دانشگاه صنعتی سهند

DOR:
20.1001.1.23223146.1402.10.1.3.6

نشریه سامانه‌های غیرخطی در مهندسی برق

دوره ۱۰ - شماره ۱

بهار و تابستان ۱۴۰۲

صفحات ۶۰ الی ۸۳

ISSN: 2322-3146

http://journals.sut.ac.ir/jnsee

انتقال یادگیری از شبکه های عصبی کانولوشنال مبتنی بر رویکرد نوین تنظیم جزیی تکاملی برای طبقه بندی اصوات محیطی

حامد ریاضتی سرشت[۱] و کریم محمدی[۲]

[۱] دانشجوی دکتری- گروه مهندسی برق- دانشگاه علم و صنعت ایران- تهران- ایران

hamed_riazati@elec.iust.ac.ir

[۲] **نویسنده مسئول**- استاد- گروه مهندسی برق- دانشگاه علم و صنعت ایران- تهران- ایران

mohammadi@iust.ac.ir

چکیده

مسئله کمبود نمونه های آموزش یکی از چالش های اصلی در به کارگیری شبکه های عصبی کانولوشنال عمیق برای طبقه بندی اصوات محیطی[۲] است. یکی از رویکرد های مورد توجه برای مواجه با چالش مذکور، انتقال یادگیری[۳] است که در آن مدلی از پیش آموزش دیده به روی دادگانی با ابعاد بزرگ[۴]، به کاربرد هدف با اعمال تنظیمات جزیی[۵] تطبیق داده می شود. در این پژوهش، ما نشان می دهیم که در هر لایه ، همه نورون/کرنل ها تأثیر یکسانی در تشخیص نمونه های کلاس های مختلف ندارند، بلکه به ازای هر کلاس زیرگروهی خاص نقش اصلی و حیاتی را در طبقه بندی بازی می کند. از این رو و با توجه به وجود شباهت های زیاد بین برخی ازکلاس های مبدأ و هدف، پیشنهاد می کنیم که تمرکز تنظیمات جزیی در هر لایه تنها معطوف به زیرگروهی از نورون/کرنل ها شود که به شدت نیازمند تغییرات هستند و مسئول اصلی خطا در طبقه بندی نمونه های ورودی هستند، و باقی دست نخورده رها شوند. برای شناسایی زیرگروه های مذکور، یک مسئله یادگیری تو در تو طرح می کنیم و یک رویکرد تکاملی مؤثر برای حل آن پیشنهاد می کنیم. ارزیابی روش پیشنهادی بیانگر بهبود مطلق به اندازه ۱.۹٪ و ۲.۳٪ در دقت طبقه بندی[۶] به ترتیب به روی دادگان های ESC-50 و DCASE-17 نسبت به روش مرسوم انتقال یادگیری است؛ بهبودی که بدون اضافه کردن داده جدید و تنها با بهره برداری مؤثرتر از دانش موجود در شبکه از پیش آموزش دیده بدست آمده است. همچنین، افزایش زمان آموزش به ازای روش تکاملی پیشنهادی کم و در حدود یک سوم زمان لازم برای آموزش شبکه از ابتدا[۷] برآورد شده است.

واژه‌های کلیدی

یادگیری عمیق

شبکه های عصبی کانولوشنال

انتقال یادگیری

الگوریتم ژنتیک

طبقه بندی اصوات محیطی

Deep convolutional neural network
[2] Environmental sound classification
[3] Transfer learning
[4] Large-scale dataset
[5] Fine-tuning
[6] Classification accuracy
[7] Training from scratch

[Downloaded from journals.sut.ac.ir on 2024-11-23]

[DOR: 20.1001.1.23223146.1402.10.1.3.6]

Sahand University of Technology

DOR:
20.1001.1.23223146.1402.10.1.3.6

Journal of Nonlinear Systems in Electrical Engineering

Vol.10, No.1

Spring and Summer 2023

ISSN: 2322 – 3146

http://journals.sut.ac.ir/jnsee

Knowledge Transfer of Convolutional Neural Networks via a Novel Evolutionary Fine-tuning Strategy (EFS) for Environmental Sound Classification

Hamed Riazati Seresht[1] and Karim Mohammadi[2]

[1]PhD Candidate, Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran,

hamed_riazati@elec.iust.ac.ir

[2]**Corresponding Author**, Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran,

mohammadi@iust.ac.ir

ABSTRACT

Insufficient training data is one of the main challenges of utilizing deep Convolutional Neural Networks (CNNs) for Environmental Sound Classification (ESC). As a promising solution, Transfer Learning (TL) has addressed this issue by adapting a network pre-trained on a large-scale dataset to the target task. In this paper, we demonstrate that not all neurons/kernels of every layer in CNN networks are equally utilized to process the inputs of different classes, but there is a specific subgroups of neurons/kernels in every layer that play the key role in classification of every output class. Based on this observation and due to similarities that exist between feature spaces of some source and target classes, we propose to concentrate the fine-tuning process only on those neurons/kernels that do need changes and have the greatest impact on misclassifying target data. To identify these neurons/kernels, we pose a nested optimization problem for which we propose an effective evolutionary approach as solution. Compared to the conventional fine-tuning approach, our proposed method achieves absolute improvements of about 1.9% and 2.3% in accuracy on ESC-50 and DCASE-17, respectively; remarkable improvements produced not by adding augmented data but with a more efficient utilization of knowledge stored in the pre-trained network. It is noteworthy that the computation time overhead of the proposed evolutionary method is rather small (about one third of the time required to train the model from scratch).

Keywords

[Downloaded from journals.sut.ac.ir on 2023]

[DOR: 20.1001.1.23223146.1402.10.1.3.6]

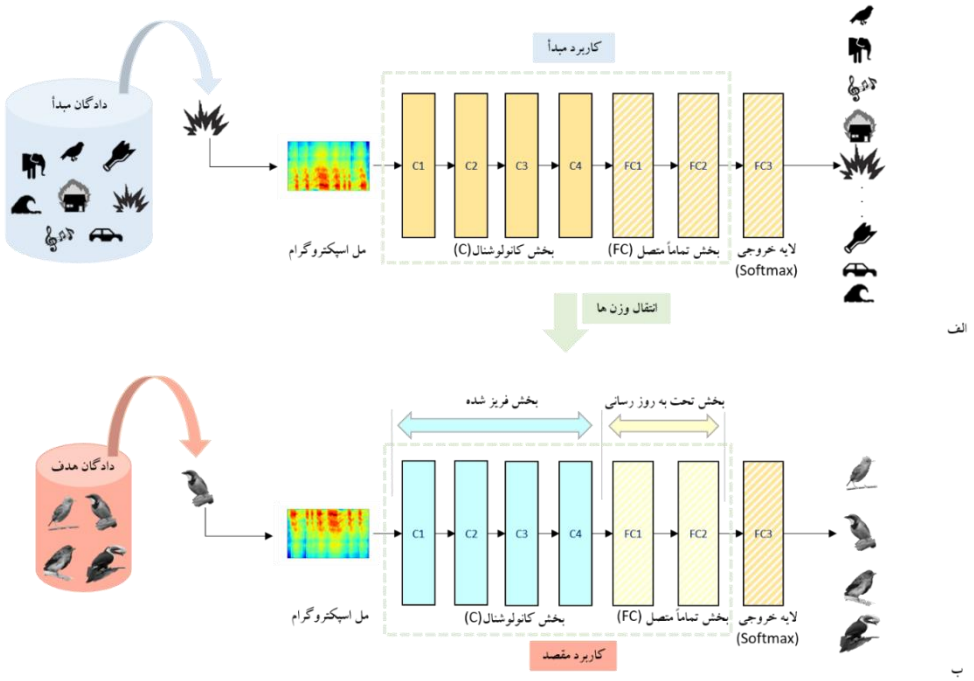
۱- مقدمه

توانمند سازی دستگاه های هوشمند در جهت ادراک محیط اطراف خود از طریق پردازش و طبقه بندی سیگنال های صوتی محیطی، در سال‌های اخیر مورد توجه فزاینده‌ای از سوی جامعه یادگیری ماشین قرار گرفته است. سیگنال های صوتی محیطی اطلاعات زیادی در مورد محیط اطراف در بردارند و با طبقه بندی آنها می توان به درک نوع محیط صوتی[۱] (همچون اتاق کار، ایستگاه قطار) و همچنین نوع وقایع صوتی[۲] شنیده شده در محیط اطراف (همچون شلیک گلوله، خودروی در حال عبور) پرداخت. توسعه روشهای پردازش سیگنال برای استخراج خودکار این اطلاعات کاربری های متنوعی دارند که به عنوان نمونه می توان به سمعک [1]، بازیابی اطلاعات [2]، سیستم های نظارت محیطی [3]، مسیر یابی ربات ها [4]، و شبکه های حسگر صوتی هوشمند [5] اشاره کرد. اما، طبقه بندی اصوات محیطی به دلیل ماهیت به شدت غیر ایستان[۳]، نسبت سیگنال به نویز پایین[۴]، و تنوع بالای مشخصه های زمان-فرکانس با چالش بسیار بیشتری نسبت به حوزه های دیگر مرتبط با صوت، همچون بازشناسی خودکار گفتار و طبقه بندی ژانر موسیقی، مواجه است.

شبکه های عصبی کانولوشنال (CNN[5]) به دلیل توانمندی بالا در یادگیری تابعی غیر خطی از فضای ورودی و همچنین استخراج ویژگی های سلسله مراتبی از الگوهای طیفی-زمانی موجود در بازنمایی مل اسپکتروگرام، به عنوان مناسب ترین گزینه برای مسئله طبقه بندی اصوات محیطی شناخته می شوند. روشهای ارائه شده بر پایه CNN ها [6, 7, 8] منجر به بهبود های قابل توجهی نسبت به روش های سنتی پایه شده اند. اما، میزان بهبودی حاصله از شبکه های CNN با عمق کم در این روشها در مقایسه با معماری های CNN عمیق تر همچون VGG [9]، DenseNet [10]، و ResNet [11] در حوزه بینایی ماشین[۶]، که در برخی موارد از توانایی بینایی انسان نیز پیشی گرفته اند، محدود بوده است. دست یابی به بهبود های بیشتر با استفاده از معماری های CNN عمیق تر به شدت وابسته به در دسترس بودن مقادیر زیادی از داده آموزشی است. با توجه به تعداد بسیار زیاد وزن های این معماری ها، آموزش شبکه با تعداد ناکافی از نمونه های آموزشی منجر به انطباق بیش از حد به داده آموزش[۷] و قابلیت تعمیم ضعیف در هنگام مواجه با نمونه جدید خواهد شد. در سال های اخیر، دادگان های جدیدی برای کاربرد طبقه بندی اصوات محیطی تهیه شده اند [12]. اما ابعاد آنها همچنان در مقایسه با دادگان های موجود در بینایی ماشین (به طور مثال در طبقه بندی تصاویر) بسیار کوچکتر است. به علاوه، تعداد نمونه های آموزش در دسترس به ازای برخی از کلاس های صوتی (همچون طبقه بندی چند گونه از پرندگان) ممکن است بسیار محدود باشد و جمع آوری داده بزرگ مورد نیاز می تواند بسیار پر هزینه و در برخی موارد غیر ممکن باشد.

[1] Sound scene
[2] Sound events
[3] Non-stationary
[4] Signal to Noise Ratio (SNR)
[5] Convolutional Neural Network
[6] Machine learning
[7] Overfitting

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]



شکل ۱- طرح کلی انتقال یادگیری برای یک شبکه CNN نمونه. الف) آموزش شبکه از ابتدا و به روی دادگان مبدأ، ب) انتقال وزن‌ها و تطبیق شبکه به دادگان مقصد. لایه های تماماً متصل (FC) با الگوی راه راه اریب از لایه های بخش کانولوشنال (C) تمییز داده شده است. همچنین، رنگ نارنجی، زرد و آبی به ترتیب بیانگر اعمال آموزش از ابتدا، اعمال تنظیمات جزیی، و فریز کردن وزن های لایه مربوطه است.

بر اساس مطالعات صورت گرفته، یکی از روش های مناسب برای مواجه با مسأله کمبود داده برچسب دار آموزش، انتقال یادگیری (TL[1]) است. برخلاف روش های سنتی که در آنها وزن های یک مدل مستقلاً و از ابتدا[2]آموزش داده می شوند (شکل ۱-الف)، در انتقال یادگیری، از دانشی که پس از آموزش شبکه به روی دادگانی جامع و بزرگ (به عنوان دادگان مبدأ) کسب و در وزن های شبکه ذخیره شده شده است، برای کاربرد هدف استفاده می شود (شکل ۱-ب). برای مثال، در [13] محققین یک شبکه CNN را ابتدا به روی دادگان بزرگ Audioset، که مشتمل بر اصوات متنوعی همچون گفتار انسان، صدای سوت/دست زدن، اصوات تولیده شده از انواع آلات موسیقی، اصوات حیوانات مختلف، باد، آب، آتش و ... است، آموزش داده و سپس آن رابه کاربرد هدف و دادگان کوچک ESC-50 [12] تطبیق داده اند. در رویکرد مرسوم برای تطبیق شبکه حاصله به روی دادگان هدف، لایه های شبکه به دو بخش لایه های عمومی و ویژه تقسیم می شوند. بر این اساس، در حالی که به تمام نورون/کرنل های لایه های بخش ویژه تنظیماتی جزیی (یعنی اعمال آموزش با نرخ یادگیری کوچک) اعمال می شود، تمام نورون/کرنل های لایه های بخش عمومی با تنظیم نرخ یادگیری در مقدار صفر به صورت دست نخورده رها می شوند[3] (شکل ۱-ب). این رویکرد برآمده از شواهد حاصل از آموزش CNN های مختلف است مبنی بر این که کرنل های لایه های اولیه گرایش به استخراج

[1]Transfer Learning (TL)
[2]Training from scratch
[3] Frozen

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

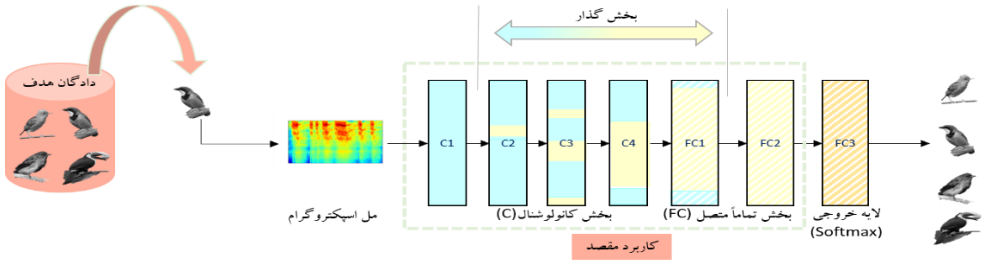
ویژگی هــای عمومی و مناسب برای کاربرد های مختلف دارند، در حالی که در لایه های بالاتر و نزدیک تر به خروجی شبکه، این عمومیت بدل به توانایی در استخراج ویژگی های ویژه و خاص برای یک کابرد به خصوص می شوند.

تطبیق شبکه به سیاق مذکور اگرچه منجر به دست یابی به دقت های بالاتر نسبت به آموزش شبکه از ابتدا شده است، اما دارای ایرادتی است. در پژوهش [14] نشان داده شده است که اعمال تنظیمات جزیی به برخی از لایه های شبکه در صورت کم بودن تعداد نمونه های دادگان هدف می تواند به طرز مخربی باعث دور شدن برخی نورون/کرنل ها از مقادیر اولیه گردد، که به معنای از دست رفتن بخشی از دانش ارزشمند اولیه موجود در شبکه است. از طرف دیگر، در [15] نشان داده شده است که وابستگی و سازگاری زیادی بین برخی از نورون/کرنل ها در لایه های همسایه در حین آموزش شبکه به روی دادگان مبدأ شکل می گیرد؛ پدیده مهمی که به نام همسازگاری[1] شناخته می شود. بنابراین، می توان نتیجه گرفت که گذار سریع از بخش عمومی (فریز شده) به بخش ویژه (در حال به روز رسانی) می تواند موجب از بین رفتن همسازگاری مذکور در صورت قرار گرفتن نورون/کرنل های همسازگار در دو سوی متفاوت شود، به خصوص وقتی تعداد نمونه های دادگان هدف کم است. در این مقاله، ما روش نوینی را به هدف ارتقاء بازدهی انتقال یادگیری تحت عنوان روش تنظیم جزیی تکاملی(EFS) [2] پیشنهاد می کنیم. برای نیل به این هدف، ما نشان می دهیم که همه نورون/کرنل های یک لایه نقش یکسان در تشخیص نمونه ها از کلاس های مختلف ندارند، بلکه به ازای هر کلاس زیرگروهی خاص نقش اصلی و حیاتی را در طبقه بندی بازی می کنند؛ موضوعی که تا جایی که ما می دانیم تاکنون مورد توجه قرار نگرفته است. از این رو، و با توجه به وجود شباهت های زیاد بین برخی ازکلاس های کاربرد مبدأ و هدف نتیجه می گیریم که شدت نیاز به اعمال تغییرات بین نورون/کرنل های هر لایه متفاوت است و چنان گذار سریعی در شبکه از بخش فریز شده به بخش تحت به روز رسانی در روال مرسوم بهینه نیست. با توجه به این مشاهده و همچنین برای رفع ایراداتی که در بالا ذکر شد (تخریب همسازگاری های موجود در شبکه و از دست رفتن دانش اولیه)، ما مطابق شکل ۲ پیشنهاد می کنیم که بخش سومی تحت عنوان بخش گذار[3] میان دو بخش فریز شده و تماماً تحت به روز رسانی تشکیل شود، که در هر لایه از آن تنها دسته ای از نورون/کرنل ها تحت به روز رسانی قرار می گیرند و باقی دست نخورده (فریز شده) رها می شوند. به بیان دیگر، گرادیان های به عقب انتشار داده شده خطا[4]در بخش گذار تنها به روی زیرگروهی از نورون/کرنل ها متمرکز می شود که بیشترین تأثیر را به روی عدم تشخیص صحیح نمونه ها دارند. برای تفکیک بهینه نورون/کرنل ها در بخش گذار به دو حالت فریز شده و تحت به روز رسانی، ما یک مسئله بهینه سازی تو در تو ایجاد می کنیم که (با توجه به چالش های مرتبط) به کمک الگوریتم ژنتیک به حل آن می پردازیم. نتایج آزمایشات ما به روی دادگان های در دسترس برتری رویکرد پیشنهادی را نسبت به روال مرسوم تأیید می کنند. شایان توجه است که بهبود نتایج بدست آمده نه از طریق اضافه کردن نمونه های مصنوعی برای آموزش و یا اضافه کردن بخش های پیچیده به معماری شبکه، بلکه تنها از طریق بهره برداری مؤثر تر از دانش موجود در شبکه پیش آموزش دیده صورت گرفته است.

[1] Co-adaptation
[2] Evolutionary Fine-tuning Strategy
[3] Transition
[4] Back-propagated gradients of errors

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

به طور خلاصه نوآوری های اصلی مقاله بدین شرح است: ۱) ما نشان می دهیم که نورون/کرنل های هر لایه در شبکه CNNنقش یکسانی در طبقه بندی کلاس های مختلف بازی نمی کنند، ۲) نشان می دهیم که با تشکیل بخش گذار پیشنهادی، که در هر لایه از آن تنها زیرگروهی بهینه از نورون/کرنل ها تحت به روز رسانی قرار می گیرند، می توان به دقت های طبقه بندی بالاتر دست یافت، ۳) روش نوینی برای ارتقاء عملکرد انتقال یادگیری بر پایه طرح یک مسئله بهینه سازی تو در تو ارائه می کنیم که برای حل آن از الگوریتم ژنتیک در کنار استفاده از گرادیان های خطا استفاده می کنیم. در ادامه مقاله، در بخش دوم به مرور مختصری بر کار های مرتبط می پردازیم. در بخش سوم جزییات روش پیشنهادی ارائه می شود. در بخش چهارم، نتایج بدست آمده از آزمایشات گزارش می شود و کارآیی روش پیشنهادی مورد ارزیابی قرار می گیرد. در نهایت، نتیجه گیری کلی از پژوهش صورت گرفته در بخش ششم ارائه می شود.



شکل ۲- طرح کلی ایده پیشنهادی مبنی بر تشکیل بخش گذار که در آن در هر لایه تنها دسته ای از نورون/کرنل ها تحت به روز رسانی قرار می گیرند (به رنگ زرد) و باقی دست نخورده (به رنگ آبی) رها می شوند.

نقش یکسانی در طبقه بندی کلاس های مختلف بازی نمی کنند، ۲) نشان می دهیم که با تشکیل بخش گذار پیشنهادی، که در هر لایه از آن تنها زیرگروهی بهینه از نورون/کرنل ها تحت به روز رسانی قرار می گیرند، می توان به دقت های طبقه بندی بالاتری دست یافت، ۳) روش نوینی برای ارتقاء عملکرد انتقال یادگیری بر پایه طرح یک مسئله بهینه سازی تو در تو ارائه می کنیم که برای حل آن از الگوریتم ژنتیک در کنار استفاده از گرادیان های خطا استفاده می کنیم. در ادامه مقاله، در بخش دوم به مرور مختصری بر کار های مرتبط می پردازیم. در بخش سوم جزییات روش پیشنهادی ارائه می شود. در بخش چهارم، نتایج بدست آمده از آزمایشات گزارش می شود و کارآیی روش پیشنهادی مورد ارزیابی قرار می گیرد. در نهایت، نتیجه گیری کلی از پژوهش صورت گرفته در بخش ششم ارائه می شود .

[Downloaded from journals.sut.ac.ir on 2024-11-23]

[DOR: 20.1001.1.23223146.1402.10.1.3.6]

۲- کارهای مرتبط

۲-۱- طبقه بندی اصوات محیطی

رویکرد های اولیه ارائه شده برای طبقه بندی اصوات محیطی، با الهام از روشهای مربوط به بازشناسی خودکار گفتار[۱]، الگوریتم های سنتی یادگیری ماشین همچون ماشین های بردار پشتیبانی[۲] و مدل های مخلوط گاوسی[۳] را برای طبقه بندی ویژگی‌هایی، همچون ضرایب کپسترال فرکانس مل[۴] و کدگذاری پیش بینی خطی[۵] [16, 17, 18, 19]، به کار برده اند. اخیراً و با شکوفایی چشمگیر یادگیری عمیق در بازشناسی گفتار [20]، شبکه های عصبی عمیق (DNN[6]) برای طبقه بندی اصوات محیطی توسط محققان بسیاری به خدمت گرفته شده اند. در [21]، Piczak برای اولین بار از CNN برای تحلیل ویژگی های مل اسپکتروگرام[۷] استفاده نمود و با تولید دقت های به مراتب بالاتری نسبت به روش های سنتی الهام بخش بسیاری از روش های بعدی شد. برخی از محققین تلاش نموده اند تا یادگیری مستقیماً از روی داده صوتی خام یک بعدی صورت پذیرد. از جمله، در [22] محققین معماری را ارائه نمودند که با استفاده از لایه های کانولوشنال یک بعدی در ابتدای شبکه، به یادگیری استخراج ویژگی های دو بعدی می پردازد که این ویژگی ها در ادامه شبکه و توسط لایه های ادغام[۸] و کانولوشنال دو بعدی طبقه بندی می شوند. در [23]، برای بهره بردن از مزایای ویژگی های دو بعدی مختلف، یک شبکه چند جریانی[۹] پیشنهاد شد که هر جریان یکی از ویژگی های ورودی را پردازش می کند.

گروهی دیگر ازمحققین تلاش نموده اند تا دقت طبقه بندی با ترکیب معماری های CNN و شبکه های عصبی بازگشتی[۱۰] ارتقا داده شود. در [7]، تلاش شده است تا با افزایش تمرکز به روی بخش های برجسته[۱۱] مل اسپکتروگرام ضمن محدود نگه داشتن بار محاسباتی و سایز شبکه به دقت های طبقه بندی بالا (در حد سیستم های سرآمد) دست یابند. همچنین، گروهی دیگر از محققین تلاش نموده اند تا با استفاده از ماژول های تمرکز[۱۲] محدوده های فرکانسی و فریم های زمانی برجسته را تشخیص و تمرکز را به روی آنها افزایش دهند [24, 25, 26, 27]. برای مواجه با چالش کمبود نمونه برچسب دار آموزش، یک راه حل ارائه تکنیک هایی برای تولید داده اضافی[۱۳] است. در [28, 29]، از طریق اعمال تغییر شکل هایی[۱۴]، همچون کشش زمانی[۱۵]، تغییر گام[۱۶]، افزودن نویززمینه[۱۷]

[1] Automatic speech recognition
[2] Support vector machines
[3] Gaussian mixture models Gaussian mixture models
[4] Mel-frequency cepstral coefficients
[5] Linear predictive coding
[6] Deep Neural Network (DNN)
[7] Mel Spectrogram
[8] Pooling
[9] Multi-stream
[10] Recurrent neural networks
[11] Salient
[12] Attention
[13] Data augmentation
[14] Deformations
[15] Time stretch
[16] Pitch shift
[17] Background noise

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

به سیگنال های صوتی، نمونه های مصنوعی تولید نمودند. در [30]، روشی بر پایه استفاده از شبکه های GAN[1] برای تولید نمونه های مصنوعی پیشنهاد شده است. انتقال یادگیری نیز به عنوان راه حلی امیدوار کننده مورد توجه در چند کار دیگر برای بهره بردن از شبکه های عمیق تر بوده است. در [31, 32, 32]، از معماری های معروفی همچون AlexNet، که از قبل به روی دادگان های بزرگ تصویر آموزش دیده بودند برای پردازش مل اسپکتروگرام استفاده شده است. با تهیه و پخش دادگان بزرگ و جامع صوتی [33]، معماری های متنوعی با عمق های بیشتر در [13, 34] به روی دادگان مذکور آموزش داده شده و بر پایه انتقال یادگیری برای کاربری اصوات محیطی به کار گرفته شدند.

۲-۲- الگوریتم های ژنتیک

الگوریتم ژنتیک (GA[2]) از معروف ترین روشهای حوزه الگوریتم های تکاملی محسوب می شوند. به خاطر ماهیت بدون گرادیان و عدم حساسیت به کمینه های محلی، الگوریتم ژنتیک به ویژه در مسائل بهینه سازی غیر محدب[3] و مشتق ناپذیر رایج هستند [35]. با الهام از تنازع بقا در طبیعت، الگوریتم ژنتیک تلاش می کند تا یک جمعیت اولیه از کروموزوم ها، که به عنوان جواب های کاندید یک مسئله در نظر گرفته می شوند، را به مرور تکامل داده به گونه ایکه کرومزوم هایی که بهترین خصوصیات را دارند به بقا خود ادامه دهند و سایرین حذف شوند تا در نهایت جواب های بهینه مسئله تولید شوند. به هر کروموزوم، که خود رشته ای از مقادیر پارامتر های مجهول مسئله هستند، یک مقدار برازش[4] اختصاص داده می شود که بیانگر میزان تناسب آن به عنوان جواب برای مسئله مورد نظر است. در هر نسل، برازش همه کروموزوم ها سنجش می شود و کاندید های آینده دارتر بر اساس نتایج آن و بر پایه یک روتین انتخاب برای تولید مثل به عنوان والد انتخاب می شوند. فرزند های نسل بعد معمولاً با اعمال عملگر های cross-over (که به عنوان جستجو محلی شناخته می شود) و جهش[5] (که به عنوان جستجو عمومی شناخته می شود) تولید می شوند.

۳- روش پیشنهادی

با در اختیار داشتن یک شبکه آموزش دیده به روی دادگانی بزرگ، رویکرد مرسوم برای تطبیق به دادگان هدف معمولاً مشتمل بر اعمال آموزش از ابتدا به لایه انتهایی همراه با تنظیم جزیی باقی لایه های شبکه می باشد (شکل ۱). در نتیجه مشاهده می شود که تمام نورون/کرنل ها در هر لایه به صورت یکسانی دستخوش تغییر می شوند. این در حالی است که مشاهدات ما نشان دهنده نقش متفاوت نورون/کرنل های هر لایه در تفکیک نمونه های کلاس های مختلف است. به بیان دیگر، معمولاً تنها بخشی از نورون/کرنل های یک لایه، به ویژه در لایه های نزدیک تر به خروجی، نقش اصلی در پردازش نمونه های یک کلاس خاص بازی می کنند. شکل ۳ هیستوگرام[6] مقادیر بیشینه تولید شده به تفکیک کلاس به ازای چند کرنل نمونه از یک شبکه از پیش آموزش

[1] Generative Adversarial Networks (GAN)
[2] Genetic Algorithm
[3] Non-convex
[4] Fitness
[5] Mutation
[6] Histogram

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

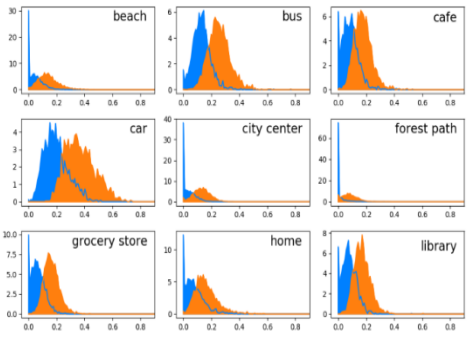
دیده را نمایش می دهد. در تحلیل این شکل باید در نظر داشت که که هر چه الگوی موجود در وزن های یک کرنل با فضای ویژگی[1] یک کلاس بیشتر تطابق و همبستگی[2] داشته باشد، صفحات ویژگی[3] تولید شده حاوی مقادیر بزرگتری خواهد بود. به بیان دیگر، بیشینه مقدار تولید شده توسط یک کرنل می تواند بیانگر میزان اهمیت کرنل مربوطه برای طبقه بندی نمونه جاری در نظر گرفته شود. در شکل ۳، با مقایسه هیستوگرام های بدست آمده به ازای کلاس های مختلف از هرکرنل پیش از تطبیق شبکه (نمودار های آبی رنگ) مشاهده می شود که توزیع هیستوگرام ها در گستره مقادیر متفاوتی قرار دارند. برای مثال، کرنل شماره ۱۰ به ازای اکثر نمونه های کلاس جاده جنگلی[4] مقادیری در حدود صفر تولید نموده، در حالی که به ازای کلاس خودرو بازه مقادیر تا مقدار ۰.۴ گسترش یافته است. همچنین، دیده می شود که رفتار کرنل ها به ازای نمونه های مربوط به یک کلاس یکسان به شدت متفاوت است. برای مثال، اگرچه غالب مقادیر تولید شده توسط کرنل های ۴۰ و ۹۵ ام مقادیری در حدود صفر به ازای نمونه های خودرو دارند، توزیع های بدست آمده از دو کرنل دیگر دارای میانگین های در حدود ۰.۲ الی ۰.۳ هستند.

با در نظر گرفتن تحلیل فوق، و با توجه به وجود شباهت بین برخی از کلاس های مبدأ و هدف، ما نتیجه می گیریم که در حالی که برخی از کرنل ها برای تطبیق به کلاس های دادگان هدف نیازمند تغییرات شدید هستند، برخی دیگر نیاز چندانی به اعمال تغییرات ندارند. البته باید توجه کرد که نیاز به اعمال تغییرات کمتر به صورت طبیعی در نتایج حاصله از کرنل نمونه ۲۲۳ ام با مقایسه توزیع های بدست آمده قبل و پس از اعمال تنظیمات جزیی به روال مرسوم به لایه های شبکه مشاهده می شود. اما، وقتی تعداد نمونه های آموزش کم باشد، این موضوع می تواند به دلیل گیر افتادن در کمینه های محلی به شکل بهینه صورت نپذیرد.

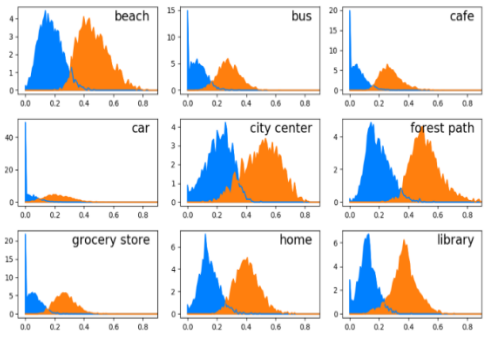
با انگیزه گرفتن از موارد فوق الذکر، در این مقاله بدنبال ارائه رویکردی خواهیم بود که در آن ارتباطات مهم موجود در بین نورون/کرنل های مختلف شبکه حفظ شوند و اعمال تغییرات به صورت هدفمند تنها معطوف به بخشی از نورون/کرنل ها شود که به شدت نیازمند تغییرات هستند و نقش کلیدی در عدم تشخیص صحیح نمونه های هدف بازی می کنند. برای نیل به این هدف، شبکه را به سه بخش فریز شده، تماماً تحت به روز رسانی، و بخشی تحت عنوان گذار[5] میان این دو که در آن تنها بخشی منتخب از نورون/کرنل ها به روز رسانی می شوند، تقسیم می کنیم. همانطور که در شکل ۲ دیده می شود، این تقسیم بندی تلاشی است در راستای پایه ریزی یک گذار تدریجی از بخش فریز شده به بخش در حال آموزش (برخلاف روال مرسوم که در آن به صورت آنی و بین دو لایه همسایه صورت می گیرد). در نتیجه این گذار تدریجی، از طرفی نمونه های دادگان هدف به دلیل کاهش تعداد پارامتر های تحت آموزش به طرز مؤثرتری برای آموزش به کار برده می شوند. همچنین، هم سازگازی های مذکور بین برخی از نورون/کرنل ها در لایه های همسایه حفظ می شوند چرا که یا هر دو با هم فریز می شوند و یا، در صورت نیاز، هر دو با هم به روز رسانی می شوند.

[1] Feature space
[2] Correlation
[3] Feature maps
[4] Forest path
[5] Transition

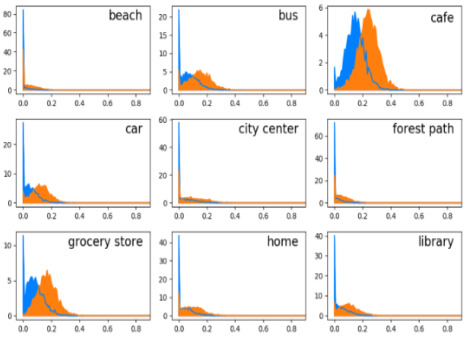
[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]



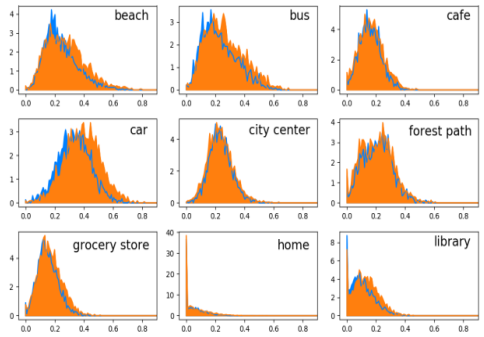
ب) کرنل شماره ۱۰



الف) کرنل شماره ۴۰



د) کرنل شماره ۹۵



ج) کرنل شماره ۲۲۳

شکل ۳- هیستوگرام های بدست آمده به ازای بیشینه مقادیر تولید شده توسط ۴ کرنل نمونه هم لایه از یک شبکه کانولوشنال به هنگام طبقه بندی نمونه هایی از ۹ کلاس مختلف. به ازای هر کلاس، دو هیستوگرام محاسبه شده است: یکی به ازای مقادیرتولید شده قبل از اعمال تنظیم جزیی (به رنگ آبی) و دیگری بعد از اعمال تنظیم جزیی (به رنگ نارنجی) به روش مرسوم. مشاهده میله ای در نزدیکی مقدار صفر در هر هیستوگرام به معنای عدم تطبیق کرنل به ازای نمونه های مربوطه و در نتیجه تولید مقادیری در حدود صفر است.

۱-۳ بیان ریاضی مسئله

اولین سوالی که به جهت تحقق ایده پیشنهادی به ذهن خطور می کند، مسئله نحوه نحوه شناسایی نورون/کرنل های نیازمند تغییرات در بخش گذار است. برای یافتن پاسخ ابتدا بیان ریاضی مسئله بهینه سازی خود را به ازای یک شبکه عصبی کانولوشنال به نام $\mathcal{A}$ که دارای L لایه است و لایه های آن به سه قسمت تماماً فریز شده (F[1])، بخش گذار (T[2])، و تماماً قابل به روز رسانی (U[3]) افراز شده است را به صورت زیر تعریف می کنیم:

$$,arg\min_{\lambda} \quad \mathcal{L}\left(\mathcal{A}_{\lambda}, \mathcal{X}^{train}, \mathcal{X}^{valid}\right) \qquad (۱)$$

[1] Frozen
[2] Transition
[3] Upgradable

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

$$\lambda = [F, T, U], \begin{cases} F = \{f_0, f_1, \dots, f_i\}, & f_l \in \{0\}^{K_l \times 1} \\ T = \{t_{i+1}, t_{i+2}, \dots, d_j\}, & t_l \in \{0,1\}^{K_l \times 1} \\ U = \{u_{j+1}, u_{j+2}, \dots, u_L\}, & u_l \in \{1\}^{K_l \times 1} \end{cases}$$

که در آن K تعداد کل نورون/کرنل های یک لایه، و f_l، t_l، و u_l به ترتیب بردار هایی تماماً صفر، مشتمل بر مقادیر صفر و یک، و تماماً یک هستند؛ به گونه ایکه مثلاً مقدار $d_l(m) = 0$ بیانگر عدم به روز رسانی وزن های نورون/کرنل m ام (w_m^l) از لایه l ام در بخش گذار است:

$$w_m^l(n+1) = w_m^l(n) + \left(\alpha \times \nabla w_m^l(n)\right) \quad , \; if \; d_l(m) = 1 \tag{۲-۱}$$

$$w_m^l(n+1) = w_m^l(n) \quad , \; if \; d_l(m) = 0 \tag{۲-۲}$$

همچنین، $\mathcal{A}_\lambda$ معرف شبکه $\mathcal{A}$ است وقتی تحت توصیف مجموعه λ به سه قسمت مورد نظر تقسیم بندی شده باشد، و $\mathcal{L}(.)$ مقدار خطای طبقه به روی مجموعه اعتبارسنجی $\mathcal{X}^{valid}$ را پس از آموزش $\mathcal{A}_\lambda$ به روی مجموعه $\mathcal{X}^{train}$ ارزیابی می کند. بنابراین، ملاحظه می شود که مسئله طرح شده در (۱) یک مسئله تو در تو است به گونه ایکه مسئله داخلی به بهینه سازی وزن های مشارکت کننده در آموزش که توسط حلقه بیرونی تعیین شده اند می پردازد، و مسئله بیرونی در تلاش برای یافتن مجموعه بهینه λ^* است که به ازای آن مقدار کمینه خطا $\mathcal{L}^*$ محاسبه می شود. از آنجا که $\mathcal{L}(.)$ نسبت به وزن های شبکه مشتق پذیر است، برای بهینه سازی حلقه داخلی می توان از الگوریتم های معمول بر پایه گرادیان همچون SGD[1] استفاده نمود. اما، بهینه سازی حلقه بیرونی از طریق مشابه امکان پذیر نیست چرا که پارامتر های بخش گذار ($T = \{t_{i+1}, t_{i+2}, \dots, t_j\}$) در مجموعه λ در محاسبات مسیر رو به جلو و تولید خروجی های شبکه نقش مستقیم نداشته و تنها در هنگام به روز رسانی وزن های شبکه (رابطه های (۲-۱) و (۲-۲)) تأثیر گذار هستند. به علاوه این که طبق تعریف هر یک از t_l ها آرایه ای از مقادیر گسسته صفر و یک در نظر گرفته شده است.

قبل از پرداختن به نحوه حل مسئله (۱)، حائز اهمیت است که به تمایز روش پیشنهادی با تکنیک Dropout [36] چه از حیث ایده اصلی و چه شرایط پیاده سازی توجه شود. ایده اصلی در تکنیک پرکاربرد Dropout، تخریب تصادفی برخی از همسازگاری های شکل گرفته در شبکه است. به بیان دیگر، هدف از این تکنیک اعمال فشار به نورون هایی است که به شدت وابسته به یک یا چند نورون دیگر آموزش دیده اند، در حالی که در صورت استقلال موقتی ممکن است این نورون ها بتوانند ظرفیت شبکه را در تفکیک نمونه های کلاس های مختلف افزایش بدهند. برای تحقق این هدف، دسته ای از نورون های لایه مورد نظر به صورت تصادفی در هر مرحله[2]آموزشی انتخاب می شوند، و هم از مشارکت در تولید مقادیر خروجی های نهایی شبکه و هم از آموزش (با صفر کردن مقدار تولید شده) به صورت موقتی محروم می شوند.در مقابل، ایده اصلی در روش پیشنهادی متمرکز کردن گرادیان های انتشار یافته تنها به روی یک دسته منتخب از نورون/کرنل ها است، به گونه ایکه اعضاء دسته مذکور دایمی بوده و در تمام ایپاک های فرآیند آموزش جایگزین نمی شوند. بنابراین، روش پیشنهادی از مشارکت تمام نورون/کرنل ها در مسیر رو به جلو استفاده نموده و مقادیر تولید شده تمام اجزاء شبکه در تولید خروجی های نهایی شبکه تأثیر خواهند داشت.

[1] Stochastic Gradient Descent
[2] Iteration

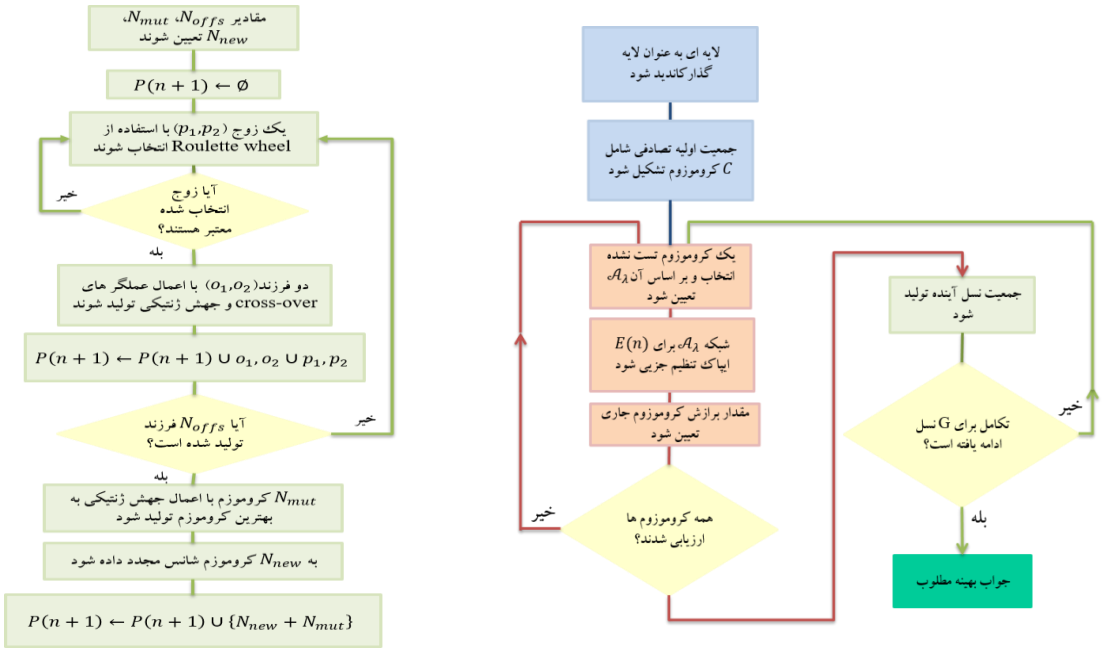
[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

۲-۳- بهینه سازی مسئله بیرونی برای شناسایی نورون/کرنل های نیازمند به روز رسانی در بخش گذار

مسئله طرح شده در رابطه (۱) شباهت زیادی به مسئله جستجو معماری عصبی (NAS[1]) [37] و یا تنظیم فراپارامتر های[2]شبکه دارد. هدف در مسئله بیرونی در NAS، بهینه سازی مجموعه λ مشتمل بر مواردی همچون تعداد لایه ها و یا نوع هر لایه است به انضمام مواردی که در بحث تنظیم فراپارامتر های شبکه (همچون نرخ یادگیری، تعداد کرنل های هر لایه و ...) وجود دارد. انواع روش های مختلفی برای بهینه سازی مسئله بیرونی درنشریات ارائـــه شده اندکه دستـــه ای ازآنها از روشهای تکاملی بهره برده اند [37, 38]. ما نیز در این پژوهش، الگوریتم تکاملی ژنتیک را برای حل مسئله خود و مواجه با چالش مذکور به کار می بریم. به جهت ساده سازی و کاهش دشواری مسئله، ما در این پژوهش محدودیتی را برای بخش گذار قائل می شویم و گستره آن را به تنها یک لایه کاهش می دهیم. بنابراین، با انتخاب M^l نورون/کرنل در لایه l ام (به ابعاد K^l) برای به روز رسانی، $\binom{K^l}{M^l}$ ترکیب مختلف فضای جستجوی مسئله را تشکیل می دهند؛ یک فضای همچنان بسیار بزرگ که به طور مثال به ازای $M = 256$ و $K = 512$ فضای جستجو از مرتبه 10^{152} خواهد بود. شکل ۴-الف نمای کلی رویکرد پیشنهادی تنظیم جزیی تکاملی (EFS[3]) را نمایش می دهد. مطابق شکل، روش پیشنهادی از سه مرحله تشکیل شده است. در مرحله اول، ابتدا لایه ای برای ایفا نقش به عنوان لایه گذار کاندید می شود. سپس، جمعیت اولیه ای ($P(0)$) از C کروموزم متفاوت، به صورت تصادفی تشکیل می شود. در مرحله دوم، همه کروموزوم ها به صورت مستقل ارزیابی می شوند. برای این منظور، مجموعه T به ازای هر کروموزوم و به تبع آن شبکه $\mathcal{A}_\lambda$ متناظر تشکیل می شود، و وزن هایش در مقادیر آموزش دیده به روی دادگان مبدأ مقدار دهی می شود. سپس، $\mathcal{A}_\lambda$ به اندازه $E(n)$ اپاک[4] به روی نمونه های $\mathcal{X}^{train}$ آموزش می بیند و دقت حاصل از آن به روی نمونه های $\mathcal{X}^{valid}$ به عنوان مقدار برازش کروموزوم متناظر در نظر گرفته می شود. در مرحله سوم، جمعیت نسل بعد ($P(n+1)$) مشتمل بر N_{offs} فرزند که از $N_{parents}$ والد منتخب بدست آمده اند، تولید می شود. مراحل دوم وسوم به اندازه G نسل ادامه داده می شوند و در نهایت کروموزوم دارای بهترین برازش به عنوان جواب مطلوب در نظر گرفته می شود. جزییات نحوه کدگذاری و ارزیابی هر کروموزوم و همچنین تولید نسل های بعد در بخش های ۳-۱ الی ۳-۳ ارائه می شوند.

[1] Neural Architectural Search
[2] Hyper-parameters
[3] Evolutionary Fine-tuning Strategy
[4] Epoch

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]



شکل ۴- الف) فلوچارت کلی روش پیشنهادی تنظیم جزیی تکاملی، ب) فلوچارت تولید جمعیت نسل آینده

۱-۲-۳- نحوه کدگذاری

نحوه کد گذاری باید به گونه ای باشد که هر کروموزوم غیر از مشخص نمودن یک ترکیب منحصر به فرد از وضعیت نورون/کرنل ها در لایه گذار، امکان اعمال عملگر های زاد و ولد را نیز فراهم نماید. از این رو، به عنوان لایه گذار برای لایه l ام، برداری به طول M^l مقدار صحیح که از بازه $[0, K^l)$ انتخاب می شوند را در نظر می گیریم که هر یک از مقادیر آن نماینده یک نورون/کرنل به خصوص است و به روز رسانی وزن ها در لایه گذار محدود به نورون/کرنل هایی می شود که اندیس های متناظرشان در کروموزوم مورد بررسی وجود دارد. برای مثال، در لایه گذاری با $K = 10$، تنها نورون/کرنل های زوج به ازای کروموزومی با بردار $[0,2,4,6,8]$ به روز رسانی می شوند.

۲-۲-۳- ارزیابی برازش

همانطور که گفته شد، برای ارزیابی هر کروموزوم می بایست شبکه مربوطه $\mathcal{A}_\lambda$ تشکیل و آموزش داده شود. اما، آموزش شبکه به ازای هر کروموزوم و تا زمانی که هیچ بهبود بیشتری حاصل نشود منابع و زمان قابل توجهی را به دلیل تعداد ایپاک های مورد نیاز (ده ها تا صدها، بسته به اندازه شبکه و مجموعه داده هدف) نیاز خواهد داشت. از این رو، ما ایده ارزیابی اولیه را پیشنهاد می‌دهیم که در آن حداکثر تعداد ایپاک برای ارزیابی هر کروموزوم از یک مقدار کوچک، مثلاً $E(0) = 4$، شروع می شود و با

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

گذر از نسلی به نسل بعد افزایش داده می شود. به این ترتیب، برازش هر کروموزوم با دقت طبقه بندی اولیه آن اندازه‌گیری می‌شود. این ایده بر پایه این فرض بنا نهاده شده است که کروموزوم هایی که دقت بهتری در ایپاک های اولیه ارائه می دهند با احتمال بالایی همچنان عملکرد بهتری در ایپاک های بعدی خواهند داشت.

۳-۲-۳- تولید نسل بعد

شکل ۴-ب نحوه تولید نسل بعد در روش پیشنهادی را به نمایش گذاشته است. قدم اول در تولید N_{offs} فرزند، انتخاب زوج والد از طریق روتین پرکاربرد Roulette wheel [39] است. برای ترکیب اطلاعات کروموزوم های زوج والد، ما ابتدا cross-over یکنواخت را به کار می بریم، که در آن هر مولفه از کروموزوم فرزند با احتمال مساوی از یکی از دو والد انتخاب می شود، و سپس با جایگزینی Q درصد از مولفه های کروموزوم فرزند با مقادیر صحیح تصادفی، کروموزوم های فرزند را دچار جهش ژنتیکی می نماییم. همچنین با الهام از طبیعت، از زوج شدن یک والد با فرزند خود، که در نسل های قبلی تولید نموده است، به دلیل مشابهت در نیمی از پارامتر ها جلوگیری می کنیم[1]. به طور مرسوم، مجموعه کروموزوم های نسل بعد متشکل از والدین انتخاب شده و فرزند هایشان است. اما، برای غنی سازی بیشتر جمعیت نسل بعد، ما دو راهکار کمکی نیز در نظر می گیریم: (۱) با اعمال جهش ژنتیکی به بهترین کروموزوم نسل جاری، N_{mut} کروموزوم جدید تولید می کنیم، (۲) با توجه به احتمال تولید دقت طبقه بندی بالاتر به ازای ایپاک های بیشتر، شانس مجددی برای N_{new} کروموزوم از نسل جاری که توسط Roulette wheel انتخاب نشده اند، برای حضور در نسل بعد در نظر می گیریم.

٤- آزمایش‌ها

در این بخش، به ارزیابی رویکرد پیشنهادی تنظیم جزیی تکاملی (EFS) و مقایسه آن با رویکرد مرسوم به عنوان روش مبنا (Baseline) بر اساس آزمایشات انجام شده می پردازیم. در ادامه، ابتدا دادگان ها و شبکه آموزش دیده مورد استفاده معرفی می شوند و سپس نتایج بدست آمده مورد بررسی و تحلیل قرار خواهند گرفت.

۱-٤- طراحی آزمایش ها

برای طبقه بندی اصوات محیطی، از دو دادگان، که به صورت عمومی در دسترس قرار دارند، استفاده می کنیم: دادگان ESC-50 [12] و DCASE-17 [40]. دادگان ESC-50 شامل ۲۰۰۰ فایل ضبط شده ۵ ثانیه است که از ۵ دسته عمده حیوانات (مثل پارس سگ)، اصوات مناظر صوتی طبیعی و صدای آب (مثل صدای بلبل و یا ریختن آب)، اصوات فضا های داخلی (مثل صدای تایپ کی بورد)، و اصوات مکان های خارجی (مثل صدای بوق خودرو) است. هر یک از این ۵ دسته دارای ۱۰ زیردسته متعادل هستند که در مجموع ۱۶۸ دقیقه داده را در بر دارند. دادگان DCASE-17 نیز به ابعاد کلی ۷۸۰ دقیقه و دارای ۱۵ کلاس

[1] اما منعی برای چند همسری قائل نمی شویم!

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

متفاوت از صحنه های صوتی همچون خانه، کتابخانه، ایستگاه قطار، ساحل است که هر کلاس دارای ۳۱۲ فایل ۱۰ ثانیه ای است. با توجه به سایز بزرگتر دادگان DCASE-17 و به جهت بررسی اثر سایز دادگان، ما نسخه دیگری از این دادگان و به ابعاد نصف نیز تهیه نموده ایم که برای این کار و برای حفظ تعادل بین کلاس ها، به سادگی از نمونه های هر کلاس تنها نمونه های با شماره زوج را جدا نموده ایم. معماری شبکه CNN مورد استفاده در این پژوهش (Vggish) که از قبل به روی فایل های صوتی دادگان بزرگ و جامع ویدیو YouTube-8M آموزش داده شده است [41]، مطابق با معماری ارائه شده در شکل ۲ و دارای ۳ لایه تماماً متصل و ۴ بلوک کانولوشنال است که در آن تعداد لایه های کانولوشنال در دو بلوک ابتدایی (C1,C2) و انتهایی (C3,C4) به ترتیب یک و دو لایه است. جزییات شبکه در جدول ۱ ارائه شده است.

جدول ۲- جزییات شبکه کانولوشنال که در آن B و C به ترتیب مشخص کننده تعداد نمونه های ورودی وکلاس های خروجی هستند.

لایه (ابعاد کرنل)@ تعداد		ابعاد خروجی
Conv1	(3×3) @ 64	B×96×64×64
Max pooling	(2×2)	B×48×32×64
Conv 2	(3×3) @ 128	B×48×32×128
Max pooling	(2×2)	B×24×16×128
Conv 3-1	(3×3) @ 256	B×24×16×256
Conv 3-2	(3×3) @ 256	B×24×16×256
Max pooling	(2×2)	B×12×8×256
Conv 4-1	(3×3) @ 512	B×12×8×512
Conv 4-2	(3×3) @ 512	B×12×8×512
Max pooling	(2×2)	B×6×4×512
Flatten		B×12288
Fc 1-1	@ 4096	B×4096
Fc 1-2	@ 4096	B×4096
Fc 2	@ C	B×C

[Downloaded from journals.sut.ac.ir on 2024-11-23]

[DOR: 20.1001.1.23223146.1402.10.1.3.6]

برای رعایت تطابق با ابعاد ورودی این شبکه (64×96)، هر فایل ضبط شده از دادگان های مذکور ابتدا به زیربخش های بدون همپوشانی به طول ۹۶۰ میلی ثانیه تجزیه می شود. سپس، نرخ نمونه برداری هر زیر بخش در مقدار 16 kHz تنظیم شده و با استفاده از تبدیل فوریه زمان کوتاه با سایز پنجره و شیفت ۲۵ و ۱۰ میلی ثانیه فریم بندی می شود. در نهایت، برای هر زیر بخش فریم بندی شده لگاریتم مل اسپکتروگرام با ۶۴ فیلتر مل و با گستره فرکانسی 0.125 KHz الی 7.5 KHz محاسبه می شود. در تمام آزمایشات، رویکرد پیشنهادی EFS بر اساس مقادیر پارامتر های گردآوری شده در جدول ۲ اجرا می شود. برای آموزش شبکه، چه تحت رویکرد پیشنهادی و چه روش مبنا، از بهینه سازی stochastic gradient descent و با نرخ یادگیری و سایز batch به ترتیب ۰.۱ و ۶۴ نمونه استفاده می شود. برای مقدار دهی تصادفی لایه آخر شبکه نیز از روش Xavier و برای محاسبه تابع هزینه نیز از cross-entropy استفاده می شود. برای جلوگیری از Overfitting نیز از معیار توقف اولیه با دوره ۱۰ اپیاک استفاده می شود. در همه موارد، هر یک از آزمایشات سه مرتبه تکرار و میانگین نتایج بدست آمده گزار ش می شوند. همچنین، همه مدل ها با استفاده از کتابخانه Tensorflow [42] و با استفاده از یک GPU (GTX 1060) پیاده سازی می شوند. در این مقاله مشابه روال مرسوم مقالات این حوزه [7, 21, 22] از دقت طبقه بندی به عنوان متر برای ارزیابی نتایج آزمایشات استفاده می کنیم:

$$Acc = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (۳)$$

۲-٤- نتایج آزمایش ها

در ابتدا، به میزان قابلیت انتقال دانش از لایه های مختلف شبکه می پردازیم. شکل ۵ نتایج بدست آمده را نمایش می دهد[۱]. در این آزمایشات، ما نسبت $M^l/_{Kl}$ را در لایه گذار کاندید شده را در مقدار ۰.۲۵ تنظیم نموده ایم. برای مثال، با انتخاب لایه پنجم (conv4-2) به عنوان لایه گذار، طول هر کروموزوم را به مقدار $M^5 = 0.25 * 512 = 128$ در نظر گرفته شده است. با توجه به نتایج، مشاهده می شودکه قابلیت انتقال یادگیری به ازای روش مبنا با در برگرفتن تمام لایه های تماماً متصل و همچنین چند لایه کانولوشنال به اوج مقدار خود می رسد و بعد از آن (conv4-2 در DCASE و conv3-2 در ESC-50) دچار افت می شود. با مقایسه نتایج حاصله به وضوح برتری رویکرد پیشنهادی EFS نسبت به روش مبنا به ازای تمام لایه های گذار کاندید مشاهده می شود. این برتری، به ازای استفاده از دادگان کوچک شده DCASE-17 نیز مشاهده می شود به گونه ای که نتایج روش پیشنهادی به ازای لایه conv4-1 حتی از نتایج روش مبنا که به ازای کل دادگان بدست آمده نیز بیشتر است. مطابق انتظار ملاحظه می شود که با افزایش تعداد لایه های حاضر در بخش تحت به روز رسانی، نتایج روش پیشنهادی به ازای عمق بیشتر و به مقداری بسیار کمتر از روش مبنا کاهش می یابند. امکان کسب دقت های بالاتر با به روز رسانی لایه های بیشتر به خوبی می تواند بیانگر موفقیت روش پیشنهادی در عدم تخریب کرنل های همسازگار و همچنین عدم از دست رفتن دانش اولیه به دلیل به روز رسانی برخی از کرنل های حساس باشد.

[۱] در شکل ۵، از عنوان لایه مرزی برای محور افقی استفاده شده است. باید توجه نمود که در هر لایه مرزی، همه نورون/کرنل ها به ازای روش مبنا، و زیرگروهی از آنها به ازای روش پیشنهادی به روز رسانی می شوند

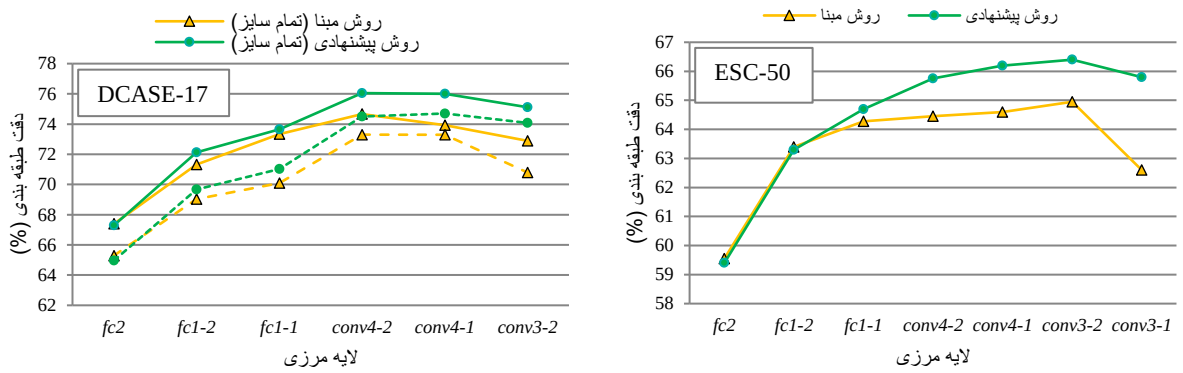
[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

جدول ۳- مقادیر استفاده شده پارامتر های رویکرد پیشنهادی EFS در آزمایشات.

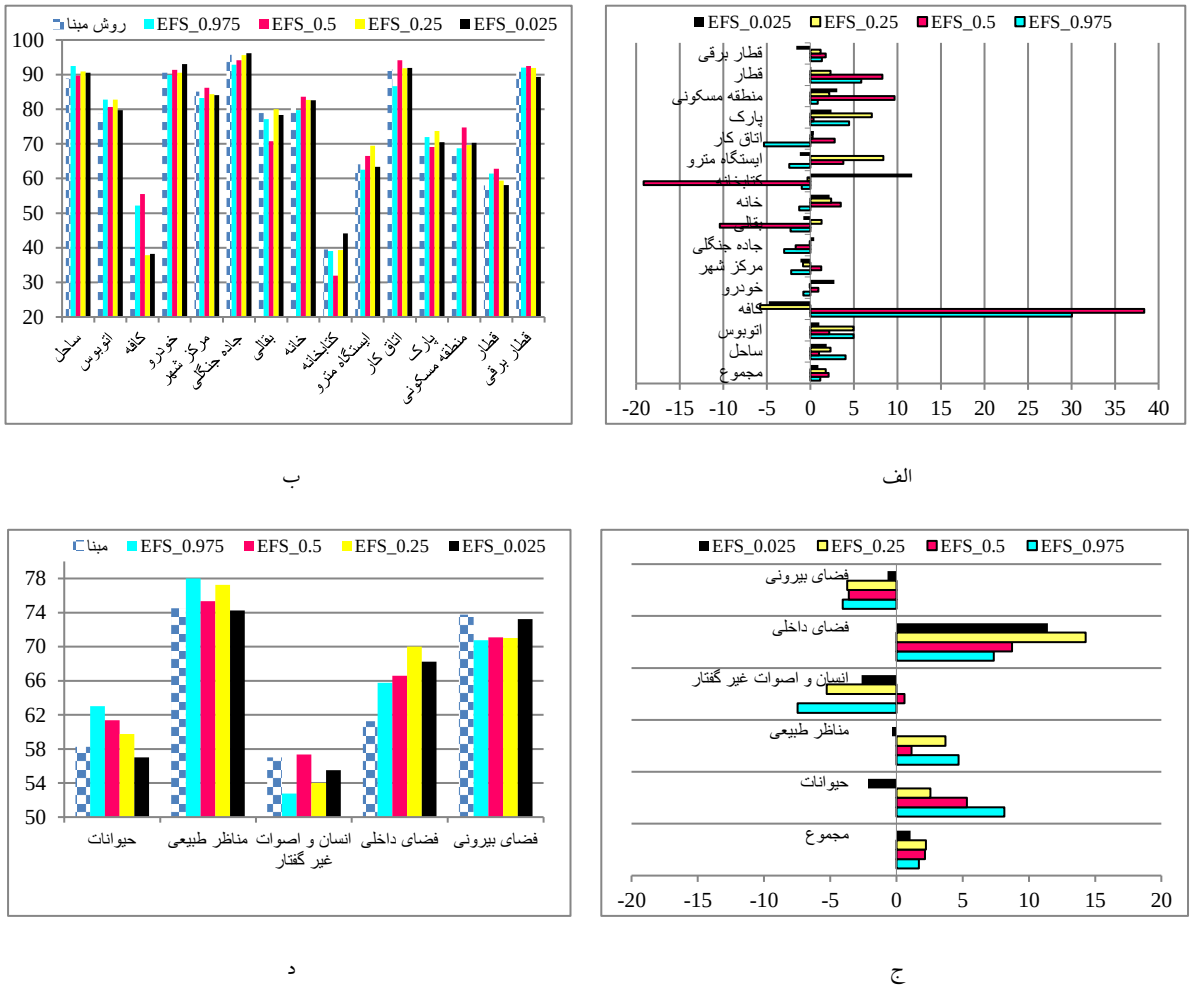
پارامتر	مقدار
G	۳
C	۱۶
Q	۰.۱
E	۴
N_{offs}	۴
N_{mut}	۴
N_{new}	۴

در شکل ۶، اثر طول کروموزوم ها یا به عبارت دیگر نسبت $M^l/_{K^l}$ را بر نتایج بدست آمده به تفکیک کلاس ها مورد تحقیق و بررسی قرار داده ایم. در این آزمایش ها لایه گذار را با توجه به بهترین نتایج بدست آمده در شکل ۵ در نظر گرفته ایم (یعنی لایه conv4-2 برای دادگان DCASE-17 و لایه conv3-2 برای دادگان ESC-50). مقدار $M^l/_{K^l} = 0.975$ تقریبی از روش مبنا، که همه نورون/کرنل های هر لایه را تحت آموزش قرار می دهد، را تقلید می کند. در حالی که، در طرف دیگر، مقدار $M^l/_{K^l} = 0.025$ تقریبی از حالت فریز کردن کامل لایه مورد بررسی در نظر گرفته شده است. با توجه به شکل ۶ دیده می شود که روش پیشنهادی تقریباً به ازای اغلب کلاس ها و مقادیر طول کروموزوم بهبودی محسوسی نسبت به روش مبنا تولید نموده است. بهبود نسبی بدست آمده به طرز قابل توجه ای به ازای برخی از کلاس ها زیاد است (مثلاً در حدود ۱۵٪ و ۴۰٪ به ترتیب به ازای کلاس های اصوات داخلی و کافه). تفاوت در بهبود ها (و بعضاً افت های) بدست آمده به ازای کلاس های مختلف تحت هر یک از نسبت های در نظر گرفته شده مجدداً مؤید ایده اصلی روش پیشنهادی است که تمام نورون/کرنل های یک لایه به طور مساوی برای پردازش نمونه های یک کلاس اهمیت ندارند. در واقع، چون میانگین نتایج بدست آمده به ازای همه کلاس ها به عنوان مقدار برازش هر کرموزوم کاندید در نظر گرفته شده است، بنابراین مجموعه λ^* تولید شده ممکن است در برگیرنده نورون/کرنل های حیاتی برای یک کلاس باشد در حالی که بخش مهمی برای کلاس دیگری را به صورت دست نخورده در نظر گرفته باشد. از این رو مشاهده می شود که به ازای هر یک از نسبت های $M^l/_{K^l}$، مجموعه λ^* تولید شده موجب بهبود دقت ها به ازای اغلب کلاس ها شده است در حالی که همان مجموعه حتی موجب افت به ازای یک یا چند کلاس دیگر شده است. مقادیر مرزی طول کروموزوم به ازای برخی از کلاس ها، همچون حیوانات و خودرو در دادگان ESC-50، جالب توجه است به گونه ایکه بهترین نتایج روش پیشنهادی تحت دو مقدار ۰.۰۲۵ و ۰.۹۷۵ بدست آمده است. این مشخص می کند که به روز رسانی و یا فریز کردن بخش کوچک ولی حیاتی از نورون/کرنل های یک لایه به جای تمام آنها نتایج بسیار بهتری به ازای برخی از کلاس ها تولید می کند. اما به صورت میانگین، نتایج حاصله حاکی از تولید بیشترین میزان پیشرفت به ازای دو مقدار ۰.۲۵ و ۰.۵ هستند.

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]



شکل ۵- میانگین دقت های طبقه بندی حاصل از تطبیق شبکه به روش مبنا و روش پیشنهادی به روی بخش توسعه[1] دادگان DCASE-17 (به ازای آموزش با تمام دادگان آموزش و نیمی از آن) ومیانگین دقت های حاصل از اعتبارسنجی ارائه شده در [12].



شکل ۶- دقت های طبقه بندی و میزان بهبودی نسبی تولید شده نسبت به روش مبنا به تفکیک کلاس ها و به ازای $\boldsymbol{M^l/K^l}$ مختلف و به ازای دو مجموعه داده DCASE-17 (الف و ب) و ESC-50 (ج و د).

[1] Development

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

conv4-2 برای دادگان DCASE-17 و لایه conv3-2 برای دادگان ESC-50). مقدار $M^l/_{K^l} = 0.975$ تقریبی از روش مبنا، که همه نورون/کرنل های هر لایه را تحت آموزش قرار می دهد، را تقلید می کند. در حالی که، در طرف دیگر، مقدار $M^l/_{K^l} = 0.025$ تقریبی از حالت فریز کردن کامل لایه مورد بررسی در نظر گرفته شده است. با توجه به شکل ۶ دیده می شود که روش پیشنهادی تقریباً به ازای اغلب کلاس ها و مقادیر طول کروموزوم بهبودی محسوسی نسبت به روش مبنا تولید نموده است. بهبود نسبی[1] بدست آمده به طرز قابل توجه ای به ازای برخی از کلاس ها زیاد است (مثلاً در حدود ۱۵٪ و ۴۰٪ به ترتیب به ازای کلاس های اصوات داخلی و کافه). تفاوت در بهبود ها (و بعضاً افت های) بدست آمده به ازای کلاس های مختلف تحت هر یک از نسبت های در نظر گرفته شده مجدداً مؤید ایده اصلی روش پیشنهادی است که تمام نورون/کرنل های یک لایه به طور مساوی برای پردازش نمونه های یک کلاس اهمیت ندارند. در واقع، چون میانگین نتایج بدست آمده به ازای همه کلاس ها به عنوان مقدار برازش هر کرموزوم کاندید در نظر گرفته شده است، بنابراین مجموعه λ^* تولید شده ممکن است در برگیرنده نورون/کرنل های حیاتی برای یک کلاس باشد در حالی که بخش مهمی برای کلاس دیگری را به صورت دست نخورده در نظر گرفته باشد. از این رو مشاهده می شود که به ازای هر یک از نسبت های $M^l/_{K^l}$، مجموعه λ^* تولید شده موجب بهبود دقت ها به ازای اغلب کلاس ها شده است در حالی که همان مجموعه حتی موجب افت به ازای یک یا چند کلاس دیگر شده است. مقادیر مرزی طول کروموزوم به ازای برخی از کلاس ها، همچون حیوانات و خودرو در دادگان ESC-50، جالب توجه است به گونه ایکه بهترین نتایج روش پیشنهادی تحت دو مقدار ۰.۰۲۵ و ۰.۹۷۵ بدست آمده است. این مشخص می کند که به روز رسانی و یا فریز کردن بخش کوچک ولی حیاتی از نورون/کرنل های یک لایه به جای تمام آنها نتایج بسیار بهتری به ازای برخی از کلاس ها تولید می کند. اما به صورت میانگین، نتایج حاصله حاکی از تولید بیشترین میزان پیشرفت به ازای دو مقدار ۰.۲۵ و ۰.۵ هستند.

در جدول ۳ به مقایسه نتایج حاصل از تطبیق شبکه با تصمیم گیری در سطح کل هر فایل ضبط شده پرداخته ایم. برای این منظور، کلاس پیش بینی شده توسط شبکه را بر اساس میانگین خروجی های شبکه به ازای همه زیربخش های هر فایل ضبط شده محاسبه نموده ایم[2]. همچنین، در این جدول نتایج حاصل از چند مدل منتقل شده دیگر به حوزه اصوات محیطی را نیز ارائه نموده ایم. همانطور که دیده می شود روش پیشنهادی موفق شده است که روش مرسوم انتقال یادگیری (روش مبنا) را به اندازه ۱.۸۶٪ و ۲.۲۹٪ به ترتیب به روی دادگان های ESC-50 و DCASE-17 بهبود دهد.

[1] Relative improvement = $\frac{Absolute\ improvement}{accuracy\ of\ baseline\ method} \times 100$

[2] با یادآوری این نکته که هر فایل ضبط شده ESC-50 و DCASE-17 به ترتیب به ۵ و ۱۰ زیر بخش افراز شده اند، لازم به ذکر است که نتایج ارائه شده در شکل های ۶ و ۷ حاصل تصمیم گیری در سطح زیر بخش های یک ثانیه ای بوده اند.

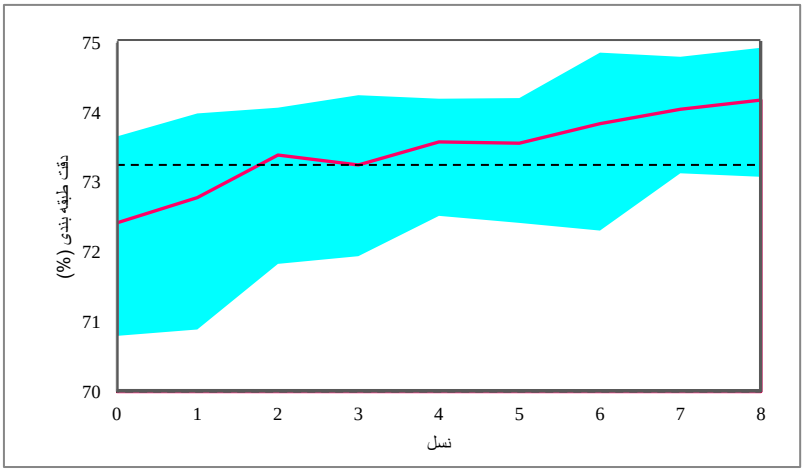
[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

جدول ۴- مقایسه دقت های بدست آمده از تطبیق شبکه Vggish به روش مبنا و پیشنهادی EFS به ازای تصمیم گیری درسطح هرفایل، ومقایسه با مدل های دیگر انتقال یافته به حوزه طبقه بندی اصوات محیطی. برای DCASE-17 نتایج به روی دادگان Evaluation، و برای ESC-50 نتایج به ازای میانگین دقت های حاصل از cross-validation گزارش شده اند.

Model	ESC-50	DCASE-17
PiczakCNN [21]	64.90	NA
SoundNet [43]	74.20	NA
AlexNet [31]	78.70	62.00
GoogleNet [31]	67.80	64.00
Vggish-Baseline	81.92	65.16
Vggish-EFS (ours)	83.78	67.45

NA: Not Available

شکل ۷ نحوه تکامل جمعیت در گذر نسل ها و میزان دقت های کمینه، متوسط و بیشینه کروموزوم های هر نسل را نمایش می دهد. با توجه به شکل دیده می شود که تنها سه نسل کافی است تا میانگین نتایج کروموزوم ها از دقت روش مبنا پیشی بگیرد. هزینه پرداخت شده در ازای بهبود های حاصله از روش پیشنهادی، افزایش زمان مورد نیاز برای تطبیق شبکه به دادگان هدف نسبت به روش مبنا است. افزایش تعداد ایپاک های مورد نیاز روش پیشنهادی نسبت به روش مبنا برابر با $C(n) \times \sum_{n=0}^{G-1} E(n)$ است، که در آن $C(n)$، G، و $E(n)$ به ترتیب تعداد کروموزوم های هر نسل، تعداد کل نسل ها، و تعداد ایپاک های آموزش در هر نسل هستند. مشاهده می شود که به ازای مقادیر استفاده شده در آزمایشات این مقاله[1]، روش پیشنهادی نسبت به روش مبنا به ۲۴۰ ایپاک بیشتر نیاز دارد. این تعداد ایپاک به ازای GPU مورد استفاده در آزمایشات این پژوهش، معادل صرف تنها ۳ ساعت زمان بیشتر است، که به عنوان زمان قابل قبولی می تواند در نظر گرفته شود.



شکل ۷- سیر تکاملی روش پیشنهادی EFS به هنگام تطبیق مدل به روی دادگان آموزش (توسعه) DCASE-17 (نیمه سایز) و به ازای یک جمعیت اولیه نمونه.

[1] $C = [16,16,16]$, $G = 3$, $E = [4,5,6]$

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

۵- نتیجه‌گیری

در این مقاله، ما یک رویکرد نوین برای انتقال یادگیری ارائه کردیم. در این راستا، به نقایصی از جمله نیاز متفاوت نورون/کرنل ها برای تغییرات به دلیل تفاوت در میزان شباهت بین کلاس های دادگان مبدأ و هدف، از دست رفتن دانش به دلیل فاصله گرفتن مخرب وزن های تحت تنظیم جزیی از مقادیر اولیه شبکه مبدأ، و تخریب همسازگاری موجود بین برخی از نورون/کرنل های لایه های همسایه به دلیل فریز کردن یک لایه و دستخوش تغییرات شدن لایه دیگر پرداختیم. ما پیشنهاد کردیم که لایه های شبکه به سه بخش کاملاً فریز شده، کاملاً تحت تنظیم جزیی و بخش گذار میان این دو افراز شوند، که در بخش گذار زیرگروهی از نورون/کرنل های آن فریز و بخش دیگر دستخوش تنظیم جزیی می شوند. برای پاسخ به مسئه طرح شده در خصوص تفکیک نورون/کرنل ها در بخش گذار از الگوریتم ژنتیک سود بردیم. نتایج آزمایش های انجام شده نشان دادند که حتی با محدود کردن بخش گذار به تنها یک لایه نیز می توان بهبود های قابل توجه ای بدست آورد. بهبود های حاصله از روش پیشنهادی به خوبی نمایانگر اهمیت تنظیم جزیی به صورت منعطف و متناسب با نیاز شبکه با توجه به فضای ویژگی دادگان هدف و میزان تفاوت آن با دادگان مبدأ است. نتایج بدست آمده به ازای تفکیک کلاس ها نشان دادند که بیشترین بهبودی حاصله در کلاس های مختلف به ازای طول یکسانی از کروموزوم ها تولید نشده است. به بیان دیگر، ممکن است به روز رسانی بخش بزرگی از یک لایه موجب بهبود نتایج یک کلاس شود، در حالی که به روز رسانی این بخش باعث افت دقت در کلاس دیگر شود. مشاهده ای که تأیید می کند که به دلیل فاصله های متفاوت کلاس های مبدأ و هدف از هم، نیاز ها به ازای کلاس های مختلف متفاوت است و روش پیشنهادی با برقراری موازنه ای بین آنها تلاش نموده است تا به بیشترین دقت میانگین نائل شود.

بهبودی حاصله از روش پیشنهادی تنها با بهره برداری مؤثرتر از دانش موجود در شبکه از پیش آموزش دیده بدست آمده است. باید توجه داشته که افزایش زمان لازم برای تطبیق، هزینه پرداخت شده به ازای بهبود های حاصله از روش پیشنهادی بوده است. اما این هزینه از آنجا که تعداد اپاک افزایش یافته در حدود یک سوم تعداد اپاک های لازم برای آموزش شبکه از ابتدا، و یا معادل صرف ۳ ساعت زمان بیشتر بر اساس سخت افزار مورد استفاده در این پژوهش بوده است، می تواند به عنوان زمان قابل قبولی در نظر گرفته شود.

[Downloaded from journals.sut.ac.ir on 2024-11-23]

[DOR: 20.1001.1.23223146.1402.10.1.3.6]

مراجع

1. Hüwel, K. Adiloğlu and J. Bach, "Hearing aid research data set for acoustic environment recognition," in in IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP), Barcelona, Spain, 2006.
2. D. Bogdanov, N. Wack, E. G'omez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata and X. Serra, "ESSENTIA: an audio analysis library for music information retrieval," in Proc ISMIR, Curitiba, Brazil, 2013.
3. M. Crocco, M. Cristani, A. Trucco and V. Murino, "Audio surveillance: A systematic review," ACM Computing Surveys (CSUR), vol. 48, no. 4, pp. 1-46, 2016.
4. F. Shkurti, W. Chang, P. Henderson, M. Islam, J. Higuera, J. Li, T. Manderson, A. Xu, G. Dudek and J. Sattar, "Underwater multi-robot convoying using visual tracking by detection," in in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017.
5. C. Mydlarz, J. Salamon and J. P. Bello, "The implementation of lowcost urban acoustic monitoring devices," Applied Acoustics, vol. 117, pp. 207-218, 2016.
6. M. Valenti, A. Diment, G. Parascandolo, S. Squartini and T. Virtanen, "DCASE2016 acoustic scene classification us-ing convolutional neural networks," in Proc. Workshop DCASE, 2016, pp. 95-99.
7. H. Riazati Seresht and K. Mohammadi, "Environmental sound classification with low-complexity convolutional neural network empowered by sparse salient region pooling," IEEE Access, vol. 11, pp. 849-862, 2022.
8. Al-Hattab, Y. A., H. F. Zaki and A. A. Shafie, "Rethinking environmental sound classification using convolutional neural networks: optimized parameter tuning of single feature extraction," Neural Computing and Applications, vol. 33, no. 21, pp. 14495-14506, 2021.
9. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. ICLR, 2015.
10. G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. CVPR, 2017, pp. 2261-2269.
11. K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, 2016, pp. 770-778.
12. J. K. Piczak, "ESC: Dataset for environmental sound," in in Proc. 23rd ACM Int. Conf. Multimed., 2015.
13. Kumar, M. Khadkevich and C. Fügen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in in IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP), 2018.
14. L. I. Xuhong, Y. Grandvalet and F. Davoine, "Explicit inductive bias for transfer learning with convolutional networks," in In Int. Conf. Mach. Learn., 2018.
15. J. Yosinski, J. Clune, Y. Bengio and H. Lipson, "How transferable are features in deep neural networks?," in Proc. NIPS, 2014, pp. 3320-3328.
16. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho and J. Huopaniemi, "Audio-based context recognition," IEEE Trans. on Audio, Speech, and Lang. Process., vol. 14, no. 1, pp. 321-329, 2005.
17. K. Lee and D. Ellis, "Audio-based semantic concept classification for consumer video," IEEE Trans. on Audio, Speech, and Lang. Process., vol. 18, no. 6, pp. 1406-1416, 2009.
18. I. McLoughlin, "Line spectral pairs," Signal Proces., vol. 88, no. 3, pp. 448-467, 2008

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

19. Temko, E. Monte and C. Nadeu, "Comparison of sequence discriminant support vector machines for acoustic event classification," in In IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP), 2006.
20. G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82-97, 2012.
21. J. Piczak, "Environmental sound classification with convolutional neural networks," in In 2015 IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP), 2015.
22. Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in In IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP), 2017.
23. X. Li, V. Chebiyyam and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," arXiv:1901.08608, 2019.
24. J. Sharma, O. C. Granmo and M. Goodwin, "Environment Sound Classification Using Multiple Feature Channels and Attention Based Deep Convolutional Neural Network," in In Interspeech, 2020.
25. T. Qiao, S. Zhang, S. Cao and S. Xu, "High Accurate Environmental Sound Classification: Sub-Spectrogram Segmentation versus Temporal-Frequency Attention Mechanism," Sensors, vol. 21, no. 16, 2021.
26. W. Mu, B. Yin, X. Huang, J. Xu and Z. Du, "Environmental sound classification using temporal-frequency attention based convolutional neural network," Scientific Reports, vol. 11, no. 1, 2021.
27. H. Wang, Y. Zou, D. Chong and W. Wang, "Environmental sound classification with parallel temporal-spectral attention," arXiv preprint arXiv:1912.06808, 2020.
28. J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Process. Letters, vol. 24, no. 3, pp. 279-283, 2017.
29. Z. Mushtaq and S. F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," Applied Acoustics, vol. 167, 2020.
30. Madhu and S. Kumaraswamy, "Data augmentation using generative adversarial network for environmental sound classification," in In 2019 27th European Signal Processing Conference (EUSIPCO), 2019.
31. S. Mun, S. Shon, W. Kim, D. K. Han and H. Ko, "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," in In IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP), 2017.
32. Z. Mushtaq, S. F. Su and Q. V. Tran, "Spectral images based environmental sound classification using CNN with meaningful data augmentation," Applied Acoustics, vol. 172, 2021.
33. J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. Moore, M. Plakal and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in In IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP), 2017.
34. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang and M. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880-2894, 2020.
35. Y. K. Wang and K. C. Fan, "Applying genetic algorithms on pattern recognition: An analysis and survey," in In Proc. 13th Int. Conf. Pattern Recognit., 1996.

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]

36. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," J. Mach. Lerning Res., vol. 15, no. 1, pp. 1929-1958, 2014.
37. Y. Sun, B. Xue, M. Zhang and G. G. Yen, "Completely automated CNN architecture design based on blocks," IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 4, pp. 1242-1254, 2020.
38. Y. Sun, G. G. Yen and Z. Yi, "Evolving unsupervised deep neural networks for learning meaningful representations," IEEE Trans. Evol. Comput., vol. 23, no. 1, pp. 89-103, 2018.
39. S. Katoch, S. S. Chauhan and V. Kumar, "A review on genetic algorithm: past, present, and future," Multimedia Tools and Applications, vol. 80, no. 5, pp. 8091-8126, 2021.
40. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in In Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), 2018.
41. S. Hershey, S. Chaudhuri, D. Ellis, J. Gemmeke, A. Jansen, R. Moore, M. Plakal, D. Platt, R. Saurous, B. Seybold and M. Slaney, "CNN architectures for large-scale audio classification," in In IEEE Int. Conf. Acoust, Speech Signal Process. (ICASSP), 2017.
42. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard and M. Kudlur, "Tensorflow: A system for large-scale machine learning," in in Proc. 12th USENIX Symp. Oper. Syst. Des. Implement., 2016, pp. 265-283.
43. Y. Aytar, C. Vondrick and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in Proc. NIPS, 2016, pp. 892-900.
44. S. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans Knowl. Data Eng., vol. 22, no. 10, pp. 1345-1359, 2010.
45. Z. Mushtaq and S. F. Su, "Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images," Symmetry, vol. 12, no. 11, 2020.

[Downloaded from journals.sut.ac.ir on 2024-11-23]
[DOR: 20.1001.1.23223146.1402.10.1.3.6]