] MobileNetV2 sinir ağı modeli dokümantasyonu, 2022. [Online]. https://keras.io/api/applications/mobilenet/#mobilenetv2-function

[15] VGG16 ve VGG19 sinir ağı modelleri dokümantasyonu, 2022. [Online]. https://keras.io/api/applications/vgg/#vgg16-function

[16] Tensorflow kütüphanesi dokümantasyonu, 2022. [Online]. https://www.tensorflow.org/api_docs

[17] Y., Aytar, C., Vondrick, A., Torralba. “SoundNet: Learning Sound Representations from Unlabeled Video “. arXiv preprint arXiv : 1610.09001v1 [cs.CV] 2016

[18] SoundNet kütüphanesinin Github İnternet Erişim Adresi, 2022, [Online]. https://github.com/JarbasAl/soundnet

[19] Altyapıda kullanılan çeşitli kütüphanelerin dokümantasyonları, 2022. [Online]. https://pypi.org

[20] DCASE (Detection and Classification of Acoustic Scenes and Events) veri kümesi, 2022. [Online]. http://dcase.community

[21] Kesit alınan bilgisayar oyununun oynanış videosu, 2022. [Online]. https://www.youtube.com/watch?v=d74REG039Dk

[22] D.S., Park, W., Chan., Y., Zhang, C.-C., Chiu, B., Zoph, E.D., Cubuk and Q.V., Le. “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition”. arXiv preprint: arXiv:1904.08779v3 [eess.AS]. 2019

[23] J. You and J. Korhonen. "Attention Boosted Deep Networks For Video Classification", 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 1761-1765. doi: 10.1109/ICIP40778.2020.9190996

[24] Z., Zhang, S., Xu, S., Zhang, T., Qiao and S., Cao. "Learning Attentive Representations for Environmental Sound Classification", IEEE Access, vol. 7, pp. 130327-130339, 2019. doi: 10.1109/ACCESS.2019.2939495

[25] B., Karasulu. “Kısıtlanmış Boltzmann makinesi ve farklı sınıflandırıcılarla oluşturulan sınıflandırma iş hatlarının başarımının değerlendirilmesi”, Bilişim Teknolojileri Dergisi, 11(3), 223-233, 2018. doi: 10.17671/gazibtd.370281