JIPET

Journal of Intelligent Procedures in Electrical Technology
Vol. 13/ No. 52/ Winter 2023
P-ISSN: 2322-3871, E-ISSN: 2345-5594, http://jipet.iaun.ac.ir/

https://dorl.net/dor/20.1001.1.23223871.1401.13.52.6.1
Research Article

Speech Emotion Recognition Using a Combination of Transformer and Convolutional Neural networks

Yousef Pourebrahim[1], *Ph.D. Student*, Farbod Razzazi[1], *Associate Professor*, Hossein Sameti[2], *Associate Professor*

[1]Department of Electrical and Computer Engineering- Science and Research Branch, Islamic Azad University, Tehran, Iran
[2]Speech Processing Laboratory- Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
y.pourebrahim@srbiau.ac.ir, razzazi@srbiau.ac.ir, sameti@sharif.edu

Abstract
Speech emotions recognition due to its various applications has been considered by many researchers in recent years. With the extension of deep neural network training methods and their widespread usage in various applications. In this paper, the application of convolutional and transformer networks in a new combination in the recognition of speech emotions has been investigated, which is easier to implement than existing methods and has a good performance. For this purpose, basic convolutional neural networks and transformers are introduced and then based on them a new model resulting from the combination of convolutional networks and transformers is presented in which the output of the basic convolutional network is the input of the basic transformer network. The results show that the use of transformer neural networks in recognizing some emotional categories performs better than the convolutional neural network-based method. This paper also shows that the use of simple neural networks in combination can have a better performance in recognizing emotions through speech. In this regard, recognition of speech emotions using a combination of convolutional neural networks and a transformer called convolutional-transformer (CTF) for RAVDESS dataset achieved an accuracy of %80.94; while a simple convolutional neural network achieved an accuracy of about %72.7. The combination of simple neural networks can not only increase recognition accuracy but also reduce training time and the need for labeled training samples.

Keywords: classification, deep neural networks, emotion recognition, speech signal processing

Received: 28 July 2021
Revised: 18 September 2021
Accepted: 22 October 2021

Corresponding Author: Dr. Farbod Razzazi

Citation: Y. Pourebrahim, F. Razzazi, H. Sameti, "Speech emotion recognition using a combination of transformer and convolutional neural networks", Journal of Intelligent Procedures in Electrical Technology, vol. 13, no. 52, pp. 79-98, March 2023 (in Persian).

https://dorl.net/dor/20.1001.1.23223871.1401.13.52.6.1
مقاله پژوهشی

بازشناسی احساسات از روی گفتار با استفاده از ترکیب شبکه‌های عصبی ترنسفورمر و کانولوشنی

یوسف پورابراهیم[۱]، دانشجوی دکتری، فربد رزازی[۱]، دانشیار، حسین صامتی[۲]، دانشیار

۱- دانشکده مهندسی برق و کامپیوتر- واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
۲- دانشکده مهندسی کامپیوتر- دانشگاه صنعتی شریف، تهران، ایران
y.pourebrahim@srbiau.ac.ir, razzazi@srbiau.ac.ir, sameti@sharif.edu

چکیده: بازشناسی احساسات از روی گفتار با توجه به کاربردهای متنوع آن امروزه مورد توجه بسیاری از محققان قرار گرفته است. با پیشرفت روش‌های آموزش شبکه‌های عصبی عمیق وگسترش استفاده از آن در کاربردهای مختلف، در این مقاله کاربرد شبکه‌های کانولوشنی و ترنسفورمر در یک ترکیب جدید در بازشناسی احساسات گفتاری مورد بررسی قرار گرفته که از لحاظ پیاده‌سازی نسبت به روش‌های موجود ساده‌تر بوده و عملکرد مطلوبی نیز دارد. برای این منظور شبکه‌های عصبی کانولوشنی و ترنسفورمر پایه معرفی شده و سپس مبتنی بر آنها یک مدل جدید حاصل از ترکیب شبکه‌های کانولوشنی و ترنسفورمر ارایه شده که در آن خروجی مدل کانولوشنی پایه ورودی مدل ترنسفورمر پایه است. نتایج حاصل نشان می‌دهد که استفاده از شبکه‌های عصبی ترنسفورمر در بازشناسی بعضی از حالت‌های احساسی عملکرد بهتری نسبت به روش کانولوشنی دارد. همچنین در این مقاله نشان داده شده که استفاده از شبکه‌های عصبی ساده به‌صورت ترکیبی عملکرد بهتری در بازشناسی احساسات از روی گفتار می‌تواند داشته باشد. در این رابطه بازشناسی احساسات گفتاری با استفاده از ترکیب شبکه‌های عصبی کانولوشنی و ترنسفورمر با نام کانولوشنال-ترنسفورمر (CTF) برای دادگان راودس دقتی برابر ۸۰/۹۴ درصد به‌دست آورد؛ در حالی‌که یک شبکه عصبی کانولوشنی ساده دقتی در حدود ۷۲/۷ درصد به‌دست آورد. همچنین ترکیب شبکه‌های عصبی ساده علاوه بر اینکه می‌تواند دقت بازشناسی را افزایش دهد، می‌تواند زمان آموزش و نیاز به نمونه‌های آموزشی برچسب دار را نیز کاهش دهد.

کلمات کلیدی: بازشناسی احساسات، پردازش سیگنال گفتار، شبکه‌های عصبی عمیق، طبقه‌بندی

تاریخ ارسال مقاله: ۱۴۰۰/۵/۶
تاریخ بازنگری مقاله: ۱۴۰۰/۶/۲۷
تاریخ پذیرش مقاله: ۱۴۰۰/۷/۳۰

نام نویسنده‌ی مسئول: دکتر فربد رزازی
نشانی نویسنده‌ی مسئول: دانشکده مهندسی برق و کامپیوتر، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

۱- مقدمه

تشخیص احساسات گفتاری[۱] به دلیل داشتن کاربردهای بی‌شماری مانند آموزش الکترونیکی، مطالعات بالینی، تشخیص دروغ، سرگرمی، بازی‌های رایانه‌ای و مراکز تماس، یک موضوع مهمی است که علاقه محققان را به خود جلب کرده است. با وجود این، تشخیص احساسات گفتاری هنوز هم به‌عنوان یک چالش مهم برای تکنیک‌های پیشرفته یادگیری ماشین باقی مانده و عملکرد آنها در حد متوسط است. یکی از دلایل چنین عملکرد متوسط، عدم اطمینان در انتخاب ویژگی‌های مناسب است. علاوه بر این، وجود نویز پس زمینه در اصوات ضبط شده، مانند صداهای دنیای واقعی، می‌تواند به طور چشم‌گیری بر اثر بخشی مدل یادگیری ماشین تأثیر بگذارد [۱]. با وجود این، ظهور مدل‌های تشخیص گفتار احساسی مناسب می‌تواند تجربه کاربر را در سیستم‌های متقابل انسان و ماشین به‌طور قابل توجهی در زمینه‌های هوش مصنوعی[۲] و یا سنجش سلامت از راه دور بهبود دهد [۲]. در واقع، توانایی تشخیص احساسات از نمونه‌های صوتی و بنابراین، توانایی تقلید از این احساسات می‌تواند تأثیر قابل توجهی در زمینه هوش مصنوعی داشته باشد.

در حال حاضر، از مدل‌های یادگیری عمیق برای کاربردهایی مانند شناسایی چهره، تشخیص صدا، تشخیص تصویر و تشخیص احساسات گفتاری استفاده می‌شود [۶-۳]. یکی از مزایای اصلی تکنیک‌های یادگیری عمیق در این است که می‌تواند یک ویژگی خاص را که در ارتباط با یک احساس مشخص است و در گفتار بیان شده نهفته است استخراج کند [۷].

در سال‌های اخیر، مدل‌های مختلفی مبتنی بر شبکه‌های عصبی عمیق[۳] برای تشخیص احساسات گفتاری معرفی شده‌اند. در برخی از این مدل‌ها، شبکه عصبی برای استخراج ویژگی‌های خاص از روی گفتار خام مورد استفاده قرار می‌گیرند [۸] و در گروهی دیگر یک سری ویژگی‌های از قبل استخراج شده از فایل صوتی، به عنوان ورودی به مدل مبتنی بر شبکه عصبی اعمال می‌شوند [۹،۱۰].

استفاده از روش‌هایی غیر از شبکه‌های عصبی در بازشناسی احساسات گفتاری نیز مورد استفاده قرار گرفته که در آنها بیشتر هدف استخراج ویژگی‌هایی بوده است که بتوانند در بازشناسی احساسات حالت‌های احساسی مختلف را نمایندگی کنند و قابلیت جداپذیری را داشته باشند. لذا با استخراج این ویژگی‌ها، ماشین برداری پشتیبان[۴] جهت طبقه‌بندی به‌کارگیری شده است [۱۱،۱۲].

ویژگی‌ها را می‌توان در دو نوع استاتیک و دینامیک تقسیم‌بندی کرد. با توجه به نوع ویژگی مورد استفاده برای بازشناسی، نوع شبکه عصبی تعیین می‌شود. برای مثال اگر ویژگی‌ها از نوع استاتیک باشند، معمول است که از شبکه‌های عصبی پرسپترون استفاده شود. اما برای ویژگی‌های دینامیک از شبکه‌های عصبی کانولوشنی[۵]، بازگشتی[۶] و یا ترکیبی از هر دو مورد استفاده قرار می‌گیرند.

شبکه‌های عصبی کانولوشنی با توجه به ابعاد ورودی آن به دو نوع شبکه عصبی کانولوشنی تک بعدی و دو بعدی تقسیم می‌شوند. در شبکه عصبی تک بعدی، ورودی‌ها می‌توانند ترکیبی از انواع ویژگی‌ها باشند که معمولا این نوع ویژگی‌ها از اعمال توابع استاتیک روی ویژگی‌های دینامیک به‌دست می‌آیند [۱۳،۱۴]. در شبکه‌های عصبی کانولوشنی دو بعدی ورودی یک تصویر دو بعدی است. در مورد بازشناسی احساسات از گفتار این تصویر معمولا طیف گفتار است که ساختار زمان-فرکانس گفتار را نشان می‌دهد و شبکه‌های کانولوشنی مبتنی بر این تصویر ورودی، ویژگی‌های مرتبط با احساسات را در لایه‌های بالاتر استخراج می‌کنند [۱۷-۱۴]. بنابراین به‌کارگیری شبکه‌های کانولوشنی دو بعدی به‌منظور استخراج ویژگی‌های تعمیم‌پذیر و متمایزگر از تصویر ورودی، یک گام به طرف بازشناسی برخط احساسات از گفتار است [۱۶]. همچنین تصویر ورودی به شبکه عصبی کانولوشنی می‌تواند به‌عنوان ورودی شبکه عصبی ترنسفورمر[۷] نیز باشد. در این حالت شبکه عصبی ترنسفورمر ارتباطات زمانی را از تصویر استخراج خواهد کرد.

با توجه به مطالعات قبلی، در این مقاله نیز تنها از یک نوع ویژگی با نام ضریب کپسترال فرکانسی در مقیاس مل[۸] (MFCC) به‌صورت تصویر برای بازشناسی احساسات گفتاری استفاده می‌شود و شبکه کانولوشنی با توجه به تصویر ورودی ساخته شده از ویژگی‌های MFCC، ویژگی‌های مختص احساسات را استخراج کرده و عمل طبقه‌بندی را انجام می‌دهد. برای محاسبه

ویژگی‌های MFFC، گفتار احساسی به فریم‌هایی تقسیم شده و در هر فریم ویژگی‌های بیان شده محاسبه می‌شوند و برای یک گفتار یک ماتریس به‌دست می‌آید که هر سطر آن ویژگی‌های محاسبه شده برای یک فریم است.

برای بازشناسی احساسات دو نوع شبکه عصبی کانولوشنی و ترنسفورمر مورد استفاده قرار گرفته است. بنابراین در مورد شبکه عصبی کانولوشنی، ماتریس ویژگی‌ها شبیه یک تصویر دو بعدی خواهد بود که به‌عنوان ورودی به شبکه داده می‌شود. در مورد شبکه ترنسفورمر نیز این ماتریس ویژگی‌ها به‌صورت یک دنباله به‌عنوان ورودی به شبکه اعمال می‌شوند.

در بازشناسی احساسات گفتاری ترکیبی از انواع ویژگی‌ها در معماری‌های مختلف مورد استفاده قرار گرفته است [۱۸] و نتایج مناسبی نیز به‌دست آورده شده است. با این وجود، استفاده از یک نوع ویژگی برای بازشناسی احساسات زمان پیش-پردازش را کاهش داده و در نتیجه یک گام رو به جلو در بازشناسی برخط احساسات از روی گفتار است. همچنین هدف اصلی این مقاله مقایسه عملکرد انواع شبکه عصبی عمیق در کاربردهای بازشناسی احساسات از روی گفتار است و مبتنی بر عملکرد آنها یک ساختار جدید ارایه شده است. لذا استفاده از ترکیب انواع ویژگی‌ها برای بازشناسی احساسات گفتاری هدف اصلی این مقاله نیست؛ زیرا که به‌کارگیری ترکیبی از ویژگی‌ها در شبکه‌های عصبی کانولوشنی دو بعدی متداول نبوده و مطالعات نشان داده‌اند که استفاده از ترکیب ویژگی‌ها در شبکه کانولوشنی تک بعدی با وجود صرف زمان زیاد برای پیش-پردازش، عملکرد زیاد مطلوبی در مقایسه با شبکه دوبعدی که تنها از تصویر طیف سیگنال گفتار احساسی استفاده می‌کند ندارد [۱۳].

شبکه ترنسفورمر در مرجع [۱۹] معرفی شده است. مدل ترنسفورمر در ابتدا برای ترجمه ماشینی پیشنهاد شده بود، اما به دلیل عملکرد بالا، به سرعت در سایر زمینه‌ها مانند تولید تصویر [۲۰]، صوت [۱۵،۲۱]، خلاصه‌سازی متن [۲۲] و موسیقی [۲۳] مورد استفاده قرار گرفت. ترانسفورمر از حالت بازگشتی و کانولوشن استفاده نمی‌کند، اما در مدل‌سازی از تاکید[۹] استفاده می‌کند. با توجه به موفقیت شبکه ترنسفورمر در کاربردهای اشاره شده، از آن در مدل پیشنهادی در این مقاله نیز استفاده شده است. با توجه به اینکه هدف این مقاله طبقه‌بندی است لذا تنها بخش کدکننده شبکه ترنسفورمر به‌کارگیری شده است. به جای شبکه ترنسفورمر امکان به‌کارگیری شبکه عصبی بازگشتی حافظه کوتاه مدت ماندگار[۱۰] (LSTM) برای یادگیری توالی طیفی[۱۱] هر احساس نیز وجود دارد، اما شبکه LSTM فقط می‌تواند پیش‌بینی تغییرات فرکانس را از روی گام‌های زمانی مجاور یاد بگیرد. در مقابل، لایه‌های چند منظوره ترنسفورمر، شبکه را قادر می‌سازد تا هنگام پیش‌بینی گام بعدی، به چندین گام زمان قبلی نیز توجه[۱۲] کند. این مساله در مدل پیشنهادی بسیار مهم است زیرا که یک احساس خاص، تنها در یک گام زمانی کوتاه وجود ندارد و در تمامی توالی فرکانس‌ها پخش شده است. این توانایی از اینجا نشأت می‌گیرد که در ترنسفورمر دنباله‌های زمانی به‌صورت یکجا به‌عنوان ورودی به شبکه داده می‌شوند؛ در حالی‌که در شبکه LSTM ورودی یک به یک[۱۳] وارد شبکه می‌شود. همچنین شبکه ترنسفورمر برخلاف LSTM از چندین مکانیسم دورن تاکیدی[۱۴] بهره می‌برد. برخلاف شبکه‌های بازگشتی، ترانسفورمر مشکلی در محو شدن شیب[۱۵] ندارد و می‌تواند بدون توجه به فاصله بین گام‌ها به هر گام گذشته دسترسی پیدا کند.

هدف مطالعه حاضر بررسی توانایی شبکه‌های عصبی به‌صورت ترکیبی در بازشناسی احساسات از گفتار است. همچنین در این مقاله از شبکه ترنسفورمر در ترکیب با شبکه عصبی کانولوشنی به‌منظور بازشناسی احساسات استفاده شده است که با توجه به مطالعات ما این اولین گزارش از کاربرد شبکه‌های عصبی ترنسفورمر در ترکیب با شبکه‌های عصبی کانولوشنی در بازشناسی احساسات از گفتار است که در آن خروجی شبکه عصبی کانولوشنی به‌عنوان ورودی شبکه ترنسفورمر بوده و شبکه صرفا با نمونه‌های گفتار احساسی آموزش می‌بینند. با این توضیحات نوآوری‌های این مقاله به‌صورت زیر هستند:

- ترکیب شبکه‌های عصبی کانولوشنی و ترنسفورمر که در آن خروجی مدل کانولوشنی به‌عنوان ورودی مدل ترنسفورمر به‌کار رفته است.
- استفاده از شبکه‌های عصبی کانولوشنی موازی
- استفاده از شبکه‌های عصبی کانولوشنی و ترنسفورمر به شکل موازی

در بخش دوم مروری بر کارهای مرتبط انجام گرفته است. در بخش سوم مدل‌های مختلف شبکه عصبی برای بازشناسی احساسات گفتاری تشریح شده است و یک مدل ترکیبی جدید مبتنی بر آنها پیشنهاد داده شده است. در بخش چهارم

داده‌های مورد استفاده به همراه معیارهای سنجش عملکرد همراه با نتایج به‌دست آمده نشریح شده است. در بخش آخر نیز جمع‌بندی آورده شده است.

۲- مرور کارهای مرتبط

بیشتر معماری‌های تشخیص احساسات گفتاری که از شبکه‌های عصبی استفاده می‌کنند مبتنی بر شبکه‌های عصبی کانولوشنی، شبکه‌های عصبی بازگشتی با دنباله‌ای بلند از حافظه‌های کوتاه مدت و یا ترکیبی از آنها هستند [۲،۷،۸]. ترکیبی از شبکه کانولوشنی و بازگشتی می‌تواند الگوهای متمایزگر را در فایل‌های صوتی در زمان استخراج ویژگی‌ها و طبقه‌بندی گویندگان تشخیص دهد [۷،۸]. یکی از اهداف اصلی در شناخت احساسات گفتاری، شناسایی ویژگی‌های خاصی برای آموزش یک مدل کارآ است. محققان برای حل این مشکل از رویکردهای مختلفی استفاده کرده‌اند. در مرجع [۸] از داده‌های صوتی خام به‌عنوان ورودی برای مدل پیشنهادی استفاده شده و شبکه عصبی کانولوشنی برای پیش-پردازش نمونه‌های صوتی با هدف کاهش نویز و تأکید بر مناطق خاصی از گفتار بیان شده به‌کارگیری شده است.

ماشین بردار پشتیبان در مرجع [۲۴] برای طبقه‌بندی نمونه‌های گفتار مردانه در دادگان راودس[۱۶] [۲۵] استفاده شده است. نویسندگان هنگام انتخاب ویژگی‌ها، تبدیل پیوسته موجک را اعمال کرده و ویژگی‌های انتخاب شده را به انواع مختلف طبقه‌بندی ماشین بردار پشتیبان تغذیه کرده‌اند. بهترین نتیجه با دقت ۶۰/۱ درصد با ماشین بردار پشتیبان درجه دوم با تکنیک اعتبار سنجی پنج دسته‌ای[۱۷] به‌دست آورده شده است [۲۴].

رویکرد دیگری به نام یادگیری چند وظیفه‌ای توسط ژانگ و همکاران به‌کارگیری شده است [۲۶]. آنها ویژگی‌های گفتارهای آهنگین و گفتارهای بدون آهنگ را از مجموعه داده راودس استخراج کرده و نشان داده‌اند که با ترکیب این ویژگی‌ها و اعمال آنها به‌عنوان ورودی به یک طبقه‌بندی کننده می‌توان به دقت بالاتری دست پیدا کرد. نویسندگان فقط از ۴ کلاس احساسات (عصبانی، شاد، خنثی و غمگین) از ۸ مورد استفاده کرده‌اند. آنها برای این چهار کلاس به دقت ۵۷/۱۴ درصد دست یافته‌اند.

همین ایده در مرجع [۲۷] با استفاده از یک شبکه عصبی عمیق اجرا شده است. آنها با استفاده از اسپکتروگرام[۱۸] تولید شده از گفتارهای آهنگین و بدون آهنگ مجموعه داده راودس به‌عنوان ورودی شبکه باقیمانده دروازه‌ای[۱۹]، به دقت ۶۵/۹۷ درصد دست یافته‌اند. از طرف دیگر، در مرجع [۲۸] از شبکه عصبی پیشنهادی توسط گروه هندسه بصری[۲۰] (VGG) دانشگاه آکسفورد با نام VGG-16 استفاده شده (عدد ۱۶ نشان‌دهنده تعداد لایه کانولوشنی استفاده شده در مدل است) و با باز آموزش بخش طبقه‌بندی کننده شبکه مذکور با استفاده از مل اسپکتروگرام‌های[۲۱] به‌دست آمده از نمونه‌های گفتار مجموعه داده راودس به دقت ۷۱ درصد دست یافته‌اند.

یک معماری جدیدی مبتنی بر شبکه‌های عصبی کانولوشنی در مرجع [۱۳] پیشنهاد شده که در آن ورودی‌ها ترکیبی از ویژگی‌های ضرایب MFCC، طیف نگاری با مقیاس مل[۲۲] (MSS)، کروماگرام[۲۳]، کنتراست[۲۴] و تونتز[۲۵] هستند. این ویژگی‌ها از فایل‌های صوتی استخراج شده و میانگین آنها به‌عنوان بردار ورودی با تعداد ۱۹۳ عنصر به یک شبکه عصبی یک بعدی کانولوشن اعمال شده است تا برای شناسایی احساسات با استفاده از نمونه‌هایی از پایگاه دادگان راودس به‌کارگیری شود. مدل پایه پیشنهادی آنها به دقت ۷۱/۶۱ درصد در حالت طبقه‌بندی ۸ کلاسی دست یافته است.

در مرجع [۲۹] یک ساختار برای تشخیص احساسات گفتاری با نام کانولوشنال-LSTM[۲۶] (CLSTM) ارائه شده که از بلوک‌های LSTM در ترکیب با شبکه کانولوشنی تشکیل شده است. نویسندگان برای استخراج ویژگی‌های عاطفی محلی، چهار بلوک کانولوشنال-LSTM را در طراحی مدل پیشنهادی به‌کار برده‌اند تا ویژگی‌های مکانی-زمانی[۲۷] را در سیگنال‌های گفتاری استخراج کنند. همچنین از یک تابع هزینه دیگری با نام تابع هزینه مرکزی[۲۸] به همراه تابع هزینه سافت‌ماکس[۲۹] برای تولید احتمال کلاس‌ها بهره برده‌اند. تابع هزینه مرکزی در بهبود نتایج طبقه‌بندی عملکرد موثری نشان داده است. دقت به‌دست آمده در بازشناسی احساسات نیز در حدود ۷۹ درصد بوده است.

در مرجع [۱۷] یک چارچوب جدید برای تشخیص احساسات گفتاری با استفاده از انتخاب یک قطعه دنباله کلیدی بر اساس اندازه‌گیری شباهت در خوشه‌ها با استفاده از شبکه مبتنی بر تابع آربی[۳۰] معرفی شده است. قطعه گفتار انتخاب شده به

اسپکتروگرام تبدیل می‌شود و برای استخراج ویژگی‌های افتراقی[۳۱] و برجسته[۳۲] به مدل شبکه عصبی کانولوشنی منتقل می‌شود. خروجی بخش شبکه عصبی کانولوشنی به یک شبکه LSTM دو جهته[۳۳] اعمال شده تا اطلاعات زمانی را برای تشخیص حالت نهایی احساسات یاد بگیرد. در این روش، قطعه‌های گفتاری کلیدی به‌جای کل گفتار پردازش می‌شود تا از پیچیدگی محاسباتی کاسته شود. روش مذکور برای دادگان راودس نتیجه ۷۷ درصد به‌دست آورده است.

۳- روش‌های پیشنهادی

در این بخش مدل‌های پیشنهادی برای بازشناسی احساسات از روی سیگنال گفتار ارایه شده است. مدل‌های پیشنهادی به‌صورت ترکیبی از مدل‌های شبکه عصبی کانولوشنی و ترنسفورمر هستند که در این مقاله هر یک از مدل‌های شبکه عصبی کانولوشنی و ترنسفور خود به‌عنوان مدل پایه به‌کار برده می‌شوند تا عملکرد روش‌های پیشنهادی علاوه بر روش‌های پیشرفته، نسبت به آنها نیز مورد بررسی قرار گیرد. به همین منظور در ادامه شبکه‌های عصبی کانولوشنی و ترنسفورمر به‌عنوان مدل پایه تشریح شده‌اند.

۳-۱- شبکه عصبی کانولوشنی

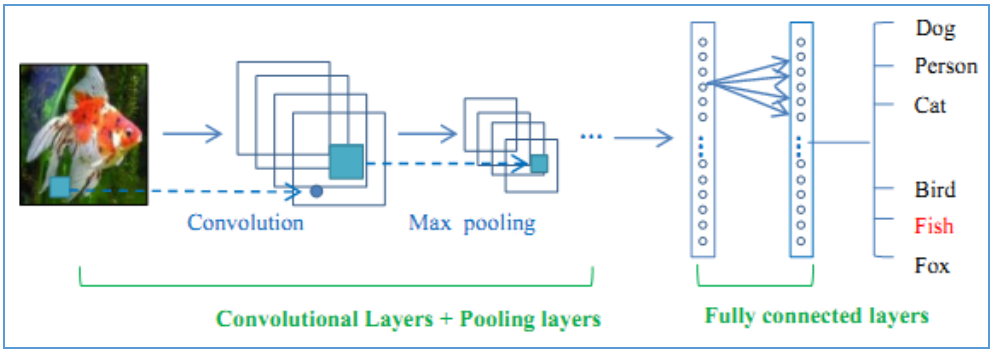
شبکه‌های عصبی کانولوشنی در زمینه پردازش تصویر موفقیت‌های زیادی کسب کرده‌اند. این شبکه‌ها می‌توانند ویژگی‌های محلی متمایزگر را از تصویر استخراج کنند [۳۰]. به‌طور کلی، یک شبکه عصبی کانولوشنی دارای سه لایه اصلی: لایه کانولوشنی، لایه انتخاب‌گر و لایه تماماً متصل است. هر یک از لایه‌های تشکیل دهنده کارهای مختلفی را انجام می‌دهند. شکل (۱) ساختار کلی یک شبکه عصبی کانولوشنی را برای طبقه‌بندی تصاویر به‌صورت لایه به لایه نشان می‌دهد.

در لایه کانولوشنی، فیلترها مهمترین نقش را دارند. این فیلترها در یک ساختار پنجره‌ای با تصاویر ورودی کانوالو می‌گردند تا ویژگی‌های محلی استخراج شوند. به این ترتیب در یک شبکه عصبی کانولوشنی چند لایه، در لایه‌های ابتدایی ویژگی‌های پایه مانند لبه و خطوط و در لایه‌های بعدی ویژگی‌های اختصاصی‌تر استخراج می‌شوند [۳۰،۳۱]. برای این منظور تعداد D' فیلتر با سایز $m\times n\times D$ که m و n طول و عرض فیلتر و D تعداد کانال‌های تصویر ورودی است، با تصویر ورودی کانوالو می‌شود. به‌عبارتی اگر داده ورودی I با ابعاد $W\times H\times D$ را با مجموعه فیلترهای $F^{d'}$ $(d'=1,2,\cdots,D')$ و ابعاد هر کدام از فیلترها $m\times n\times D$ است) در یک لایه کانولوشن کانوالو کنیم، خروجی G با ابعاد $W'\times H'\times D'$ توسط رابطه (۱) به‌دست می‌آید [۳۲].

$$G^{d'}(x,y)=f(b^{d'}+I*F^{d'})=f\left(b^{d'}+\sum_{s=-(m-1)/2}^{(m-1)/2}\sum_{t=-(n-1)/2}^{(n-1)/2}\sum_{d=1}^{D}F_d^{d'}(s,t)I_d(x+s,y+t)\right) \quad (۱)$$

$$\forall d'=1,2,\ldots,D' \qquad \forall x=1,2,\ldots,W' \qquad \forall y=1,2,\ldots,H'$$

که در آن $I_d(x,y)$ مقدار ماتریس در موقعیت (x,y) و کانال d از داده ورودی I، $G^{d'}(x,y)$ مقدار ماتریس در موقعیت (x,y) و کانال d' از داده خروجی G، $F_d^{d'}(s,t)$ مقدار وزن فیلتر در موقعیت (s,t) و کانال d از فیلتر $F^{d'}$، $b^{d'}$ مقدار بایاس در نظر گرفته شده برای کانال d' از خروجی G، f تابع فعال‌سازی و $*$ نشانگر عمل کانولوشن دو بعدی هستند.



شکل (۱): نمایی از ساختار یک شبکه کانولوشنی

Figure (1): View of the structure of a convolutional network

لایه‌های استخراج به‌منظور کاهش ابعاد ویژگی‌های ورودی به لایه‌های بعدی مورد استفاده قرار می‌گیرند و در نتیجه، پیچیدگی محاسباتی را در لایه‌های بعدی کاهش می‌دهند. از انواع روش‌های استخراج می‌توان به استخراج حداکثر[۳۴]، استخراج حداقل[۳۵] و یا استخراج میانگین[۳۶] اشاره کرد. استخراج حداکثر بیشتر مورد استفاده قرار می‌گیرد که در آن حداکثر مقدار به‌عنوان خروجی در نظر گرفته می‌شود. به عبارتی دیگر، یک پنجره لغزان با ابعاد m×n بر روی داده دو بعدی ورودی با اندازه گام s و در جهت‌های عمودی و افقی حرکت می‌کند و حداکثر مقدار داده را به عنوان خروجی پنجره در نظر می‌گیرد. لایه‌های تماماً متصل یک فضای m بعدی از ویژگی‌ها را با استفاده از یک ترکیب غیرخطی به یک فضای n بعدی دیگر از ویژگی‌ها $(R^m \to R^n)$ نگاشت می‌کند. به عبارتی اگر $x \in R^m$ یک بردار ویژگی باشد که در خروجی نودهای لایه یکم ایجاد شده‌اند، در این صورت ویژگی تولید شده در خروجی نود iام از لایه (l+1) به‌صورت رابطه (۲) است:

$$y_i^{l+1} = f\left(\sum_{j=1}^{m} w_{i,j} x_j^k + b_i\right) \qquad \forall i = 1,2,\ldots,n \qquad (۲)$$

که $w_{i,j}$ و b_i به‌ترتیب وزن و بایاس قابل آموزش بوده و f تابع فعال‌سازی است. از کنار هم قرار گرفتن لایه‌های تماماً متصل، یک شبکه تماماً متصل[۳۷] ایجاد می‌شود. در شکل (۲) مدل پایه شبکه عصبی کانولوشنی پیشنهادی نشان داده شده است. این مدل شامل سه لایه کانولوشنی بوده که بعد از هر لایه کانولوشنی جهت کاهش اندازه بازنمایی‌های به‌دست آمده یک لایه انتخاب مقدار بیشینه[۳۸] قرار دارد. خروجی این لایه در لایه سوم از شکل ماتریسی به شکل برداری تغییر داده می‌شود تا جهت طبقه‌بندی به لایه بعدی که یک لایه از نوع سافت ماکس است، اعمال شود. با توجه به اینکه تعداد هایپر پارامترهای یک شبکه عصبی زیاد هستند، لذا در این مقاله تعداد لایه‌های شبکه ثابت فرض شده است و بقیه پارامترها از طریق آزمون و خطا با استفاده از جستجو در یک مجموعه مشخص و مبتنی بر دقت به‌دست آمده برای داده‌های اعتبارسنجی[۳۹] تعیین شده‌اند. داده‌های اعتبار سنجی در انتهای هر دوره آموزش شبکه به‌منظور پیگیری روند آموزش استفاده می‌شوند. هر زمان که دقت به‌دست آمده برای نمونه‌های اعتبارسنجی افزایش قابل توجهی نداشته باشد و نمودار مربوط به آن به شکل صاف دربیاید، نشان دهنده این است که شبکه ظرفیت بیشتری برای آموزش ندارد و ادامه روند آموزش باعث آموزش بیش از حد شبکه می‌شود و درنتیجه کارآیی آن برای نمونه‌های آزمون به شدت کاهش می‌یابد. برای تمامی لایه‌های کانولوشنی فیلترها از اندازه ۳×۳ تشکیل شده‌اند. تعداد فیلترها در لایه اول برابر ۱۶ و در لایه دوم برابر ۳۲ و در آخرین لایه برابر ۶۴ انتخاب شده است. اولین لایه انتخاب مقدار بیشینه از فیلترهای ۲×۲ تشکیل شده است. هدف این لایه انتخاب مقدار بیشنه در هر پنجره ۲×۲ است و به این ترتیب خروجی این لایه باعث کاهش اندازه تصویر ورودی می‌شود. دومین و سومین لایه انتخاب مقدار بیشینه نیز دارای فیلترهایی از اندازه ۴×۴ هستند. سایر جزییات در شکل آورده شده است. هدف از به‌کارگیری لایه نرمالایزر جهت افزایش سرعت آموزش شبکه بوده و هدف از لایه حذفی[۴۰] جلوگیری از بیش آموزش[۴۱] شبکه است.

۲-۳- شبکه عصبی ترنسفورمر

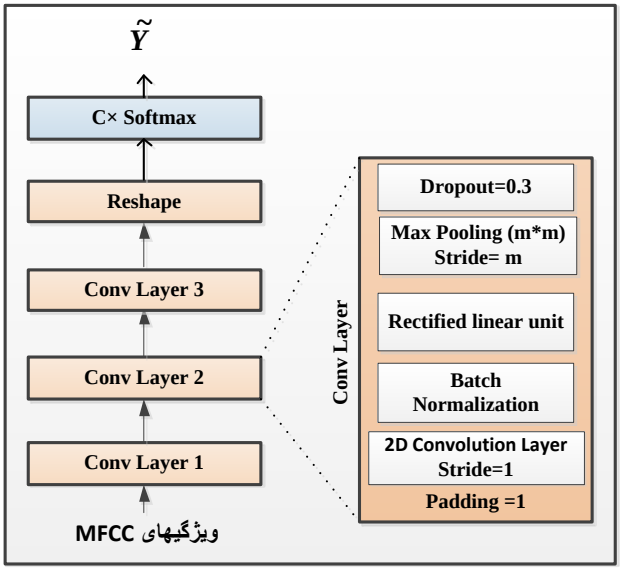
در هر شبکه ترنسفورمر تعداد N کدکننده به‌صورت پشت سر هم قرار دارد که مهمترین بخش تشکیل دهنده هر کدکننده در یک شبکه ترنسفورمر بخش درون تاکیدی چند شاخه[۴۲] است که خود متشکل از تعداد h تاکید است. ساختار شبکه کدکننده در شکل (۳) نشان داده شده است. روابط ورودی و خروجی در هر کدکننده برای یک دنباله ورودی $x \in R^{t \times d}$ به‌صورت رابطه (۳) است. t تعداد گام‌های زمانی بوده و d تعداد ویژگی‌ها در هر گام زمانی را نشان می‌دهد:

$$\text{Attention}_i(Q,K,V) = \text{head}_i = Z_i = \text{soft max}\left(\frac{QK^T}{\sqrt{n}}\right)V \qquad 1 < i \le h \qquad (۳)$$

$$Q = XW_{Q_i} \qquad (۴)$$

$$K = XW_{K_i} \qquad (۵)$$

$$V = XW_{V_i} \qquad (۶)$$



شکل (۲): مدل پایه شبکه عصبی کانولوشنی تشکیل شده از از سه لایه کانولوشنی (C تعداد کلاس‌ها را نشان می‌دهد.)

Figure (2): The basic model of canvolutional neural network consists of three layers of cannulation. C indicates the number of classes

در رابطه‌های (۳) تا (۶) $W \in R^{d\times n}$ وزن‌ها هستند و n=d/h است و $Z_i \in R^{t\times n}$ است. در رابطه (۳) بالا تابع سافت ماکس مقدار تاکید را برای بخش‌های خاصی از ورودی محاسبه می‌کند و بنابراین Z_i ویژگی‌هایی هستند که با استفاده از تاکیدهای محاسبه شده با استفاده از Q و K وزن‌دهی شده‌اند. خروجی‌های حاصل از m تاکید به یکدیگر الحاق[۴۳] شده و وارد یک لایه خطی می‌شوند که به‌صورت رابطه‌های (۷) و (۸) آورده شده‌اند:

$$MultiHead(Q,K,V) = (head_1,\ldots,head_m)W_O = (Z_1,\ldots,Z_m)W_O = ZW_O = Linear(Z) \quad (۷)$$

$$Z_O = LayerNorm\,(X + Linear(Z)) \quad (۸)$$

در رابطه (۷) $W_o \in R^{(hn)\times d}$ و در رابطه (۸) $Z_o \in R^{t\times d}$ هستند. لایه پیشرو خود از دولایه خطی تشکیل شده است که ما بین آنها از تابع فعال‌ساز یکسوساز استفاده شده است. ورودی این لایه برداری با ابعاد 1×d بوده و خروجی آن نیز از همان ابعاد است.

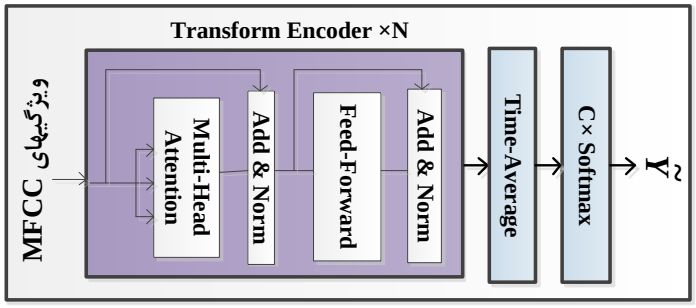
$$FN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (۹)$$

در نهایت خروجی یک کدکننده در شبکه ترنسفورمر به‌صورت زیر به‌دست می‌آید:

$$H = LayerNorm\,(Z_O + Feedforwar\,d(Z_O)) \quad (۱۰)$$

که در رابطه (۱۰) $H \in R^{t\times d}$ است که نسبت به زمان از آن میانگین گیری شده است تا یک بردار از اندازه d به‌دست آید.

با توجه به اینکه در مرجع [۱۹] تعداد ۶ کدکننده مورد استفاده قرار گرفته و افزایش تعداد کدکننده‌ها باعث افزایش زمان مورد نیاز برای آموزش شبکه می‌شود بدون آنکه کارآیی شبکه را افزایش دهد، لذا در این مقاله تعداد کدکننده‌ها از مجموعه {۴،۵،۶} و تعداد نورن‌های شبکه پیشرو[۴۴] نیز از مجموعه {۵۱۲،۱۰۲۴} انتخاب گردیده‌اند.



شکل (۳): مدل شبکه ترنسفورمر تشکیل شده از N کدکننده (این مدل به عنوان مدل پایه است.)

Figure (3): The structure of the transformer network which consists of N encoders. This model is used as the basic model

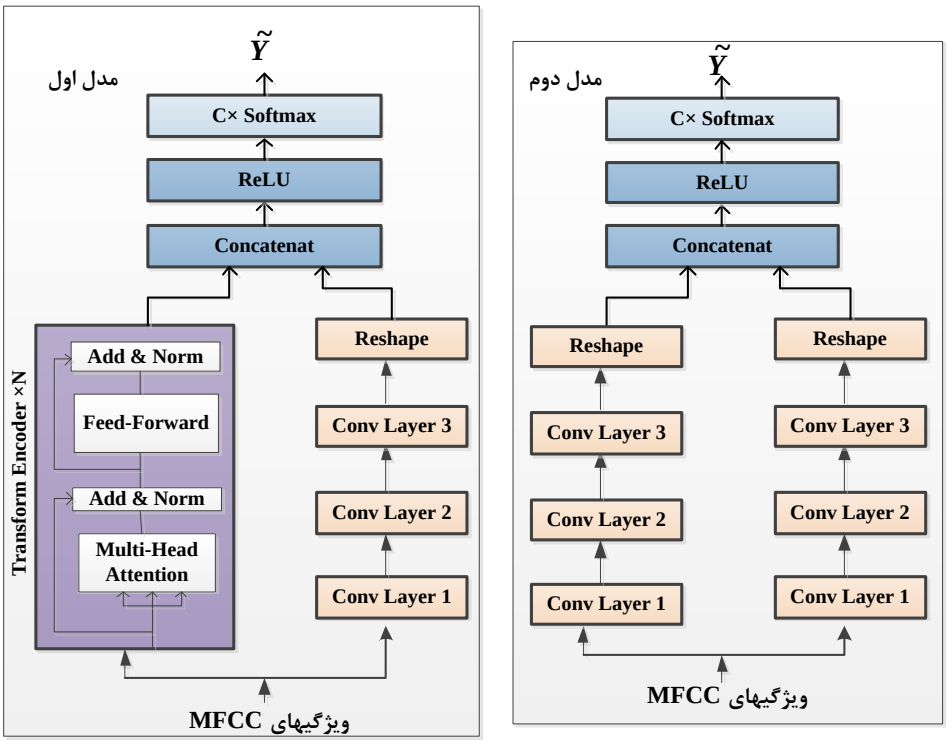
۳-۳- مدل‌های پیشنهادی

مدل اول در شکل (۴) اولین مدل پیشنهادی در این مقاله است که از موازی سازی دو مدل شبکه عصبی کانولوشنی و ترنسفورمر پایه حاصل شده است. خروجی مدل‌ها به یکدیگر الحاق شده و تشکیل یک بردار می‌دهند و سپس با عبور از یک لایه خطی، با استفاده از لایه سافت ماکس طبقه‌بندی می‌شوند. هدف از ارایه این مدل بررسی تاثیر موازی سازی شبکه ترنسفورمر با شبکه عصبی کانولوشنی در عملکرد طبقه‌بندی است.

مدل دوم در شکل (۴) دومین مدل پیشنهادی با نام شبکه عصبی کانولوشنی موازی[۴۵] (PCNN) را نشان می‌دهد که از موازی سازی دو شبکه عصبی کانولوشنی ایجاد شده است. خروجی‌های این شبکه ابتدا وارد یک لایه دنس[۴۶] با تعداد نورون برابر نصف اندازه بردار ورودی با تابع فعال‌ساز یکسوساز[۴۷] شده و سپس با عبور از یک لایه خطی وارد لایه سافت ماکس جهت طبقه‌بندی می‌شوند. هدف از ارایه این مدل بررسی تاثیر موازی سازی دو شبکه کانولوشنی با پارامترهای یکسان در عملکرد طبقه‌بندی نسبت به یک شبکه عصبی کانولوشنی پایه است. شکل (۵) مدل حاصل از ترکیب مدل ترنسفورمر و کانولوشنی پایه را با نام کانولوشنال-ترنسفورمر[۴۸] (CTF) نشان می‌دهد که در آن خروجی مدل کانولوشنی به عنوان ورودی مدل ترنسفورمر است. هدف از این مدل استخراج روابط مکانی-زمانی ویژگی‌های ورودی است. به عبارتی، مدل کانولوشنی روابط فرکانسی بین فریم‌های مجاور را در شکل MFCC ورودی استخراج می‌کند و مدل ترنسفورمر روابط زمانی بین فریم‌ها را در کل گفتار ورودی استخراج می‌کند.

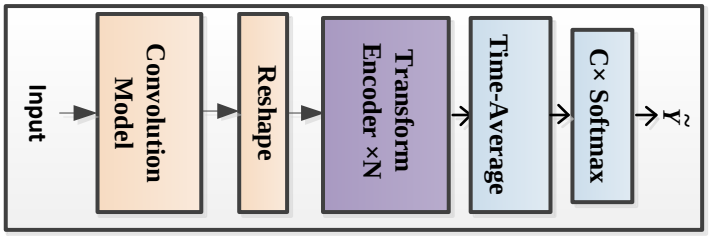
۴- تنظیمات مدل‌ها و بررسی نتایج حاصل از شبیه‌سازی‌ها

در این بخش دادگان مورد استفاده برای بررسی عملکرد مدل پیشنهادی تشریح شده است و همچنین نحوه استخراج ویژگی‌ها به همراه معیارهای سنجش عملکرد ارایه شده است. در نهایت نتایج حاصل مورد بحث قرار گرفته است.



شکل (۴): مدل اول: ساختار پیشنهادی اول که از موازی سازی شبکه‌های عصبی کانولوشنی و ترنسفورمر تشکیل شده است. مدل دوم: ساختار پیشنهادی دوم که از موازی سازی دو شبکه عصبی کانولوشنی تشکیل شده است.

Figure (4): The first model: The first proposed structure, which consists of the parallelism of the convolutional and transformer neural networks. The second model: The second proposed structure, which consists of the parallelism of two convolutional neural networks.



شکل (۵): ساختار مدل CTF. در این مدل خروجی مدل کانولوشنی پایه به عنوان ورودی مدل ترنسفورمر پایه به‌کار رفته است.
Figure (5): CNN-TF model structure. In this model, the output of the basic convolutional model is used as the input of the basic transformer model.

۴-۱- دادگان مورد استفاده در شبیه‌سازی

دادگان راودس مجموعه نمونه‌هایی از گفتار و آواز احساسی در زبان انگلیسی است که در آن گفتارهای بیان شده توسط ۲۴ بازیگر حرفه‌ای (۱۲ مرد و ۱۲ زن) مورد ضبط قرار گرفته است [۲۵]. گفتارها از روی متن‌های از پیش تعریف شده در حالت-های احساسی متفاوت بیان شده‌اند. این مجموعه شامل هشت احساس در حالت‌های: غمگین، آرام، شاد، عصبانی، متعجب، خنثی، ترس و انزجار است. در مجموع ۱۴۴۰ صدای صوتی با نرخ نمونه برداری ۴۸۰۰۰ هرتز ضبط شده است.

۴-۲- تابع هزینه

تابع هزینه در شبکه‌های عصبی، به‌منظور بهینه‌سازی شبکه عصبی از لحاظ مقدار خطا در زمان آموزش شبکه به‌کارگیری می‌شود. به عبارتی دیگر در هر دور آموزش شبکه به‌منظور تنظیم وزن‌های آن، متوسط مجموع اختلاف مقدار پیش‌بینی (مقدار احتمال رخداد هر کلاس) با مقدار واقعی در مسیر پیش‌رو مورد محاسبه قرار گرفته و به‌عنوان یک معیار از روند آموزش شبکه ارایه می‌شود. بنابراین، بهینه‌سازی تابع هزینه به معنی بهینه‌کردن مقدار متوسط خطای شبکه است. لذا تابع هزینه صرفا یک معیار برای شبکه عصبی است که در زمان آموزش شبکه مورد استفاده قرار می‌گیرد تا وزن‌های شبکه به نحوی تنظیم گردند که مقدار خطا در هر مرحله از آموزش نسبت به مرحله قبل کوچک‌تر شود. در این مقاله به‌منظور تنظیم وزن‌های شبکه از تابع هزینه کراس آنتروپی[۴۹] استفاده شده است که ورودی آن بردار احتمال پیش‌بینی شده در خروجی لایه سافت ماکس به همراه بردار احتمال واقعی است.

$$J^{sof} = -\frac{1}{N}\sum_{i=1}^{N}\mathbf{p}_i \ln\hat{\mathbf{p}}_i = -\frac{1}{N}\sum_{i=1}^{N}p_i[k]\ln\hat{p}_i[k] = -\frac{1}{N}\sum_{i=1}^{N}\ln\hat{p}_i[k] \qquad (۱۱)$$

در رابطه (۱۱) بردار p_i بردار احتمال واقعی نمونه ورودی iام بوده که تنها یک عنصر آن برابر یک بوده و بقیه عناصر آن صفر هستند. به‌منظور محاسبه بردار احتمال پیش‌بینی شده، فرض کنید که خروجی آخرین لایه قبل از لایه سافت ماکس به‌ازای یک نمونه ورودی به‌صورت بردار $[h_1, h_2, \cdots, h_m]$ است. در خروجی شبکه از تعداد K نود (در این مقاله K برابر هشت است) با تابع فعال‌ساز سافت‌ماکس استفاده شده است. خروجی هر نود مقدار احتمال یک کلاس را به‌صورت معادله (۱۲) محاسبه می‌کند:

$$\hat{p}_i = f(z_i) = \frac{\exp(z_i)}{\sum_i z_i} \qquad \forall i = 1,2,\ldots,K \qquad (۱۲)$$

که در رابطه (۱۲) مقدار z_i از رابطه (۱۳) محاسبه می‌شود.

$$z_i = \sum_{j=1}^{m} w_{i,j}h_j + b_i \qquad \forall i = 1,2,\ldots,K \qquad (۱۳)$$

در رابطه (۱۳) مقادیر w_{ij} و b_i به‌ترتیب وزن و بایاس قابل آموزش هستند. با توجه به رابطه (۱۲) به‌ازای هر نمونه ورودی یک بردار احتمال به‌صورت رابطه (۱۴) در خروجی به‌دست می‌آید.

$$\hat{\mathbf{p}} = [\hat{p}_1, \cdots, \hat{p}_k] \qquad (۱۴)$$

لذا تابع هزینه به‌صورت زیر ساده‌سازی می‌شود:

$$J^{sof} = -\frac{1}{N}\sum_{i=1}^{N} p_i[k]\ln\hat{p}_i[k] = -\frac{1}{N}\sum_{i=1}^{N}\ln\hat{p}_i[k] \qquad (۱۵)$$

در رابطه فوق مقدار $\hat{p}_i[k]$ مقدار پیش‌بینی شده برای کلاس kام به ازای ورودی iام است. $p_i[k]$ نیز مقدار احتمال واقعی نمونه iام است که مقدار آن برابر یک است. متغیر N نیز تعداد نمونه‌های آموزشی را نشان می‌دهد.

۳-۴- استخراج ویژگی‌ها

برای بررسی عملکرد مدل پیشنهادی، مجموعه نمونه‌های صوتی به‌صورت تصادفی به سه دسته آموزشی، اعتبارسنجی و آزمون تقسیم می‌شوند که ۸۰ درصد نمونه‌ها برای آموزش به‌کار می‌روند و ۱۰ درصد به‌عنوان داده‌های اعتبارسنجی و ۱۰ درصد بقیه به‌عنوان آزمون استفاده می‌شوند. به‌منظور افزایش تعداد نمونه‌های آموزشی برای جلوگیری از آموزش بیش از حد مدل و همچنین افزایش مقاومت مدل در برابر تغییرات محیطی، مقداری نویز سفید گوسی به نمونه‌های آموزشی اضافه می‌شود تا نمونه‌های جدیدی حاصل شود. در این مقاله به‌ازای هر یک نمونه آموزشی اصلی، دو نمونه نویزی جدید محاسبه می‌شود و در نتیجه تعداد کل نمونه‌های آموزشی سه برابر می‌شوند. در نتیجه تعداد نمونه‌های آموزشی برابر ۳۴۴۱، تعداد نمونه‌های اعتبارسنجی برابر ۴۲۹ و تعداد نمونه‌های آزمون برابر ۴۵۰ خواهد بود. شکل (۶) یک نمونه فایل صوتی اصلی را به همراه نمونه‌های نویزی شده را نشان می‌دهد. اضافه کردن نویز باعث می‌شود که کارآیی مدل در محیط‌های واقعی نیز تا حد ممکن حفظ شود. ویژگی‌های مورد استفاده برای آموزش مدل پیشنهادی از نوع ویژگی‌های MFCC هستند. برای استخراج این ویژگی‌ها هر فایل صوتی با استفاده از پنجره همینگ[۵۰] به طول ۵۱۲ نمونه قطعه‌بندی می‌شود. برای هر قطعه از گفتار تعداد ۴۰ ویژگی MFCC محاسبه می‌شود. با توجه به اینکه هر فایل صوتی در دادگان طولی برابر ۳ ثانیه دارد و نرخ نمونه‌برداری هم برابر ۴۸۰۰۰ نمونه در ثانیه است، لذا برای هر فایل صوتی یک ماتریس ویژگی از اندازه ۲۸۲×۴۰ به‌دست می‌آید. از آنجا که توزیع ویژگی‌ها نامشخص است لذا با استفاده از عمل نرمالیزه‌سازی، میانگین و انحراف از معیار آنها به مقدار صفر و یک تغییر داده می‌شود.

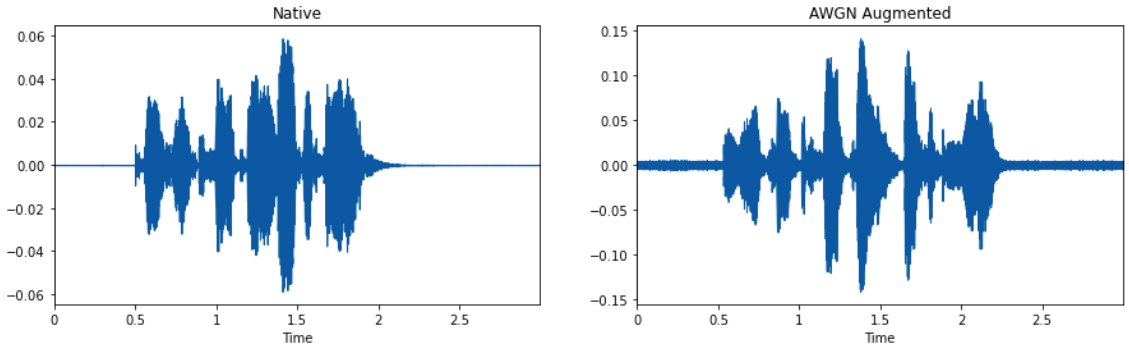
۴-۴- نتایج حاصل از شبیه‌سازی

در این بخش عملکرد چهار مدل مورد بررسی قرار گرفته است. مدل اول از نوع کانولوشنی است. مدل دوم از نوع ترنسفورمر است. این دو مدل به‌عنوان مدل پایه در نظر گرفته شده‌اند. مدل سوم نیز با موازی سازی دو مدل قبلی به‌دست آمده است. مدل چهارم از موازی سازی دو مدل پایه از نوع شبکه عصبی کانولوشنی تشکیل شده است. برای مدل ترنسفورمر مقدار h برابر ۴ انتخاب گردیده است. افزایش مقدار h باعث افزایش زمان آموزش می‌شود و تغییر چندانی در بازدهی شبکه در مدل پیشنهادی این مقاله ایجاد نمی‌کند. همچنین برای کاهش زمان آموزش شبکه ترنسفورمر، در ورودی این شبکه از یک لایه استخراج مقدار بیشینه استفاده گردیده است تا ابعاد ماتریس ویژگی ورودی را به اندازه ۷۰×۴۰ کاهش دهد. جهت بهینه‌سازی شبکه روش SGD با پارامترهای h برابر ۰/۰۰۱ و کاهش وزن برابر $1e^{-3}$ و مقدار حرکت برابر ۰/۸ به‌کارگیری شده است.

۱-۴-۴- نتایج حاصل برای مدل ترنسفورمر پایه

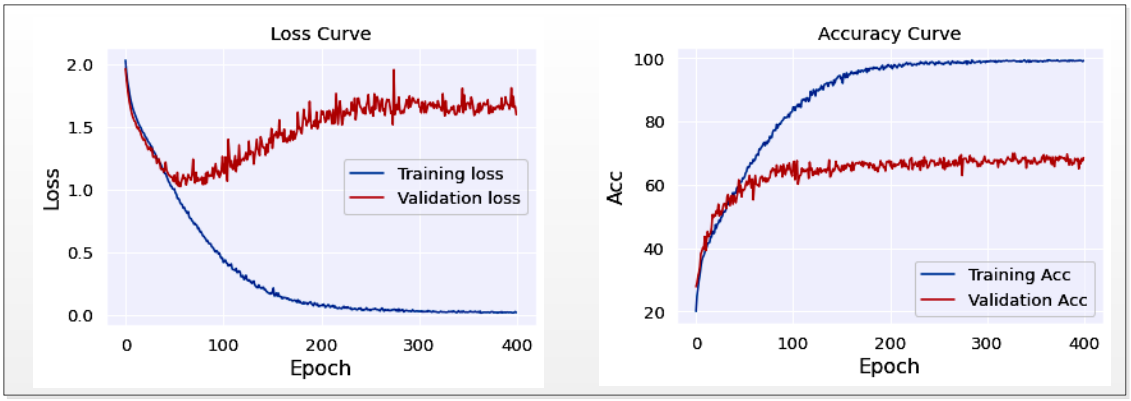
شکل (۷) منحنی کاهش خطا و دقت حاصل شده برای شبکه ترنسفورمر به‌ازای هر دوره آموزشی[۵۱] برای داده‌های اعتبارسنجی را نشان می‌دهد. شکل (۸) نیز ماتریس درهم‌ریختگی[۵۲] را برای داده‌های آزمون نشان می‌دهد. با توجه به این ماتریس، متوسط دقت حاصل برابر ۶۶ درصد به‌دست آمده است.

منحنی خطا در شکل (۷) نشان می‌دهد که مدل بعد از گذشت حدود ۲۰۰ دوره آموزشی دچار اشباع می‌شود که نشان دهنده آموزش کامل شبکه است. همچنین مشاهده می‌شود که خطای آموزشی برای داده‌های اعتبارسنجی بعد از گذشت حدود ۷۰ دوره افزایشی شده است. این بدان معنی است که سیستم قابلیت تعمیم‌پذیری برای داده‌های جدید را ندارد.



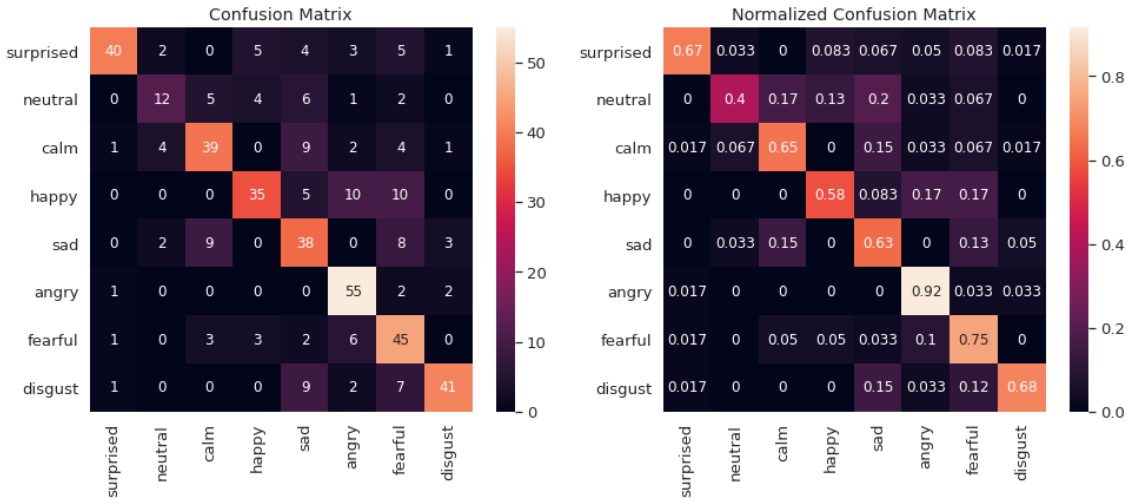
شکل (۶): فایل صوتی الف) سمت چپ فایل صوتی اصلی را نشان می‌دهد. ب) سمت راست فایل صوتی را بعد اضافه کردن نویز سفید گوسی نشان می‌دهد. بخش‌های بی‌صدا تفاوت دو فایل صوتی را از لحاظ وجود نویز نشان می‌دهد.

Figure (6): Audio file a)The left side shows the original audio file b) The right side Shows the audio file after adding Gaussian white noise. Silent sections show the difference between two audio files in terms of noise.



شکل (۷) : سمت چپ: منحنی کاهش خطا و سمت راست: منحنی دقت حاصل برای شبکه ترنسفورمر

Figure (7): Left: Error reduction curve and right: Accuracy curve for the transformer network.



شکل (۸) : ماتریس درهم‌ریختگی برای مدل ترنسفورمر

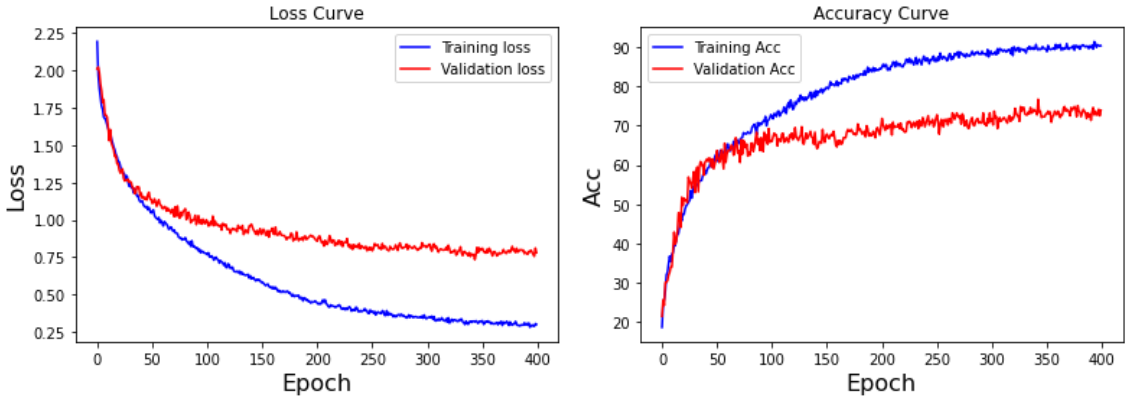
Figure (8): The confusion matrix for the transformer model.

دلیل این موضوع را می‌توان به اضافه شدن نویز به داده‌های آموزشی نسبت داد که در نتیجه آن سیستم ترنسفورمر قادر به استخراج ویژگی‌هایی که قابلیت تعمیم‌پذیری داشته باشند را ندارد. از موارد دیگر می‌توان به تعداد ویژگی‌های استخراجی در خروجی مدل ترنسفورمر دانست که با توجه به اندازه ماتریس ویژگی در ورودی سیستم، اندازه بردار ویژگی در خروجی سیستم

یاد شده برابر ۴۰ است. این تعداد ویژگی نتوانسته است مشخصات تمامی کلاس‌ها را در خود داشته باشد. همچنین تعداد نمونه‌های آموزشی کم و طول زمانی هر نمونه نیز تاثیر بسزایی در دقت چنین سیستم‌هایی دارد.
ماتریس درهم‌ریختگی مربوط به مدل ترنسفورمر نشان می‌دهد که بازشناسی احساسات گفتاری حالت عصبانی با دقت بالایی انجام گرفته است، در حالی‌که در مورد احساسات گفتاری حالت خنثی بدترین دقت را داشته است. لذا می‌توان گفت که این سیستم در بازشناسی احساساتی که در آن سطح انرژی تاثیر بیشتری دارد، موفق‌تر عمل می‌کند.

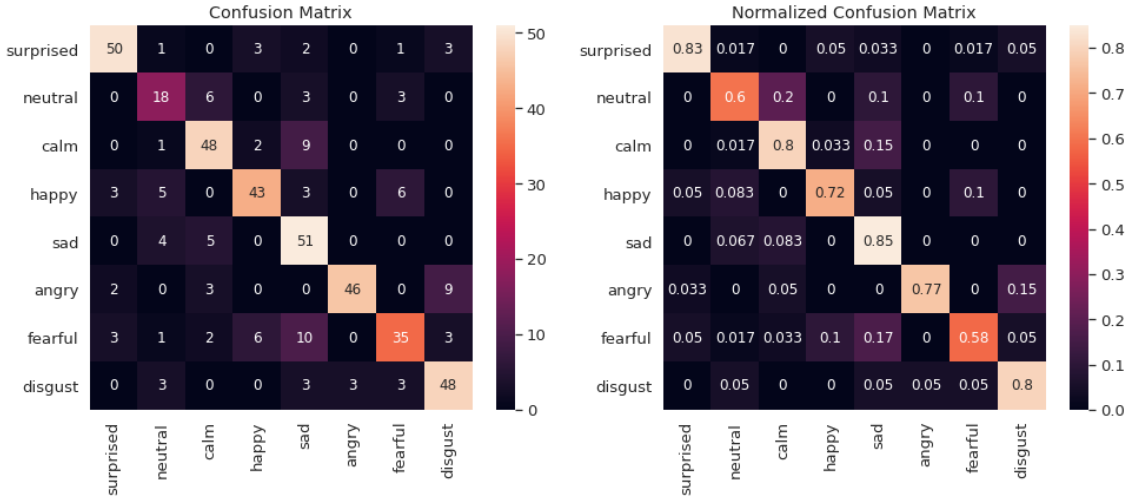
۲-۴-۴- نتایج حاصل برای مدل کانولوشنی پایه

شکل (۹) منحنی کاهش خطا و دقت حاصل شده برای شبکه کانولوشنی به‌ازای هر بازه آموزشی برای داده‌های اعتبارسنجی را نشان می‌دهد. شکل (۱۰) نیز ماتریس درهم‌ریختگی را برای داده‌های آزمون نشان می‌دهد. با توجه به این ماتریس، متوسط دقت حاصل برابر ۷۴ درصد است. برای این مدل کانولوشنی پایه، بعد از گذشت حدود ۳۵۰ دوره آموزشی سیستم توانسته است تا حدودی نسبت به نمونه‌های آموزشی و اعتبارسنجی به‌ترتیب به دقت حدود ۹۳ درصد و ۷۵ درصد برسد. در مقایسه با مدل ترنسفورمر پایه، این مدل در ارتباط با نمونه‌های آموزشی به دقت پایین‌تری رسیده است و در مورد نمونه‌های اعتبارسنجی به دقت بهتری دست یافته است.
همچنین مدت زمان آموزشی آن بیشتر طول کشیده است. با توجه به ماتریس درهم‌ریختگی دقت حاصل برای نمونه‌های آزمون نیز برابر ۷۲/۷ درصد حاصل شده است. همچنین این مدل در بازشناسی احساسات گفتاری حالت عصبانی و حالت ترس ضعیف‌تر از مدل ترنسفورمر عمل کرده است ولی در مورد حالت احساسی خنثی عملکرد به مراتب بهتری داشته است.



شکل (۹): سمت چپ: منحنی خطا و سمت راست: منحنی دقت حاصل برای شبکه کانولوشنی پایه

Figure (9): Left: Error reduction curve and right: Accuracy curve for the basic convulotional neural network.



شکل (۱۰): ماتریس درهم‌ریختگی برای مدل کانولوشنی پایه

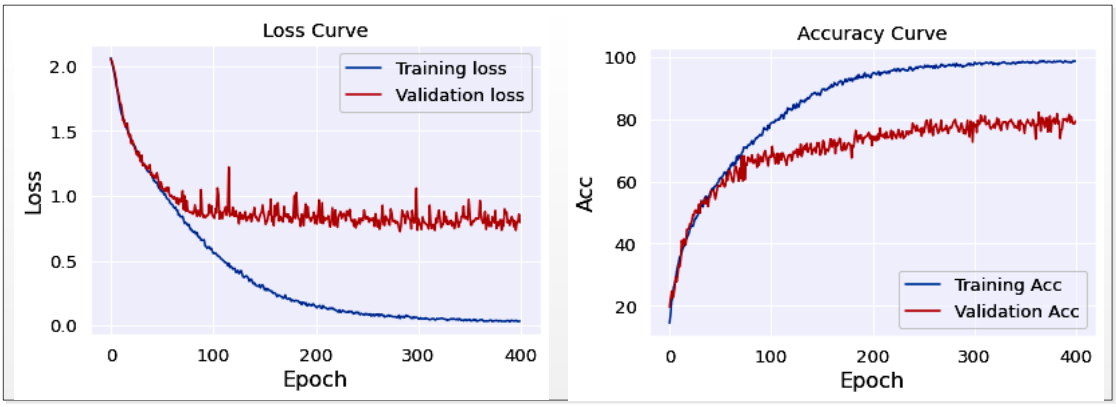
Figure (10): The confusion matrix for the basic convulotional neural network

۴-۴-۳- نتایج حاصل برای مدل ترکیبی ترنسفورمر پایه و کانولوشنی پایه

شکل (۱۱) منحنی کاهش خطا و منحنی دقت حاصل شده برای شبکه حاصل از موازی سازی شبکه‌های ترنسفورمر به همراه شبکه کانولوشنی به ازای هر دوره آموزشی برای داده‌های اعتبارسنجی را نشان می‌دهد. شکل (۱۲) نیز ماتریس درهم‌ریختگی را برای داده‌های آزمون نشان می‌دهد. با توجه به این ماتریس، متوسط دقت حاصل برابر ۷۵/۷ درصد است. مقایسه منحنی خطا و منحنی دقت مدل ترکیبی حاصل از موازی‌سازی مدل‌های ترنسفورمر پایه و کانولوشنی پایه با مدل‌های پایه نشان می‌دهد که این مدل عملکرد بهتری را از لحاظ زمان آموزشی دارد بدون اینکه مدل دچار آموزش بیش از حد شود. همچنین در مورد حالت‌های احساسی که هر دو مدل پایه نسبت به بازشناسی آنها عملکرد ضعیفی داشته‌اند، مدل ترکیبی توانسته این نقطه ضعف را تا حد زیادی برطرف نماید. دلیل این مساله را می‌توان در ترکیب ویژگی‌های زمانی و مکانی توسط مدل ترکیبی در بازشناسی احساسات در نظر گرفت.

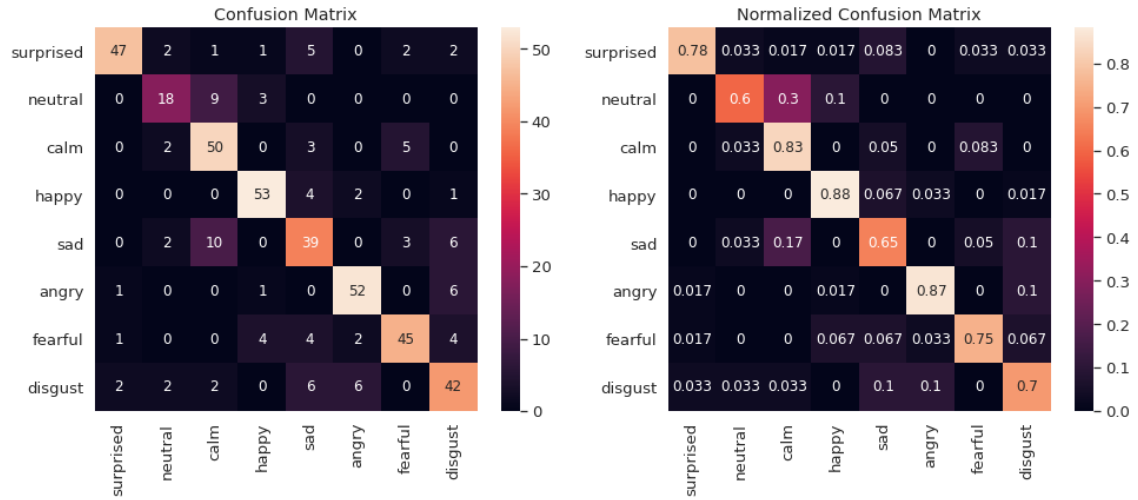
۴-۴-۴- نتایج حاصل برای مدل PCNN

شکل (۱۳) منحنی کاهش خطا و منحنی دقت حاصل شده برای شبکه حاصل از موازی‌سازی دو شبکه کانولوشنی به‌ازای هر دوره آموزشی برای داده‌های اعتبارسنجی را نشان می‌دهد. شکل (۱۴) نیز ماتریس درهم‌ریختگی را برای داده‌های آزمون نشان می‌دهد. با توجه به این ماتریس، متوسط دقت حاصل برابر ۷۹ درصد به‌دست آمده است.



شکل (۱۱): سمت چپ: منحنی خطا و سمت راست: منحنی دقت حاصل برای شبکه ترنسفورمر + کانولوشنی پایه

Figure (11): Left: Error reduction curve and right: Accuracy curve for the basic transformer network + convulotional neural network.



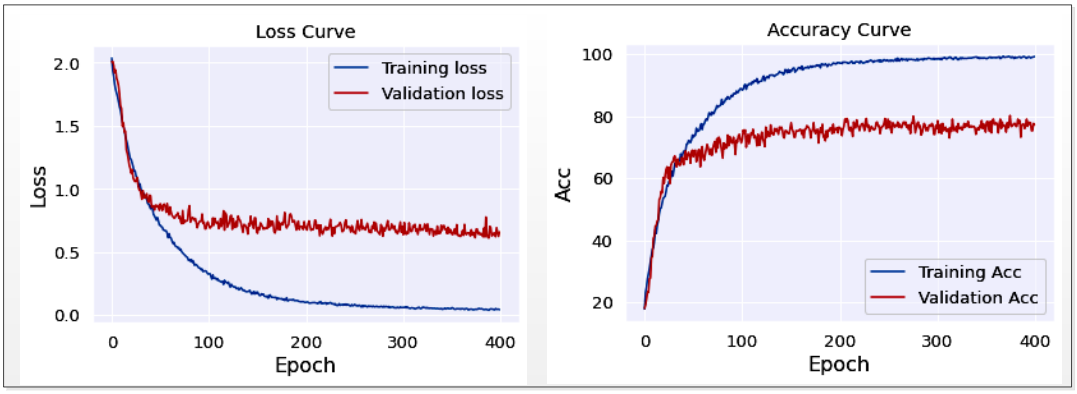
شکل (۱۲): ماتریس درهم‌ریختگی برای مدل ترنسفورمر+کانولوشنی

Figure (12): The confusion matrix for the basic transformer network + convulotional neural network.

توجه در منحنی خطا و منحنی دقت نشان می‌دهد که این مدل نسبت به مدل‌های قبلی هم از لحاظ مدت زمان آموزشی و هم از لحاظ دقت به‌دست آمده عملکرد بسیار بهتری را به‌دست آورده است. همچنین ماتریس درهم‌ریختگی نیز نشان می‌دهد که سیستم در مورد بازشناسی احساسات نسبت به سایر مدل‌ها، به دقت بالاتری دست یافته است. این مساله نشان می‌دهد که موازی‌سازی دو مدل پایه ساده با ساختار مشابه می‌تواند در یادگیری ویژگی‌هایی که بتوانند حالت‌های احساسی را نمایندگی کنند، بهتر عمل نماید و در نتیجه باعث افزایش دقت بازشناسی شود. همچنین عمل موازی‌سازی باعث می‌شود که پیاده‌سازی سخت‌افزاری و نرم‌افزاری چنین سیستم‌هایی در مقایسه با مدل‌هایی که از چندین لایه مختلف به تعداد زیاد در ساختار خود استفاده می‌کنند، بسیار راحت‌تر باشد و همچنین مدت زمان آموزش آنها بسیار کمتر باشد و تعداد نمونه‌های آموزشی مورد نیاز نیز کمتر شود. با توجه به اینکه دست‌رسی به تعداد نمونه‌های آموزش برچسب‌دار هزینه‌بر بوده و زمان زیادی لازم است تا جمع‌آوری گردند، لذا به‌کارگیری مدل‌های ساده به شکل موازی می‌تواند در حل این مشکل راه‌گشا باشد.

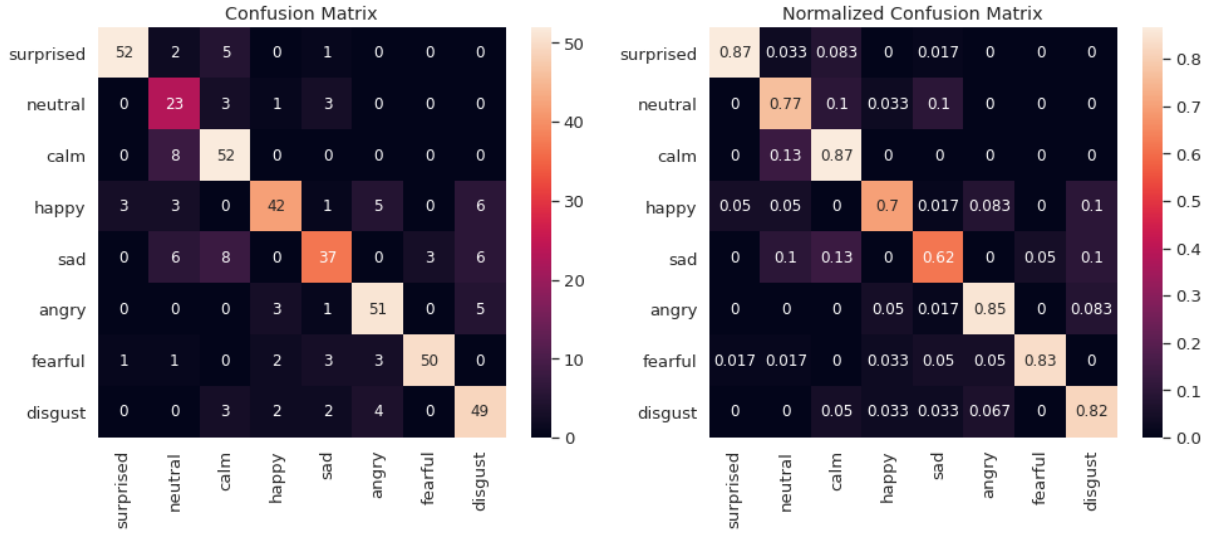
۵-۴-۴- نتایج حاصل برای مدل CTF

شکل (۱۵) منحنی کاهش خطا و منحنی دقت حاصل شده برای شبکه CTF به ازای هر دوره آموزشی برای داده‌های اعتبارسنجی را نشان می‌دهد. شکل (۱۶)شکل (۱۴ نیز ماتریس درهم‌ریختگی را برای داده‌های آزمون نشان می‌دهد. با توجه به این ماتریس، متوسط دقت حاصل برابر ۸۱/۷ درصد به‌دست آمده است.



شکل (۱۳): سمت چپ: منحنی خطا و سمت راست: منحنی دقت حاصل برای شبکه PCNN

Figure (13): Left: Error reduction curve and right: Accuracy curve for the PCNN network

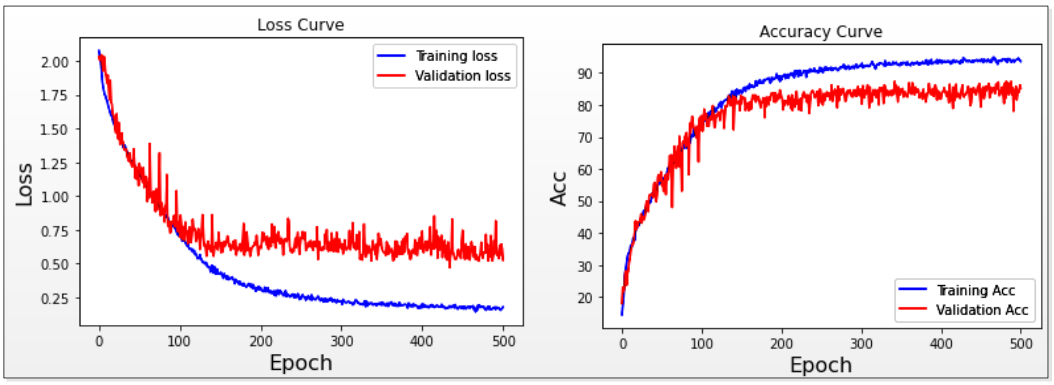


شکل (۱۴): ماتریس درهم‌ریختگی برای مدل PCNN

Figure (14): The confusion matrix for the PCNN model

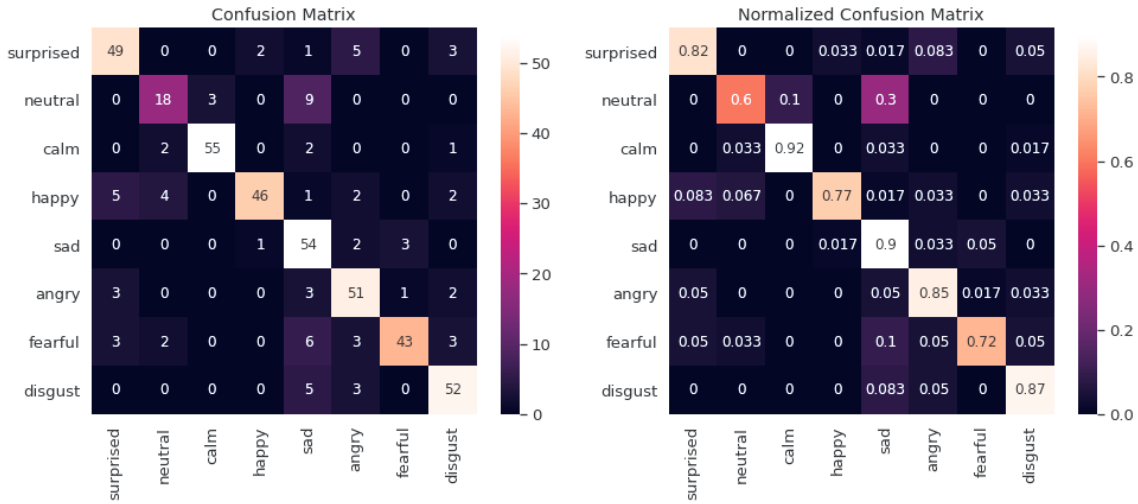
توجه در ماتریس درهم ریختگی مدل CNN-TF نشان می‌دهد که این مدل به جهت استخراج روابط مکانی-زمانی در تصویر MFCC ورودی، توانسته است به دقتی بالاتر از تمامی مدل‌های بیان شده در قبل برسد. مطالعات قبلی نیز نشان داده‌اند که مدل‌های ترکیبی از مدل کانولوشنی و مدل LSTM کارآیی نسبتا بهتری از دیگر مدل‌ها در بازشناسی احساسات از گفتار دارند. مهمترین مزیت مدل CTF نسبت به مدل CNN-LSTM این است که پیاده‌سازی مدل ترنسفورمر نسبت به مدل LSTM راحت‌تر بوده و همچنین مدل ترنسفورمر ارتباط تمامی فریم‌ها را در بازشناسی مورد سنجش قرار می‌دهد؛ در حالی‌که در مدل LSTM با افزایش فاصله بین فریم‌ها این سنجش ارتباط ضعیف‌تر می‌شود. جدول (۱) متوسط دقت به‌دست آمده برای نمونه‌های آزمون را برای پنج مدل تشریح شده در قبل نشان می‌دهد که با تکرار ۱۰ بار آزمایش به‌دست آمده است. در هر بار آزمایش نمونه‌ها به‌صورت تصادفی به مجموعه‌های آموزش، اعتبارسنجی و آزمون تقسیم شده‌اند. همچنین هیچ نمونه‌ای به‌صورت تکراری در مجموعه نمونه‌های دیگر آورده نشده است. با توجه به این جدول می‌توان نتیجه گرفت برای بازشناسی احساسات گفتاری از روی گفتار، در مواقعی که نمونه‌های برچسب‌دار کم باشند و هم چنین نمونه‌ها دارای نویز باشند، مدل ترکیبی از چندین مدل ساده می‌تواند به دقت بازشناسی بهتری دست یابد.

با توجه به این جدول مدل حاصل از موازی سازی دو مدل کانولوشنی به دقت حدود ۷۸ درصد دست یافته است که در مقایسه با مدل کانولوشنی و ترنسفورمر افزایش دقت به‌ترتیب در حدود ۵ درصد و ۱۴ درصد را برای نمونه‌های آزمون نشان می‌دهد. در مقایسه با مدل‌های پیشرفته‌تر، مدل پیشنهادی CTF به دقتی بالاتر رسیده است. برای مثال در مقایسه با روش مبتنی بر شبکه‌های عصبی کانولوشنی آورده شده در مرجع [۱۳]، افزایش دقتی در حدود ۹ درصد حاصل شده است.



شکل (۱۵): سمت چپ: منحنی خطا و سمت راست: منحنی دقت حاصل برای شبکه CTF

Figure (15): Left: Error reduction curve and right: Accuracy curve for the CTF network



شکل (۱۶): ماتریس درهم‌ریختگی برای مدل CTF

Figure (16): The confusion matrix for the CTF model

در مقایسه با روش مبتنی بر انتخاب ویژه‌گی [۱۶] که در آن در کنار شبکه عصبی کانولوشنی از یک ماشین بردار پشتیبان برای انتخاب ویژگی‌ها استفاده شده، روش CTF توانسته دقت را تا حدود ۴ درصد افزایش دهد. مقایسه این دو روش نشان می‌دهد که ترکیب شبکه‌های کانولوشنی و ترنسفورمر می‌تواند در استخراج ویژگی‌های مختص هر حالت احساسی کمک کننده باشد. مدل PCNN تنها در یک مورد در مقایسه با روش آورده شده در مرجع [۲۹] با نام CLSTM، کاهش دقتی در حدود دو درصد دارد. در مورد این تفاوت دقت می‌توان گفت که در مدل آورده شده در مرجع [۲۹] از نمونه‌های کاملا عاری از نویز در آموزش شبکه استفاده شده است که این یک نقطه ضعف برای این مدل است؛ زیرا که در محیط‌های واقعی وجود نویز پس زمینه در سیگنال گفتار معمول بوده و درنتیجه عملکرد آن را در کاربردهای واقعی به شدت تحت تاثیر قرار می‌دهد. همچنین در مدل CLSTM از تابع هزینه مرکزی بهره برده شده است. این تابع هزینه در مواردی که یک نمونه از یک کلاس خاص در فاصله بیشتری از نمونه‌های هم کلاسی خود قرار داشته باشد، باعث افزایش خطا خواهد شد و کارآیی خود را ازدست خواهد داد. زیرا که مرکز کلاس به طرف نمونه‌مرزی کشیده شده و درنتیجه مرزبندی بین کلاس‌ها دشوارتر می‌گردند. بنابراین در مواقعی که نمونه‌ها دارای نویز باشند، مدل پیشنهادی این مقاله از لحاظ کاربردی بودن نسبت به روش‌های قبلی مزیت بیشتری خواهد داشت. همچنین مدل پیشنهادی PCNN به جهت پیاده سازی از روش CLSTM ساده‌تر است؛ به این جهت که در ساخت آن تنها از بلوک‌های کانولوشنی استفاده شده است در حالی‌که در مدل CLSTM علاوه بر بلوک‌های کانولوشنی از بلوک‌های LSTM هم استفاده شده که پیاده‌سازی این بلوکها دشواری خاص خود را دارند. در جدول (۲) نتایج حاصل از به‌کارگیری دو نوع ویژگی با تعداد ویژگی‌های متفاوت در هر فریم آورده شده است. نتایج نشان می‌دهند که مدل‌های مبتنی بر شبکه‌های کانولوشنی و ترنسفورمر توانسته‌اند اطلاعات بیشتری را در ارتباط با احساسات از ویژگی‌های MFCC در مقایسه با ویژگی‌های مل‌اسپکتروگرام استخراج کنند. همچنین نتایج نشان می‌دهند که افزایش تعداد ویژگی در هرفریم تاثیر چندانی در افزایش دقت مدل کانولوشنی PCNN ندارد.

Table (1): Comparison of the average accuracy obtained for the four neural networks over ten trials

جدول (۱): مقایسه متوسط دقت حاصل برای چهار شبکه عصبی به ازای ۱۰ بار آموزش

مدل	دقت حاصل برای نمونه‌های آموزشی بر حسب درصد	دقت حاصل برای نمونه‌های اعتبارسنجی برحسب درصد	دقت حاصل برای نمونه‌های آزمون برحسب درصد
ترنسفورمر	۹۹/۱	۶۸/۷	۶۴/۲
کانولوشنی	۹۲/۷	۷۵/۵	۷۲/۷
کانولوشنی [۱۳]	--	--	۷۱/۶
کانولوشنی مبتنی بر انتخاب ویژگی [۱۶]	--	--	۷۷/۶
کانولوشنی+ LSTM [۱۷]	--	--	۷۷/۰۱
CLSTM [۲۹]	--	--	۷۹/۰۱
ماشین بردار پشتیبان [۳۳]	--	--	۷۵/۷۹
کانولوشنی [۲۷]	--	--	۶۴/۴۸
ترنسفورمر + مدل کانولوشنی	۹۸/۵	۷۷/۴	۷۳/۹
PCNN	۹۹/۷	۷۹/۴	۷۸/۱
CTF	۹۷/۶	۸۳/۳	**۸۰/۹۴**

Table (2): Results for the basic convolutional model and two hybrid models for two types of features with a different number of features per frame in percentage

جدول (۲): نتایج حاصل برای مدل کانولوشنی پایه و دو مدل ترکیبی به ازای دو نوع ویژگی با تعداد ویژگی‌های متفاوت در هر فریم برحسب درصد

نوع ویژگی	تعداد ویژگی در هر فریم	مدل		
		کانولوشنی پایه	PCNN	CTF
MFCC	۴۰	۷۲/۷	۷۸/۱	۸۰/۹
	۶۰	۷۳/۲	۷۸/۳	۸۰/۶
	۸۰	۶۹/۲	۷۷/۶	۸۱/۱
مل‌اسپکتروگرام	۴۰	۶۷/۵	۷۴/۸	۷۷/۲
	۶۰	۶۸/۵	۷۵/۹	۷۸/۷
	۸۰	۶۸/۳	۷۶/۲	۷۸/۸

۵- نتیجه‌گیری

در این مقاله به‌کارگیری مدل‌های ترنسفورمر و کانولوشنی در بازشناسی احساسات از روی گفتار مورد مطالعه قرار گرفت. نشان داده شد که استفاده از مدل موازی از مدل‌های پایه می‌تواند در بازشناسی احساسات گفتاری عملکرد بهتری داشته باشند. همچنین مدل‌های ترنسفورمر و کانولوشنی در بازشناسی یک سری از حالت‌های احساسی نسبت به هم متفاوت عمل می‌کنند و ترکیب این دو مدل می‌تواند نقاط ضعف آنها را تا حدودی برطرف سازد. افزایش تعداد مدل‌های پایه در حالت موازی و همچنین اضافه شدن نویزهای متفاوت به نمونه‌های آموزشی و بررسی تاثیر آنها در دقت بازشناسی احساسات گفتاری می‌تواند به عنوان یک مساله باز در تحقیقات بعدی مورد مطالعه قرار گیرد. همچنین بازشناسی احساسات از گفتار پیوسته که در آن احساسات متفاوت به‌صورت پشت سر هم بروز می‌کنند به‌عنوان یک چالش اساسی در مطالعات باقی مانده است.

سپاسگزاری

این مقاله از رساله دوره دکتری در دانشگاه آزاد اسلامی واحد علوم و تحقیقات استخراج شده است. نویسندگان بر خود لازم می‌دانند مراتب تشکر صمیمانه خود را از همکاران حوزه پژوهشی دانشگاه آزاد اسلامی و داوران محترم که ما را در انجام و ارتقای کیفی این مقاله یاری نموده‌اند، اعلام نمایند.

References

مراجع

[1] K. Han, D. Yu, I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine", Proceeding of the ISCA, pp. 223-227, Singapore, Malaysia, Sept. 2014 (doi: 10.21437/Interspeech.2014-57).

[2] A. M. Badshah, J. Ahmad, N. Rahim, S.W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network", Proceeding of the IEEE/PlatCon, pp. 1-5, Busan, South Korea, Feb. 2017 (doi: 10.1109/PlatCon.2017.7883728).

[3] S. Mittal, S. Agarwal, M.J. Nigam, "Real time multiple face recognition: A deep learning approach", Proceedings of the ICDMIP, pp. 70-76, Okinawa, Japan, Nov. 2018 (doi: 10.1145/3299852.3299853).

[4] H.S. Bae, H.J. Lee, S.G. Lee, "Voice recognition based on adaptive MFCC and deep learning", Proceeding of the IEEE/ICIEA, pp. 1542-1546, Hefei, China, June 2016 (doi:10.1109/ICIEA.2016.7603830).

[5] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", Proceedings of the IEEE/CVPR, pp. 770-778, Las Vegas, NV, USA, June 2016 (doi: 10.1109/CVPR.2016.90).

[6] K.Y. Huang, C.H. Wu, Q.B. Hong, M.H. Su, Y.H. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds", Proceeding of the IEEE/ICASSP, pp. 5866-5870, Brighton, UK, May 2019 (doi: 10.1109/ICASSP.2019.8682283).

[7] W. Lim, D. Jang, T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks", Proceeding of the IEEE/APSIPA, pp. 1-4, Jeju, Korea (South), Dec. 2016 (doi: 10.1109/APSIPA.2016.782-0699).

[8] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network", Proceeding of the IEEE/ICASSP, pp. 5200-5204, Shanghai, China, March 2016 (doi: 10.1109/ICASSP.2016.7472669)
[9] Y. Pourebrahim, F. Razzazi, H. Sameti, "Parallel shared hidden layers auto-encoder as a cross-corpus transfer learning approach for unsupervised persian speech emotion recognition", Signal Processing and Renewable Energy, 2021 (Accepted Manuscript).
[10] Y. Pourebrahim, F. Razzazi, H. Sameti, "Semi-supervised parallel shared encoders for speech emotion recognition", Digital Signal Processing, vol. 118, Article Number: 103205, Nov. 2021 (doi: 10.1016/j.dsp.2-021.103205).
[11] N. Yazdanian, H. Mahmodian, "Emotion recognition of speech signals based on filter methods", Journal of Intelligent Procedures in Electrical Technology, vol. 7, no. 27, pp. 3-12, Dec. 2016 (dor: 20.1001.1.2322-3871.1395.7.27.1.4).
[12] M. Kadkhodaei Elyaderani, S.H. Mahmoodian, G. Sheikhi, "Wavelet packet entropy in speaker-independent emotional state detection from speech signal", Journal of Intelligent Procedures in Electrical Technology, vol. 5, no. 20, pp. 67-74, March 2015 (dor: 20.1001.1.23223871.1393.5.20.6.1).
[13] D. Issa, M.F. Demirci, A. Yazici, "Speech emotion recognition with deep convolutional neural networks", Biomedical Signal Processing and Control, vol. 59, Article Number: 101894, May 2020 (doi: 10.1016/j.bspc.2020.101894).
[14] J. Zhao, X. Mao, L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks", Biomedical Signal Processing and Control, vol. 47, pp. 312-323, Jan. 2019 (doi: 10.1016/j.bspc.2018.08.03-5).
[15] S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition", Sensors, vol. 20, no. 1, Article Number: 183, Dec. 2020 (doi: 10.3390/s20010183).
[16] M. Farooq, F. Hussain, N.K. Baloch, F.R. Raja, H. Yu, Y.B. Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network", Sensors, vol. 20, no. 21, Article Number: 6008, Oct. 2020 (doi: 10.3390/s20216008).
[17] M. Sajjad, S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM", IEEE Access, vol. 8, pp. 79861-79875, April 2020 (doi: 10.1109/ACCESS.2020.2990405).
[18] M.S. Fahad, A. Ranjan, J. Yadav, A. Deepak, "A survey of speech emotion recognition in natural environment", Digital Signal Processing, Article Number: 102951, March 2020 (doi: 10.1016/j.dsp.2020.102951).
[19] A. Vaswani, N, Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need", Advances in Neural Information Processing Systems, pp. 5998-6008, Dec. 2017.
[20] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, "Image transformer", Proceeding of the PLMR, pp. 4055-4064, Stockholm, Sweden, July 2018.
[21] D. Povey, H. Hadian, P. Ghahremani, K. Li, S. Khudanpur, "A time-restricted self-attention layer for ASR", Proceeding of the IEEE/ICASSP, pp. 5874-5878:, Calgary, AB, Canada, April 2018 (doi: 10.1109/IC-ASSP.2018.8462497).
[22] P.J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, N. Shazeer, "Generating wikipedia by summarizing long sequences", arXiv preprint, pp. 1-18, Jan. 2018.
[23] C. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A.M. Dai, M.D. Hoffman, D. Eck, "Music transformer", arXiv preprint, 2018.
[24] P. Shegokar, P. Sircar, "Continuous wavelet transform based speech emotion recognition", Proceeding of the IEEE/ICSPCS, pp. 1-8, Surfers Paradise, QLD, Australia, Dec. 2016 (doi: 10.1109/ICSPCS.2016.7843-306).
[25] S.R. Livingstone, F.A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english", Plosone, vol. 13, no. 5, Article Number: 0196391, 2018 (doi: 10.1371/journal.pone.0196391).
[26] B. Zhang, E.M. Provost, G. Essl, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach", Proceeding of the IEEE/ICASSP, pp. 5805-5809, Shanghai, China, March 2016 (doi: 10.1109/ICASSP.2016.7472790).
[27] Y. Zeng, H. Mao, D. Peng, Z. Yi, "Spectrogram based multi-task audio classification", Multimedia Tools and Applications, vol. 78, no. 3, pp. 3705-3722, Feb. 2019 (doi: 10.1007/s11042-017-5539-3).
[28] A.S. Popova, A.G. Rassadin, A.A. Ponomarenko, "Emotion recognition in sound", Proceeding of the ICN pp. 117-124, Moscow, Russia, Oct. 2017 (doi: 10.1007/978-3-319-66604-4_18).
[29] S. Kwon, "CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network", Mathematics, vol. 8, no. 12, Article Number: 2133, Nov. 2020 (doi: 10.3390/math8122133).
[30] F. Chollet, "Deep learning with python", New York, NY: Manning Publications, 2017.

[31] M.S. Seyfioğlu, A.M. Özbayoğlu, S.Z. Gürbüz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities", IEEE Trans. on Aerospace and Electronic Systems, vol. 54, no. 4, pp. 1709-1723, Feb. 2018 (10.1109/TAES.2018.2799758).
[32] V. Verma, N. Agarwal, N. Khanna, "DCT-domain deep convolutional neural networks for multiple JPEG compression classification", Signal Processing: Image Communication, vol. 67, pp. 22-33, Sept. 2018 (doi: 10.1016/j.image.2018.04.014).
[33] A. Bhavan, P. Chauhan, R.R. Shah, "Bagged support vector machines for emotion recognition from speech", Knowledge-Based Systems, vol. 184, Article Number: 104886, Nov. 2019 (doi: 10.1016/j.knosys.2019.10-4886).

زیرنویس‌ها

1. Speech emotion recognition (SER)
2. Artificial intelligent (AI)
3. Deep neural network (DNN)
4. Support vector machine (SVM)
5. Convolutional neural network (CNN)
6. Recurrent neural network (RNN)
7. Transformer
8. Mel-frequency cepstral coefficients (MFCCs)
9. Attention
10. Long-short term memory (LSTM)
11. Spectrogram
12. Attention
13. One by one
14. Self-attention
15. Vanishing gradient
16. RAVDESS
17. 5-fold
18. Spectrogram
19. Gated residual networks (GResNets)
20. Visual Geometry Group (VGG)
21. Mel-spectrogram
22. Mel-scaled spectrogram (MSS)
23. Chromagram
24. Contrast
25. Tonnetz
26. Convolutional LSTM (CLSTM)
27. Spatiotemporal
28. Center loss (CL)
29. Softmax
30. Redial based (RB)
31. Discriminative
32. Salient
33. Bi-directional LSTM
34. Maxpooling
35. Minpooling
36. Average pooling
37. Fully connected network (FCN)
38. Maxpooling
39. Validation set
40. Dropout
41. Overfitting
42. Multi-head self-attention
43. Concatenate
44. Feed-forward network (FFN)
45. Parallel CNN (PCNN)
46. Dense
47. Relu

48. Convolutional-transformer (CTF)
49. Cross-entropy
50. Hamming
51. Epoch
52. Confusion