مجله ماشین بینایی و پردازش تصویر
سال دوازدهم، شماره سوم، پاییز ۱۴۰۴

طراحی شبکه عصبی سبک برای تشخیص جعل عمیق چهره با استفاده از تقطیر دانش و هرس مبتنی بر ضریب همبستگی پیرسون

سارا عسکری همت[۱]، مهدی افتخاری[۲]

چکیده

امروزه شبکه‌های عصبی پیچشی به‌طور گسترده برای تشخیص جعل عمیق چهره به کار می‌روند. اما به دلیل تعداد زیاد پارامترها و هزینه محاسباتی سنگین چالش‌برانگیز هستند. هدف این مقاله، طراحی شبکه‌ای سبک و با دقت قابل قبول است تا در دستگاه‌های با منابع محدود قابل استفاده باشد. بدین منظور، از روش‌های تقطیر دانش و هرس فیلترها برای فشرده‌سازی شبکه بهره گرفته شد. از شبکه پیش‌آموزش‌دیده ResNet50 به‌عنوان معلم برای انتقال دانش به دانش‌آموز استفاده شد. همچنین، از ماسک‌های دودویی برای هرس فیلترها استفاده گردید. ایده اصلی این مقاله، بهره‌گیری از ضریب همبستگی پیرسون برای شناسایی فیلترهای زائد و هدایت فرآیند هرس است. این روش روی پنج مجموعه‌داده اعمال و سپس با دو روش که یکی از آنها الهام‌بخش این پژوهش بود، مقایسه شد. در این مقایسه، حداکثر میزان هرس ممکن در روش‌ها اعمال گردید. روش پیشنهادی در اکثر موارد، فشردگی بیشتر و قدرت تعمیم قابل قبولی دارد. به عنوان مثال در مجموعه‌داده 140k Real and Fake Faces علاوه بر بهبود ۴/۳۱ درصدی دقت نسبت به معلم، به کاهش ۸۵/۳۳ درصد در پارامترها و ۸۳/۹۲ درصد در محاسبات دست یافت و شبکه فشرده‌شده قدرت تعمیم مشابه با روش پایه از خود نشان داد.

کلید واژه‌ها

جعل عمیق چهره، تقطیر دانش، هرس، همبستگی ، پارامتر، محاسبات.

۱- مقدمه

یکی از شاخه‌های نوظهور در حوزه پردازش تصویر، جعل عمیق چهره است که به معنای استفاده از الگوریتم‌های یادگیری عمیق برای ایجاد یا تغییر ویژگی‌های چهره به صورت واقع‌بینانه است؛ به‌گونه‌ای که تشخیص میان ویژگی‌های واقعی و جعلی دشوار شود. این فناوری پیشرفت چشمگیری داشته و کاربردهای بسیاری در صنعت فیلم سازی، بهبود جلوه‌های بصری و بازی‌های ویدئویی داشته است[۱]. جعل عمیق چهره در کنار این جنبه‌های مثبت، در فعالیت‌های مجرمانه نیز استفاده می‌شود. به همین دلیل استفاده از روش‌های تشخیص جعل عمیق چهره[۱] به دلیل گسترش روش‌های تولید محتواهای جعل عمیق و جنبه‌های منفی آن به سرعت مورد توجه قرار گرفت. از کاربردهای تشخیص جعل عمیق چهره می‌توان به حفاظت از امنیت و حریم خصوصی، جلوگیری از فریب افکار عمومی، حفاظت از شهرت افراد، پیشگیری از باج‌خواهی، آسیب‌های روانی و حتی پیشگیری از تنش‌های سیاسی بین کشورها اشاره کرد.

تشخیص جعل عمیق چهره با کمک شبکه‌های عصبی پیچشی[۲] منجر به دقت قابل توجهی می‌شود. با این حال تعداد پارامترهای زیاد و هزینه محاسباتی بالا در این شبکه‌ها از چالش‌های قابل اشاره هستند و این شبکه‌ها برای استفاده در دستگاه‌های دارای سخت افزار محدود مناسب نیستند. اگر شبکه‌های تشخیص جعل عمیق چهره به صورت فشرده و در عین حال مؤثر بتوانند روی دستگاه‌های

این مقاله در تاریخ ۱ مهر ماه ۱۴۰۴ دریافت شد.

[۱] بخش مهندسی کامپیوتر، دانشگاه شهید باهنر کرمان.
رایانامه: sarahemmat@eng.uk.ac.ir
[۲] بخش مهندسی کامپیوتر، دانشگاه شهید باهنر کرمان.
رایانامه: m.eftekhari@uk.ac.ir

نویسنده مسئول: مهدی افتخاری

[1] Deepfake Face Detection
[2] Convolutional Nural Network (CNN)

کوچك و قابل حمل استفاده شوند، به‌طور قابل توجهی به مقابله با انتشار جعل عمیق کمک خواهند کرد.

یکی از روش‌هایی که برای فشرده کردن شبکه‌ها به کار می‌رود، استفاده از رویکرد تقطیر دانش[1] است. روش تقطیر دانش فرآیندی است که طی آن یک شبکه بزرگ و پیچیده (معلم[2]) دانش خود را به یک شبکه کوچک‌تر و سبک‌تر (دانش‌آموز[3]) منتقل می‌کند [۲]. این روش به شبکه کوچک‌تر اجازه می‌دهد تا با حفظ دقت قابل قبول، با منابع محاسباتی و حافظه کمتر، عملکرد مشابه شبکه بزرگ (معلم) را داشته باشد. تقطیر دانش شامل رویکردهای گوناگونی از جمله تقطیر دانش براساس تعداد معلم‌ها، براساس نوع داده‌ها، برخط[4] و بدون معلم، براساس برچسب‌ها و براساس رویکردهای نوین قابل ارزیابی است [۳].

اطلاعاتی که در شبکه معلم نهفته است دانش تاریک[5] نام دارد و همان اطلاعاتی است که باید تقطیر شود. این دانش شامل اطلاعات پنهان مانند درصد احتمال هر دسته است و شبکه دانش‌آموز یاد می‌گیرد پیش‌بینی‌هایش را با معلم هماهنگ کند. این پیش‌بینی‌ها از طریق تابع softmax که مقادیر خام شبکه را به درصدهای احتمال هر دسته تبدیل می‌کند، مقایسه می‌شوند. دانش‌آموز طوری آموزش می‌بیند که مقادیر خام آن به مقادیر خام معلم نزدیک شود تا خروجی‌های تابع softmax آنها شبیه هم شوند[۳].

در بسیاری از مواقع، خروجی تابع softmax روی مقادیرخام معلم به گونه‌ای است که دسته درست با احتمال بسیار بالا انتخاب می‌شود و سایر دسته‌ها دارای احتمالاتی بسیار نزدیک به صفر هستند. در چنین شرایطی اطلاعات مفیدی از سایر دسته‌ها و چگونگی یادگیری آنها توسط معلم به دست نمی‌آید تا بتوان به دانش‌آموز انتقال داد. برای رفع این مشکل، پارامتر softmax temperature معرفی شد که می‌تواند هدف را نرم‌تر کند[۲]:

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \qquad (۱)$$

در این عبارت z_i دسته فعلی، z_j سایر دسته‌ها، T دما و q_i احتمال نهایی است. هرچه دما بیشتر شود، توزیع احتمالاتی که تولید می‌شود نرم‌تر خواهد بود و اطلاعات بیشتری درباره دسته‌هایی که شبکه معلم آنها را مشابه دسته پیش‌بینی‌شده یافته است ارائه می‌دهد.

روشی دیگر برای فشرده کردن و همچنین کاهش تعداد پارامترهای شبکه، روش هرس[6] است. به‌طور کلی هرچه شبکه بیشتر هرس شود، سبک‌تر می‌شود؛ اما همزمان عملکرد آن تحت تأثیر قرار گرفته و دقت کاهش می‌یابد. بنابراین، تعادل میان میزان هرس کردن و از دست دادن دقت امری بسیار مهم است. روش‌های هرس به‌طور کلی به دو نوع هرس غیر ساختاری[7] و هرس ساختاری[8] تقسیم می‌شود.

در هرس غیر ساختاری وزن‌ها، نورون‌ها یا اتصالات خاص حذف شده و وزن‌های منفرد (اتصالات بین نورون‌ها) حذف می‌شوند، بدون اینکه تعداد نورون‌ها یا ساختار کلی شبکه تغییر کند. در هرس ساختاری بخش‌های بزرگ‌تری مانند فیلترها، کانال‌ها و لایه‌ها حذف می‌شوند[۴]. هرس غیرساختاری به دلیل دسترسی نامنظم به حافظه ممکن است سرعت استنتاج عملی شبکه را کاهش دهد اما هرس ساختاری اینگونه نیست[۵]. همچنین، هرس فیلترها را می‌توان به دو دسته هرس سخت و نرم تقسیم کرد. در هرس سخت فیلترهای هرس‌شده در فرآیند آموزش ثابت نگه داشته می‌شوند. اما این روش انعطاف‌پذیر نیست و حذف اشتباه فیلترهای مهم غیرقابل بازگشت است. در هرس نرم فیلترهای زائد به‌صورت پویا هرس شده و فیلترهایی که به اشتباه هرس شده‌اند می‌توانند بازیابی شوند[۵].

در این مقاله از ترکیب روش‌های تقطیر دانش و هرس فیلترها برای ایجاد یک شبکه دانش‌آموز سبک و با دقت مشابه معلم برای تشخیص جعل عمیق چهره استفاده شده است.

۲- مرور کارهای پیشین

۲ـ۱ـ روش‌های تشخیص جعل عمیق چهره

روش‌های تشخیص جعل عمیق چهره به دو دسته تقسیم می‌شوند: روش‌های مبتنی بر یادگیری ماشین و روش‌های مبتنی بر شبکه‌های عصبی عمیق. روش‌های مبتنی بر یادگیری ماشین از مدل‌های آماری و تحلیل الگوها برای شناسایی ناهنجاری‌ها و ناسازگاری‌های موجود در محتواهای جعلی استفاده می‌کنند. در این رویکرد، ویژگی‌هایی مانند خصوصیات آماری، توزیع رنگ، الگوهای بافتی و سایر خصوصیت‌های قابل مشاهده از تصاویر و ویدئوها استخراج می‌شوند[۶]. بدین منظور، مدل‌هایی مانند درخت تصمیم[9]، ماشین بردار پشتیبان[10] و جنگل تصادفی[11] به‌طورگسترده مورد استفاده قرار می‌گیرند. همچنین، مطالعات اولیه که بر تحلیل شاخص‌هایی مانند الگوهای پلک زدن[۸,۷] و زاویه سر [۹] تمرکز داشته‌اند، نقش مهمی در شکل‌گیری پایه نظری این حوزه ایفا کرده‌اند.

با این حال، این روش‌ها هرچند در مجموعه‌داده‌های محدود عملکرد قابل قبولی داشته‌اند، اما در برابر محتواهای جعل عمیق جدید که با استفاده از روش‌های پیشرفته تولید محتوا ساخته می‌شوند، به‌راحتی فریب می‌خورند[۱۰]. به‌منظور غلبه بر این چالش‌ها، روش‌های مبتنی بر شبکه‌های عصبی عمیق شامل

[1] Knowledge Distillation
[2] Teacher
[3] Student
[4] Online
[5] Dark Knowledge
[6] Pruning
[7] Unstructured Pruning
[8] Structured Pruning
[9] Decision Tree
[10] Support Vector Machine (SVM)
[11] Random Forest

شبکه‌های عصبی پیچشی به‌طور گسترده برای تحلیل ویژگی‌های فضایی به کار گرفته شده‌اند[۱۱].

۲ـ۲ـ کاربرد تقطیر دانش در تشخیص جعل عمیق چهره

در زمینه تقطیر دانش چند رویکرد کلیدی برای شناسایی محتواهای جعل عمیق معرفی شده‌اند. مقاله [۱۲] یک روش انتقال تطبیقی نمایش ویژگی‌ها ارائه داده است که بین شبکه معلم (دامنه منبع) و شبکه دانش‌آموز (دامنه هدف) ارتباط برقرار می‌کند. این روش امکان تطبیق کارآمد آشکارسازها با انواع جدید جعل عمیق را فراهم می‌سازد. بر پایه اصول مشابه، [۱۳] مشکل فراموشی فاجعه‌آمیز[1] در تطبیق‌های دامنه‌ای متوالی[2] را با استفاده از معماری معلم ـ دانش‌آموز کاهش می‌دهد. این روش، دانش را هم در سطح ویژگی‌های نهان و هم در سطح پیش‌بینی نهایی حفظ می‌کند.

مرجع [۱۴] یک چارچوب تقطیر دانش دو در یک معرفی کرده‌است که اطلاعات مکانی و فرکانسی را از یک شبکه دو شاخه‌ای[3] به یک شبکه تک شاخه‌ای انتقال می‌دهد. این روش با استفاده از همگن‌سازی گرادیان تضادهای اطلاعاتی بین این ویژگی‌ها را کاهش داده و موجب بهبود عملکرد شبکه می‌شود.

در سال ۲۰۲۲ چارچوبی مبتنی بر یادگیری تضاد نظارت‌شده[4] و تقطیر دانش برای مواردی که داده‌های برچسب‌خورده محدودی دارند ارائه شده است [۱۵]. این چارچوب از یک رویکرد سه مرحله‌ای شامل پیش‌آموزش خودنظارتی با نمونه‌های بدون برچسب، آموزش نظارت‌شده با داده‌های برچسب‌خورده محدود و تقطیر دانش برای ایجاد یک شبکه دانش‌آموز فشرده با قابلیت تعمیم بهتر در میان مجموعه‌های داده بهره می‌برد.

برای شناسایی جعل عمیق در دامنه‌های مختلف به‌طور خاص، یک چارچوب تطبیقی طراحی شده است. این روش با استفاده از بهینه‌سازی همزمان شبکه‌های معلم و دانش‌آموز، انتقال دانش پنهان بین دامنه‌ها را آسان می‌کند که هدف آن کاهش شکاف عملکردی میان مجموعه‌های داده منبع و هدف، بهبود تعمیم‌پذیری شبکه و افزایش دقت شناسایی جعل عمیق در موارد با تغییر دامنه است[۱۶].

در [۱۷] چارچوبی ارائه شده که از تقطیر دانش مبتنی بر توجه[5] همراه با یادگیری در حوزه فرکانس و بهره‌گیری از نظریه انتقال بهینه استفاده می‌کند تا عملکرد تشخیص جعل عمیق را در تصاویر فشرده و با کیفیت پایین بهبود دهد. این چارچوب شامل تقطیر توجه فرکانسی برای بازیابی اجزای فرکانس بالای از دست‌رفته در تصاویر فشرده و تقطیر توجه چنددیدگاهی[6] جهت انتقال مؤثرتر توزیع تنسورها بین معلم و دانش‌آموز است.

۲ـ۳ـ هرس شبکه و ترکیب آن با تقطیر دانش

در زمینه هرس شبکه‌های تشخیص جعل عمیق چهره، مرجع [۱۸] با استفاده از هرس تکراری مبتنی بر اهمیت[7] بر روی معماری‌های مختلف نشان داد که زیرشبکه‌های سبک حتی در سطوح بالای هرس می‌توانند دقت تشخیص را حفظ کنند و همچنان بر نواحی کلیدی چهره متمرکز می‌مانند و قابلیت انتقال این زیرشبکه‌ها بین مجموعه‌دادهای مختلف را دارد.

نویسندگان مقاله [۱۹] نشان دادند شبکه هرس‌شده‌ای که از طریق تقطیر دانش و با کمک شبکه اصلی بازیابی شده، دقت بهتری نسبت به حالتی دارد که از طریق تنظیم دقیق و با استفاده از برچسب نمونه‌ها بازیابی شده است. همچنین استفاده از شبکه اصلی به عنوان معلم، عملکرد بهتری نسبت به استفاده از شبکه‌هایی با ساختار متفاوت (اما با همان دقت شبکه معلم) دارد.

بررسی اثر هرس در تقطیردانش برای بهبود فشرده‌سازی شبکه‌های عصبی انجام شده است [٤]. طبق نتایج، دانش‌آموزهایی که از معلمان هرس‌شده آموزش دیده‌اند عملکرد بهتری نسبت به دانش‌آموزهایی دارند که از معلمان بدون هرس به دست آمده‌اند دارند. حتی زمانی که دقت معلم هرس‌شده کمتر باشد. همچنین، یک طرح فشرده‌سازی پیشنهاد شده که دانش‌آموز بر اساس ساختار معلم هرس‌شده طراحی می‌شود.

محققان در چارچوب هرس تکرارشونده، مفهوم تقطیر با دستیار[8] را معرفی کرده اند [۲۰]. در این روش، شبکه‌های میانی که در طول فرآیند هرس تکرارشونده ذخیره شده‌اند به عنوان دستیار معلم عمل می‌کنند. ابتدا دانش از معلم به دستیار و سپس از دستیار به دانش‌آموز منتقل می‌شود .این شبکه‌های میانی شکاف ظرفیت را پر کرده و به انتقال دانش روان‌تر و بازیابی مؤثرتر دقت کمک می‌کنند.

به طور خاص مطالعاتی در مورد مشکل هرس کردن معماری‌های ResNet انجام شده است [۲۱]. تنها لایه‌هایی از ResNet که وابستگی ابعادی ندارند هرس می‌شوند. برای فشرده‌سازی لایه‌هایی که قابل هرس نبودند، از تقطیر دانش استفاده شد. سپس شبکه هرس‌شده در مرحله اول به عنوان معلم عمل می‌کند و یک شبکه دانش‌آموز جدید ساخته می‌شود که در آن، ابعاد لایه‌های هرس‌نشده به صورت دستی کاهش یافته است.

در کار چن و همکاران [۲۲] با هدف عدم نیاز به معلم پیش‌آموزش دیده، هرس زودهنگام شبکه و روش خودتقطیری با هم ترکیب می‌شوند. برای تصمیم‌گیری در مورد هرس از یک تابع زیان خودتقطیری استفاده می‌شود. این کار باعث می‌شود وزن‌هایی که برای فرآیند خودتقطیری مهم‌تر هستند در مرحله هرس زودهنگام حفظ شوند.

---

[1] Catastrophic Forgetting
[2] Sequential Domain Adaptation
[3] Dual Branch
[4] Supervised Contrastive Learning
[5] Attention
[6] Multi View
[7] Iterative Magnitude-Based Pruning
[8] Assistant

اخیراً یک روش هرس تکرارشونده ارائه شده که در آن اهمیت فیلترها بر اساس ترکیبی از ویژگی‌های پیدا شده از نقشه‌های ویژگی سنجیده می‌شود [۲۳]. از ماتریس گرام[1] برای سنجش مرتبط بودن ویژگی و از مقدار آنتروپی برای اندازه‌گیری ظرفیت اطلاعاتی استفاده کردند تا فیلترهای زائد شناسایی شوند. سپس برای بازیابی دقت در هر مرحله از یک روش تقطیر دانش ترکیبی بهره بردند که ویژگی‌های لایه‌های میانی و خروجی نهایی معلم را به دانش‌آموز منتقل می‌کند.

برخی محققان نشان دادند که شبکه دانش‌آموز در فرآیند تقطیر دانش، خود می‌تواند یک ساختار شامل افزونگی داشته باشد. بر این اساس، روش هرس در حین تقطیر پیشنهاد داده شد [۲۴]. نوآوری اصلی این روش طراحی یک ماسک پویا و مشتق پذیر برای هر لایه پیچشی شبکه دانش‌آموز است تا تصمیم گرفته شود کدام کانال ها حذف و کدام‌ها حفظ شوند. این ماسک ها با تابع ApproxSign به صورت باینری تبدیل می‌شوند و در طول آموزش بر اساس زیان کل شبکه به‌روزرسانی می‌شوند. دانش‌آموز به صورت خودکار یاد می‌گیرد کدام کانال‌هایش مهم هستند و کدام یک تکراری و قابل حذف‌اند. پس از اتمام آموزش، کانال‌های غیرضروری حذف شده و شبکه نهایی برای تثبیت عملکرد، تنظیم دقیق می‌شود.

برای بهینه‌سازی فرآیند هرس فیلتر در شبکه‌های پیچشی، چارچوبی به نام KDFS[3] معرفی شد[۵]. این چارچوب با هدف کاهش بار محاسباتی و حافظه طراحی شده که شبکه دانش‌آموز با بهره‌گیری از دانش یک شبکه معلم هرس می‌شود. در این چارچوب، فرآیند هرس به یک مسئله بهینه‌سازی سراسری و گرادیان‌محور تبدیل شده است تا دانش شبکه معلم به درستی به دانش‌آموز منتقل شود. از یک عبارت جریمه برای اندازه‌گیری و کنترل هزینه محاسباتی شبکه دانش‌آموز استفاده شده است. به این صورت که یک نرخ فشرده سازی اعمال می‌شود و تابع هدف، دانش‌آموز را به سمت رسیدن به آن نرخ هدایت می‌کند. روش پیشنهادی از همین چارچوب وام گرفته است و در بخش‌های بعدی به تغییرات اعمال شده بر این روش پرداخته خواهد شد.

۲ـ۴ـ استفاده از ضریب همبستگی پیرسون در هرس شبکه

روش CorrNet برای هرس فیلترهای شبکه‌های عصبی پیچشی ارائه شده است که بر اساس ضریب همبستگی پیرسون بین نقشه‌های ویژگی متوالی عمل می‌کند [۲۵]. یک ماژول انتخاب ویژگی مبتنی بر همبستگی بین لایه‌های پیچشی گنجانده می‌شود تا فیلترهای غیرضروری را با محاسبه همبستگی بین نقشه‌های ویژگی خروجی هر لایه شناسایی کند تا یکی از فیلترهای متناظر بدون آسیب جدی به عملکرد شبکه حذف شود . پس از هرس، شبکه با فرآیند تنظیم دقیق بهبود یافته و این فرآیند به‌صورت تکرارشونده انجام می‌شود تا به معماری بهینه‌ای دست یابد.

در مطالعه دیگری، جفت فیلترهای با همبستگی بالا شناسایی شده و با حذف یکی از فیلترهای هر جفت، افزونگی در شبکه کاهش پیدا می‌کند [۲۶]. نوآوری این کار در تعریف یک تنظیم کننده مبتنی بر ضریب همبستگی پیرسون در تابع هزینه است که قبل از هرس، شباهت بین جفت‌های فیلتر را به حداکثر می‌رساند [۲۶]. این بهینه‌سازی منجر به انتقال دانش بین فیلترهای مشابه می‌شود و امکان حذف ایمن یکی از فیلترها بدون اتلاف قابل توجه اطلاعات فراهم می‌آید. پس از هرس، یک دوره کوتاه برای تنظیم دقیق شبکه انجام می‌شود تا عملکرد شبکه بازیابی گردد. این روش به‌صورت تکرارشونده اجرا می‌شود تا دستیابی به نرخ فشرده‌سازی مطلوب ممکن باشد.

نویسندگان مقاله [۲۷] راهبردی برای هرس شبکه‌های تشخیص اشیا با هدف کاهش همزمان تعداد پارامترها و هزینه محاسباتی پیشنهاد دادند. در طول آموزش شبکه، یک محدودیت عدم همبستگی چندسطحی به تابع زیان اصلی اضافه می‌شود که با کمک ضریب همبستگی پیرسون، همبستگی بین نقشه‌های ویژگی کانال‌های مختلف را کاهش می‌دهد. این امر منجر به یادگیری ویژگی‌های با همبستگی کم و مکمل هم می‌شود که اطلاعات غنی‌تری را دارند. پس از آموزش با این محدودیت، روش پیشنهادی یک معیار پویای ارزیابی اهمیت کانال‌ها بر اساس نُرم L1 نقشه‌های ویژگی خروجی هر لایه تعریف می‌کند. این روش اهمیت کانال‌ها را بر اساس مقدار خروجی نقشه‌های ویژگی قضاوت می‌کند.

در پژوهشی که اخیراً انجام شده است، ضریب همبستگی پیرسون در تابع زیان لحاظ شده است [۲۸]. از این زیان در مرحله قبل از آموزش استفاده می‌شود تا شبکه برای افزایش همبستگی بین پیکسل‌های نقشه‌های ویژگی آموزش ببیند که بعداً بتواند فیلترهای افرونه را شناسایی کند [۲۸]. برای هر کانال، امتیاز اهمیت با محاسبه تفاوت در نُرم هسته‌ای[2] قبل و بعد از حذف کانال تعیین می‌شود. کانال‌هایی که همبستگی بالایی با کانال‌های دیگر دارند به عنوان کانال‌های افزونه شناسایی می‌شوند.

تحقیق به روز دیگری از ضریب همبستگی پیرسون برای شناسایی فیلترهای زائد در شبکه‌های عصبی پیچشی استفاده می‌کند [۲۹]. در این روش، برای هر فیلتر، میانگین k بزرگ‌ترین ضریب همبستگی آن با سایر فیلترهای همان لایه محاسبه می‌شود تا معیاری برای سنجش اهمیت نسبی آن فیلتر به‌دست آید [۲۹]. با در نظر گرفتن جریمه‌های مبتنی بر تعداد پارامترها و هزینه محاسباتی، اهمیت فیلترها در لایه‌های مختلف به‌گونه‌ای نرمال‌سازی می‌شود که امکان تعیین هرس در هر لایه فراهم گردد. ضریب همبستگی پیرسون تنها در مرحله ارزیابی اهمیت فیلترها به‌کار گرفته می‌شود. پس از

[1] Gram Matrix
[2] Nuclear Norm
[3] Knowledge-driven Differential Filter Sampler

انتخاب فیلترهای زائد، شبکه با استفاده از یک فرآیند هرس تکرارشونده و یک طرح تنظیم دقیق چندهدفه بهینه می‌گردد.

نویسندگان مقاله [۳۰] معیاری به نام نسبت واریانس ویژگی را پیشنهاد دادند. این معیار بر مبنای ضریب همبستگی پیرسون بنا شده است که به طور ایده‌آل میزان اتلاف اطلاعات در یک نقشه ویژگی پس از حذف یک کانال را اندازه‌گیری می‌کند. با توجه به اینکه محاسبه مستقیم ضریب پیرسون برای هزاران کانال بسیار پرهزینه است، این معیار به عنوان یک تقریب کارآمد و مؤثر عمل می‌کند. واریانس ویژگی برای یک کانال، نسبت واریانس خروجی فیلترشده آن کانال به واریانس نقشه ویژگی نهایی را محاسبه می‌کند. پس از آموزش شبکه، کانال‌ها به صورت سراسری بر اساس این معیار رتبه‌بندی شده و کم‌اهمیت‌ترین‌ها حذف می‌شوند. این رویکرد کانال‌هایی را شناسایی می‌کند که به طور ذاتی اطلاعات مفیدی را منتقل می‌کنند.

۳- روش پیشنهادی

شکل ۱ چارچوب کلی روش پیشنهادی را نشان می‌دهد. ابتدا چارچوب KDFS و سپس روش پیشنهادی توضیح داده می‌شود. هدف، هرس کردن شبکه دانش‌آموز است تا به شبکه سبک‌تر تبدیل شود، در حالی که معماری آن مشابه معلم است و دانش خود را از آن دریافت می‌کند. از نقشه‌های ویژگی[1] شبکه معلم به عنوان دانش برای هدایت دانش‌آموز استفاده می‌شود. با استفاده از یک نمونه‌بردار[2] مبتنی بر گرادیان، هرس نرم اعمال می‌شود. یعنی تصمیم‌گیری برای حذف یا نگه‌داشتن هر فیلتر، بخشی از فرآیند یادگیری شبکه می‌شود.

در مرحله آموزش، ابتدا ماتریس پارامتر نمونه بردار P با روش Kaiming Initialization به ازای هر لایه پیچشی تولید می‌شود:

$$P_l \in R^{C_l \times 2} \qquad (۲)$$

که C_l تعداد فیلترها در هر لایه l است. در هر لایه پیچشی l،یک ماتریس پارامتر P وجود دارد. این ماتریس دو ستون و به تعداد C_l، سطر دارد. در مرحله رو به جلو[3]، برای هر لایه یک ماسک دودویی m_l نیاز است که مشخص می‌کند کدام فیلترها باید حفظ و کدام حذف شوند. برای تخمین ماسک دودویی، تابع Gumbel-Softmax به ماتریس P اعمال می‌شود :

$$\pi_l^{(i,k)} = \frac{\exp[(P_l^{(i,k)}, G_l^{(i,k)})/\tau]}{\sum_{j=1}^{2} \exp[(P_l^{(i,j)}, G_l^{(i,j)})/\tau]} \qquad (۳)$$

که در آن بردار $\pi_l^{(i,k)}$ به معنای احتمال تعلق فیلتر iام به وضعیت k(حفظ یا حذف) در لایه l ام و $G_l^{(i,k)}$ نویز تصادفی است که از توزیع Gumbel نمونه برداری می‌شود. دما با τ نشان داده شده و به صورت نمایی طبق رابطه زیر کاهش می یابد:

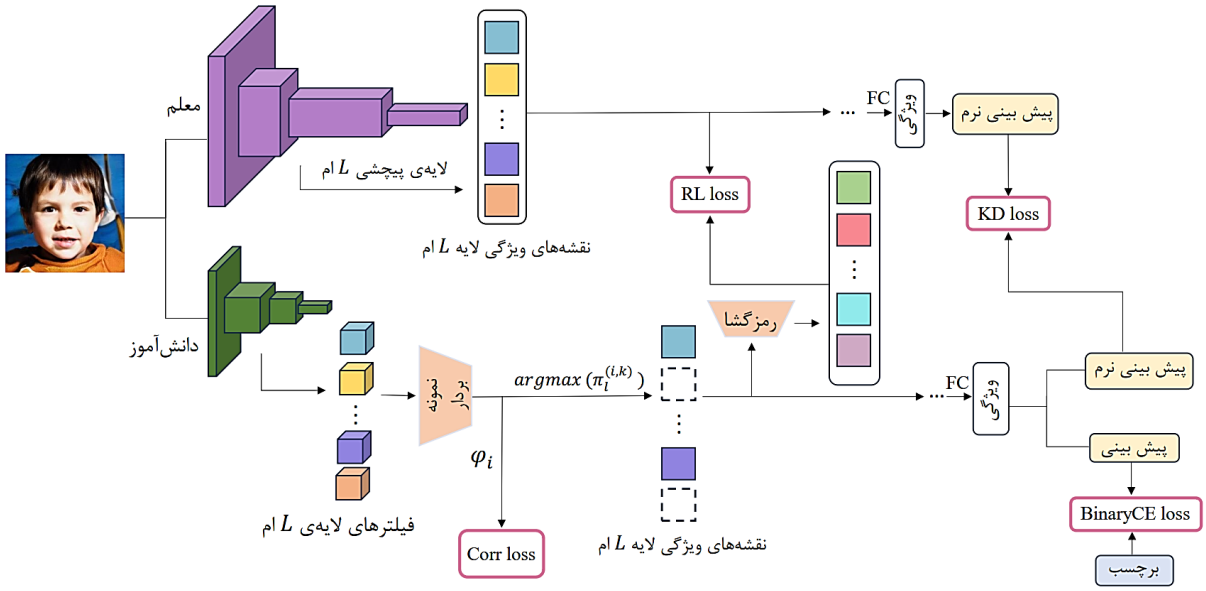
$$\tau(e) = \tau_0 \left(\frac{\tau_E}{\tau_0}\right)^{\frac{e}{E}} \qquad (۴)$$

که e شماره دور فعلی، E تعداد کل دورها، τ_0 دمای اولیه و τ_E دمای پایان است. سپس به بردار π_l عملگر $argmax$ اعمال می‌شود تا ماسک‌های دودویی حاصل شوند. در مرحله پس انتشار[4]، به‌روز رسانی ماسک‌ها انجام می‌شود؛ اما چون این ماسک دودویی است، امکان محاسبه مقدار گرادیان از آن وجود ندارد. در نتیجه از بردار π_l که شامل مقادیر پیوسته است برای گرادیان گرفتن استفاده می‌شود.

شبکه با تابع هدف که شامل سه عبارت زیان و یک عبارت جریمه است آموزش می‌بینند. این چهار بخش به صورت همزمان بهینه‌سازی می‌شوند تا شبکه دانش‌آموز به بهترین توازن بین دقت و هزینه محاسباتی دست یابد. مطلوب است این تابع هدف به کمترین مقدار ممکن برسد. برای یادگیری بهینه دانش‌آموز، نقشه‌های ویژگی معلم و خروجی‌های نمونه‌بردار دانشجو هم‌تراز می‌شوند. بدین منظور از یک رمزگشا[5] در هر مرحله ی کاهش ابعاد استفاده می‌کند تا شبکه دانش‌آموز را مجبور کند خروجی‌های پیچیده شبکه معلم را از روی اطلاعات فشرده خودش بازسازی کند. سپس شبکه با هدف کم کردن تفاوت بین خروجی بازسازی‌شده دانش‌آموز و خروجی اصلی معلم آموزش داده می‌شود. این تفاوت با زیان بازسازی اندازه‌گیری می‌شود.

برای انتقال دانش معلم به دانش‌آموز از تقطیر دانش استفاده شده است. دانش‌آموز از خروجی‌های نرم‌شده معلم، که شامل اطلاعات اضافی هستند کمک می‌گیرد و طوری آموزش داده می‌شود که سعی کند تفاوت خروجی‌های نرم‌شده خودش را تا حد امکان به خروجی‌های نرم‌شده معلم کم کند. با استفاده از زیان بی‌نظمی متقابل[6]، میزان انطباق توزیع احتمالات خروجی دانش‌آموز با جواب واقعی سنجیده می‌شود. با تعیین یک نرخ فشرده‌سازی، هزینه محاسباتی (عملیات ممیز شناور[7]) در دانش‌آموز تقریبا به همان اندازه کاهش می‌یابد. اگر هزینه محاسباتی دانش‌آموز از این هدف دور باشد، یک جریمه به آن تعلق می‌گیرد. در مرحله استنتاج، ماسک باینری نهایی که از ماتریس P به دست آمده مستقیما به شبکه اعمال شده و هرس نهایی انجام می‌شود.

[1] Feature Maps
[2] Sampler
[3] Forward
[4] Backward
[5] Decoder
[6] Cross-entropy loss
[7] Floating Point Operations (FLOPs)



شکل (۱): ساختار کلی روش پیشنهادی

در روش پیشنهادی، به جای جریمه بر اساس عملیات ممیز شناور، از یک عبارت زیان مرتبط با ضریب همبستگی پیرسون استفاده شده تا فیلترهای زائد در هر لایه برای شبکه دانش‌آموز شناسایی شوند. در این مرحله نیز از همان تابع Gumbel-Softmax در فرمول ۳ استفاده شده و فقط احتمال نگه داشتن فیلتر مورد نظر از بردار π_l در نظر گرفته می‌شود. احتمال نگه داشتن هر فیلتر با φ_i نشان داده می‌شود. سپس ماتریس ضریب همبستگی پیرسون ρ بین تمام جفت فیلترهای i و j در هر لایه محاسبه می‌شود:

$$\rho_{i,j} = \frac{1}{W-1}\sum_{k=1}^{W}\left(\frac{\omega_{ik}-\mu_i}{s_i}\right)\left(\frac{\omega_{jk}-\mu_j}{s_j}\right) \quad (۵)$$

که W تعداد کل وزن‌ها در هر فیلتر، ω_{ik} و ω_{jk} وزن k ام در فیلتر i و j، μ_i و μ_j میانگین وزن‌های فیلتر i و j، s_i و s_j انحراف معیار وزن‌های فیلتر i و j است.

حال برای هر فیلتر، امتیاز افزونگی آن با استفاده از میانگینِ مربعِ همبستگی آن با سایر فیلترها محاسبه می‌شود. این امتیاز نشان می‌دهد که یک فیلتر به طور متوسط چقدر با تمام فیلترهای دیگر در آن لایه همبستگی دارد. در محاسبه این امتیاز، قطر اصلی ماتریس همبستگی در نظر گرفته نمی‌شود. فرمول امتیاز همبستگی برای فیلتر i در لایه l به صورت زیر است:

$$Score_{l,i} = \frac{1}{C-1}\sum_{j=1,j\neq i}^{C}\left(\rho_{i,j}\right)^2 \quad (۶)$$

که در آن C تعداد کل فیلترهای لایه l است. هر چه امتیاز بیشتر باشد، همبستگی با سایر فیلترها هم بیشتر است. زیان همبستگی هر لایه از طریق ضرب امتیازهای همبستگی در φ_i به دست می‌آید:

$$L_{corr,l} = \frac{1}{L}\sum_{i=1}^{L} Score_{l,i}.\varphi_i \quad (۷)$$

در این روش اگر یک فیلتر، هم امتیاز افزونگی بالایی با فیلتر دیگر داشته باشد و هم شبکه تمایل به نگه داشتن آن (احتمال بالا) داشته باشد، زیان بزرگی تولید می‌شود. این زیان، شبکه را مجبور می‌کند تا تمایل خود برای نگه داشتن آن فیلتر زائد را کاهش دهد. در نهایت برای هر لایه این عملیات محاسبه شده و با میانگین‌گیری از زیان‌های به دست آمده از هر لایه، یک عدد واحد را به عنوان زیان همبستگی کل برمی‌گرداند. این زیان به تابع هدف اصلی اضافه شده و در فرآیند پس‌رو استفاده می‌شود.

از آنجایی که مسئله تشخیص جعل عمیق یک مسئله ی دسته بندی دودویی است، از تابع زیان بی نظمی متقاطع دودویی به همراه تابع فعال‌سازی سیگموئید استفاده شده است[۳۱]:

$$q_i = \frac{1}{1+\exp(-z_i/T)}, p_i = \frac{1}{1+\exp(-v_i/T)} \quad (۸)$$

که z_i ورودی iام لایه آخر معلم و v_i ورودی iام لایه آخر دانش‌آموز است. در نهایت، تابع هزینه تقطیر دانش چنین است:

$$L_{KD} = -\frac{1}{N}\sum_{i=1}^{N} q_i\, log(p_i) + (1-q_i)\log(1-p_i)].T^2 \quad (۹)$$

که N تعداد نمونه‌هاست. حال می‌توان تابع هزینه نهایی را به صورت زیر تعریف کرد:

(۱۰)

$$argmin_{\theta,P}\frac{1}{N}\sum_{i=1}^{N} L_{CE}\left(y_i, \emptyset(x_i, W, m(P))\right) + \lambda_1\sum_{i=1}^{N} L_{KD}(p_i, q_i) + \lambda_2\sum_{l=1}^{L} L_{RL}\left(\mathbb{Z}_l^{(t)}, D_{\theta l}\left(\mathbb{Z}_l^{(t)}; \theta_l\right)\right) + \lambda_3\left(\frac{1}{L}\sum_{l=1}^{L} L_{corr,l}\right)$$

که y_i مقدار واقعی برچسب داده است. عبارت $\emptyset(x_i, W, m)$ به معنی پیش بینی شبکه با ورودی x_i ،همراه با معلم از پیش آموزش دیده با وزن‌های W و ماسک دودویی m است. ضرایب λ_1وλ_2وλ_3 به عنوان ابرپارامتر برای شدت اثر زیان‌ها استفاده می‌شوند. عبارت $D_{\theta l}(\mathbb{Z}_l^{(t)}; \theta_l)$ به معنی ویژگی‌های بازسازی‌شده از یک رمزگشا با پارامتر θ در لایه lام است که اندازه‌ای برابر با نقشه‌های ویژگی معلم $\mathbb{Z}_l^{(t)}$ دارد. تفاوت روش پیشنهادی با کارهای پیشین که از ضریب همبستگی پیرسون استفاده کرده‌اند در این است که در مراجع [۲۵]، [۲۹] و [۳۰]، همبستگی تنها در مرحله ارزیابی اهمیت فیلترها و به صورت پس‌پردازشی استفاده می‌شود و سپس فرآیند هرس تکرارشونده انجام می‌گیرد. اما در روش پیشنهادی، همبستگی به صورت یک عبارت زیان در تابع هدف ادغام شده و شبکه در طول فرآیند آموزش به صورت پویا یاد می‌گیرد کدام فیلترها را هرس کند.

برخلاف [۲۳] که از ترکیب ماتریس گرام و آنتروپی برای هرس تکرارشونده استفاده می‌کند، روش پیشنهادی با ترکیب امتیاز همبستگی و احتمالات نمونه‌بردار ، یک فرآیند بهینه‌سازی سراسری و مشتق‌پذیر ایجاد می‌کند که همزمان با تقطیر دانش انجام می‌شود.

در مرجع [۲۶] اگرچه از همبستگی در تابع زیان استفاده می‌شود، اما هدف حداکثرسازی همبستگی بین فیلترها در مرحله پیش از هرس است و سپس به صورت تکرارشونده هرس و تنظیم دقیق انجام می‌شود. اما در روش پیشنهادی، همبستگی مستقیماً برای محاسبه امتیاز افزونگی هر فیلتر به کار می‌رود و به صورت یک عبارت زیان در تابع هدف ادغام شده است؛ به‌گونه‌ای که شبکه در طول یک فرآیند آموزش واحد و به صورت پویا یاد می‌گیرد کدام فیلترها را هرس کند.

مراجع [۲۷] و [۲۸] از همبستگی پیرسون برای تغییر الگوی یادگیری شبکه در مرحله آموزش استفاده می‌کنند و سپس بر اساس معیارهای دیگری (نُرم L1 یا نُرم هسته‌ای) هرس انجام می‌دهند. اما روش پیشنهادی به طور مستقیم از همبستگی برای محاسبه امتیاز افزونگی و ایجاد زیان هرس استفاده می‌کند که همزمان با تقطیر دانش بهینه می‌شود.

۴- ارزیابی روش پیشنهادی

۴ـ۱ـ مجموعه‌داده

در این مقاله از مجموعه‌داده‌های Real vs Fake Faces-10k [۳۲]، 140k Real and Fake Faces [۳۳]، deepfake and real images [۳۴] (شامل ۱۹۰ هزار داده)، GRAVEX-200k [۳۵] و Deepfake-dataset [۳۶] (شامل ۳۳۰ هزار داده) استفاده شده است. ابعاد همه تصاویر ۲۵۶×۲۵۶ است. در مجموعه‌داده Real vs Fake Faces-10k، ۸۰ درصد داده‌ها برای آموزش، ۱۰ درصد برای ارزیابی و ۱۰ درصد برای آزمون جدا شدند. سایر مجموعه‌داده‌ها به طور جداگانه شامل داده‌های آموزش، ارزیابی و آزمون بودند. در تمامی مجموعه‌داده‌ها تعداد داده‌های واقعی و جعلی در هر تقسیم بندی با هم برابرند. جزئیات مجموعه‌داده‌ها در جدول ۱ آمده است.

مقادیر میانگین و انحراف معیار برای هر مجموعه‌داده به منظور نرمال‌سازی محاسبه شده است. برای افزایش داده‌ها در داده‌های آموزش از قرینه سازی افقی[1]، چرخش تصادفی[2]، برش تصادفی[3]، تغییرات تصادفی در رنگ[4] و تبدیل تصادفی آفین[5] استفاده شد.

جدول (۱): تعداد داده‌ها در تقسیم‌بندی مجموعه‌داده‌ها

مجموعه‌داده	آموزش	ارزیابی	آزمون
Real vs Fake Faces - 10k	۷۰۰۰	۱۵۰۰	۱۵۰۰
140k Real and Fake Faces	۱۰۰۰۰۰	۲۰۰۰۰	۲۰۰۰۰
deepfake and real images	۱۴۰۰۰۲	۳۹۴۲۸	۱۰۹۰۵
GRAVEX-200K	۱۴۰۰۰۰	۴۰۰۰۰	۲۰۰۰۰
Deepfake-dataset	۲۴۰۰۰۲	۵۹۴۲۸	۳۰۹۰۵

۴ـ۲ـ جزئیات پیاده سازی

از دو پردازنده گرافیکی Tesla T4 به صورت موازی و هرکدام دارای ۱۶ گیگابایت حافظه استفاده شد و آموزش شبکه با موازی‌سازی داده‌ها در محیط توزیع‌شده[6] انجام گرفت. در روش مرجع یا همان KDFS، از مجموعه‌داده‌های Cifar10، Cifar100 و Imagenet برای مسئله دسته بندی استفاده شده است. برای مقایسه این روش با روش پیشنهادی، از مجموعه‌داده‌های ذکر شده در جدول ۱ برای آموزش شبکه در روش KDFS استفاده شد. همچنین در هر دو روش از شبکه ResNet50 به عنوان معلم استفاده شد.

جهت انجام مقایسه بیشتر، از روش PDD [۲۴] استفاده گردید. در این روش شبکه‌های معلم و دانش‌آموز معماری متفاوتی دارند. ما برای مقایسه عادلانه، شبکه معلم را Resnet50 شبکه دانش‌آموز را Resnet18 در نظر گرفتیم و از همین پنج مجموعه‌داده اعلام شده برای آموزش و هرس شبکه دانش‌آموز استفاده کردیم.

ابتدا شبکه پیش‌آموزش‌دیده ResNet50 روی مجوعه داده‌های چهره تنظیم دقیق[7] شد. پارامترهای آخرین بلوک پیچشی به همراه لایه تمام متصل نهایی برای آموزش فعال شدند. سایر لایه‌های پیشین شبکه در تمام طول فرآیند آموزش ثابت[8] باقی ماندند. لایه آخر برای دسته‌بندی دودویی تغییر داده شد. بهینه‌ساز Adam با نرخ

[1] Random Horizontal Flip
[2] Random Rotation
[3] Random Crop
[4] Color Jitter
[5] Random Affine
[6] Distributed Data Parallel
[7] Fine-Tune
[8] Freez

یادگیری ۰/۰۰۰۱ برای پارامترهای فعال و نرخ کاهش وزن ۰/۰۰۰۱ اعمال شد. بدین ترتیب شبکه معلم برای هر مجموعه‌داده به دست آمد. جدول ۲ دقت آزمون هر معلم روی مجموعه‌داده‌ها را نشان می‌دهد.

برای آموزش شبکه دانش‌آموز در روش‌های KDFS و پیشنهادی از بهینه ساز Adam استفاده شد. ابرپارامترهای مناسبِ هر مجموعه‌داده به صورت تجربی به دست آمدند. اندازه دسته‌ها[1] ۱۲۸ تعیین شد. نرخ یادگیری[2] در هر دو روش برای GRAVEX-200K ، ۰/۰۱ و برای دیگر مجموعه‌داده‌ها ۰/۰۰۵ در نظر گرفته شد. تعداد دورها[3] در هر دو روش برای مجموعه‌داده‌های Real vs Fake Faces-10k و deepfake and real images و Deepfake-dataset به ترتیب ۴۰۰ و۲۶ و ۳۵ و برای دیگر مجموعه‌داده‌ها ۸۰ در نظر گرفته شد. مقادیر سایر ابرپارامترها در جدول‌های ۳ و ۴ آمده است. شبکه دانش‌آموز حاصل از هرس با مجموعه‌داده‌های Real vs Fake Faces-10k، GRAVEX-200K و Deepfake-dataset به منظور بهبود دقت به ترتیب با نرخ‌های یادگیری ۰/۰۰۰۰۸، ۰/۰۰۰۰۲ و ۰/۰۰۰۰۴ تنظیم دقیق شدند.

جدول (۲): دقت آزمون شبکه پیش‌آموزش دیده (معلم) روی هر مجموعه‌داده

مجموعه‌داده	دقت آزمون(%)
Real vs Fake Faces-10k	۹۶
140k Real and Fake Faces	۹۴/۷۳
deepfake and real images	۸۵/۶۴
GRAVEX-200K	۹۳/۴
Deepfake-dataset	۹۳/۹۴

جدول (۳): ابرپارامترهای روش KDFS. این مقادیر به صورت تجربی تعیین شدند.

مجموعه‌داده	T	τ_0	τ_E	λ_1	λ_2	λ_3
Real vs Fake Faces - 10k	۳	۲	۰/۲	۱	۱	۰/۵
140k Real and Fake Faces	۲	۱	۰/۱	۰/۵	۱۰	۱۰۰
deepfake and real images	۲	۱	۰/۱	۰/۵	۱	۱
GRAVEX-200K	۲	۱	۰/۱	۰/۵	۱۰	۱۰۰
Deepfake-dataset	۳	۳	۰/۷	۰/۱	۱	۵

جدول (٤): ابرپارامترهای روش پیشنهادی. این مقادیر به صورت تجربی تعیین شدند.

مجموعه‌داده	T	τ_0	τ_E	λ_1	λ_2	λ_3
Real vs Fake Faces-10k	۳	۲	۰/۲	۱	۱	۰/۵
140k Real and Fake Faces	۱	۱	۰/۰۵	۰/۵	۱	۱
deepfake and real images	۲	۱	۰/۱	۰/۵	۱	۱
GRAVEX-200K	۲	۱	۰/۱	۰/۵	۱	۱
Deepfake-dataset	۱/۵	۲	۰/۵	۰/۵	۱	۱

در روش PDD برای آموزش دانش‌آموز نرخ یادگیری ۰/۰۱ و مقدار تکانه[4] ۰/۹ اعمال شد. تعداد دورها برای مجموعه‌داده Real vs Fake Faces-10k ۵۰ و برای دیگر مجموعه‌داده‌ها ۳۰ در نظر گرفته شد. شبکه‌های دانش‌آموز حاصل از هرس نیز با نرخ یادگیری ۰/۰۰۱ تنظیم دقیق شدند.

به منظور بررسی حداکثر ظرفیت هرس، آزمایش‌های متعددی با ابرپارامترهای مختلف انجام شد. براساس آزمایش‌ها مشاهده شد که اگر کاهش پارامتر و عملیات ممیز شناور بیش از ۹۰ درصد باشد، باعث حذف کامل برخی از لایه‌ها در شبکه می‌شوند که با تعریف هرس فیلترها سازگار نیست. همچنین مشاهده شد حداکثر کاهش پارامتر و عملیات ممیز شناورِ ممکن در محدوده ۷۸ تا ۸۸ درصد است.

همچنین قدرت تعمیم[5] هر یک از شبکه‌های دانش‌آموز به دست آمده نیز بر روی سایر مجموعه‌داده‌ها بررسی شد. بدین منظور مقادیر نرمال‌سازی مربوط به مجموعه‌داده‌ای که شبکه دانش‌آموز با آن آموزش دیده، به داده‌ها اعمال شدند. پارامترهای آخرین بلوک پیچشی به همراه لایه تمام متصل نهایی برای آموزش فعال شدند. سایر لایه‌های پیشین شبکه در تمام طول فرآیند آموزش ثابت ماندند و لایه آخر برای دسته بندی دودویی تغییر داده شد. بهینه ساز AdamW با کاهش نرخ یادگیری در هر ۵ دور با ضریب ۰/۱ اعمال شد و نتایج براساس بهترین نرخ‌های یادگیری با هم مقایسه شدند.

۴ـ۳ـ نتایج

با توجه به اختلاف اندکی که در دقت روش پیشنهادی و KDFS در برخی موارد مشاهده شد، جهت بررسی معناداری آماری این تفاوت، از آزمون تی[6] همراه با ۱۰ بار اجرای آزمایش با seedهای متفاوت استفاده گردید و مقدار احتمال[7] گزارش شد.

نتایج آزمایش‌ها در جدول ۵ نشان می‌دهد که روش پیشنهادی در یک مجموعه‌داده عملکرد آماری یکسان با KDFS، در سه مجموعه‌داده دقت قابل رقابت اما با فشرده‌سازی بیشتر، و در یک مجموعه‌داده عملکرد پایدارتر نسبت به PDD دارد. در ادامه جزئیات نتایج آمده است.

در مجموعه‌داده Real vs Fake Faces-10k، روش پیشنهادی از نظر آماری عملکردی مشابه با KDFS دارد و از نظر کاهش پارامترها نسبت به دو روش دیگر برتر است؛ به‌طوری‌که با حفظ دقتی معادل روش KDFS، موفق به کاهش ۲/۴۹ درصدی بیشتر در تعداد پارامترها شده است. اگرچه روش PDD در این مجموعه‌داده بیشترین میزان کاهش محاسبات را دارد، اما افت شدید دقت آن در

[1] Batch Size
[2] Learning rate
[3] Epochs
[4] Momentum
[5] Generalization
[6] t-test
[7] P-Value

حدود ٤ درصد نسبت به روش پیشنهادی، کارایی عملی آن را با چالش مواجه می‌کند.

روش KDFS برتری آماری جزئی و معناداری در مجموعه‌داده Real and Fake Faces-140k دارد. با این حال روش پیشنهادی با بهبود ۰/۹۷ درصدی در کاهش پارامترها و ۲/۱۵ درصدی در کاهش محاسبات نسبت به KDFS، تعادل[1] خوبی ایجاد کرده است.

در مجموعه‌داده Deepfake and Real Images، اگرچه روش PDD بالاترین دقت را دارد اما انحراف معیار بسیار بالای آن نشان دهنده نوسانات شدید عملکرد و عدم پایداری در شرایط مختلف است. در مقابل، روش پیشنهادی با انحراف معیار به‌مراتب کمتر و در عین حال افزایش ۳/٦۱ درصدی در کاهش پارامترها نسبت به PDD، عملکردی پایدارتر و قابل‌اعتمادتر دارد.

در مجموعه‌داده GRAVEX-200K روش پیشنهادی رویکردی تهاجمی‌تری در فشرده‌سازی شبکه دارد. هرچند دقت آن در مقایسه با KDFS کاهش یافته است، اما این روش موفق شده است بالاترین میزان کاهش پارامتر و محاسبات را در میان تمامی مجموعه‌داده‌ها به‌دست آورد. پس در کاربردهایی که محدودیت حافظه و توان مصرفی از اهمیت بالاتری برخوردار است، روش پیشنهادی می‌تواند انتخابی مناسب باشد.

تنها در مجموعه‌داده Deepfake-dataset روش KDFS در همه موارد برتر است. با این وجود، روش پیشنهادی در مقایسه با روش PDD، هم از نظر دقت با اختلاف ٤/۱۲ درصد و هم از نظر کاهش پارامترها عملکرد به‌مراتب بهتری دارد.

در مجموع، نتایج نشان می‌دهد که روش پیشنهادی در اکثر مجموعه‌داده‌ها با دستیابی به بالاترین نرخ فشرده‌سازی و حفظ پایداری عملکرد موفق به ایجاد تعادلی بهینه میان دقت و هزینه‌های محاسباتی شده است؛ این ویژگی، روش پیشنهادی را به گزینه‌ای مناسب برای پیاده‌سازی در سامانه‌های با محدودیت منابع سخت‌افزاری تبدیل می‌کند.

شکل ۲ منحنی ROC به همراه مقدار سطح زیر منحنی[2] را برای مجموعه‌داده Real vs Fake Faces-10k نشان میدهد. روش پیشنهادی با سطح زیر منحنی برابر با ۰/۹۹ بهترین عملکرد را دارد، در مقابل، روش PDD مقدار سطح زیر منحنی برابر با ۰/۹٦۱ دارد و کمی ضعیف تر از دو روش ذکر شده عمل می کند.

از منظر بررسی پیچیدگی محاسباتی، تفاوت روش پیشنهادی با KDFS در بخش سوم تابع هزینه است. در دو بخش اول تابع هزینه در هر دو روش پیچیدگی محاسبتی به صورت خطی است.پیچیدگی اصلی روش پیشنهادی در محاسبه ماتریس همبستگی است که برای L لایه با تعداد C فیلتر و W تعداد وزن در هر فیلتر، پیچیدگی زمانی $O(L.C^2.W)$ دارد. در مقابل، روش KDFS از یک تابع جریمه بر اساس FLOPs در بخش سوم تابع هزینه استفاده می‌کند که پیچیدگی زمانی آن $O(L.C)$ است که خطی و براساس تعداد لایه‌ها است.

روش پیشنهادی به دلیل نیاز به نگهداری ماتریس همبستگی در حافظه، پیچیدگی فضایی $O(L.C^2)$ دارد. در حالی که روش KDFS تنها نیازمند ذخیره مقادیر FLOPs برای هر لایه است.پس پیچیدگی فضایی آن $O(L.C)$ است.

جدول (۵): مقایسه میزان دقت، درصد کاهش محاسبات(FLOPs) و پارامترها در روش KDFS، روش پیشنهادی و روش PDD با مجموعه‌داده‌های متفاوت. از معماری ResNet50 با ۲۳/۵ میلیون پارامتر و ۵۳۹۰ مگافلاپس به عنوان شبکه معلم در روش پیشنهادی و KDFS استفاده شد. از معماری ResNet18 برای روش PDD با ۱۱/۱۷ میلیون پارامتر ۲۳۸۱ مگافلاپس استفاده شد. اعداد پررنگ به معنی برتری روش پیشنهادی است.

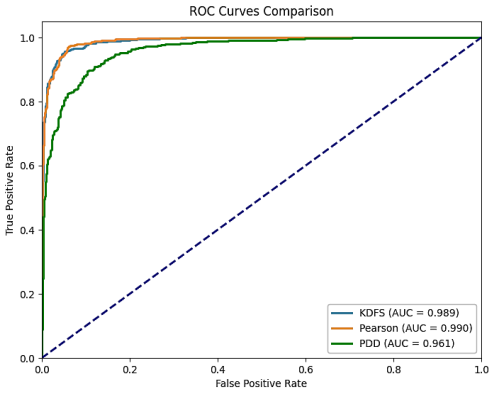
روش	مجموعه‌داده	نرخ فشرده سازی	دقت آزمون (%)	P-value	افزایش دقت(%)	تعداد پارامتر (میلیون)	فلاپس (مگا فلاپس)	کاهش پارامتر (%)	کاهش FLOPs (%)
KDFS	Real vs Fake Faces-10k	۰/۸	۹۴/۷۱±۰/۳۱	۰/۸	-۱/۲۹	۴/۲۸	۱۱۳۹/۸	۸۱/۸۱	۷۸/۸۵
پیشنهادی		-	۹۴/۷۳±۰/۲۳		**-۱/۲۷**	**۳/۶۹**	**۱۱۰۰/۵**	**۸۴/۳**	۷۹/۵۸
PDD		-	۹۰/۸۲±۰/۴۷	–	-۶/۰۷	۲/۰۴	۴۱۴/۹۲	۸۱/۶۹	۸۲/۵۸
KDFS	140k Real and Fake Faces	۰/۷	۹۹/۲۵±۰/۰۳	< ۰/۰۰۱	۴/۵۲	۳/۶۸	۹۸۲/۶۱	۸۴/۳۶	۸۱/۷۷
پیشنهادی		-	۹۹/۰۴±۰/۰۳		۴/۳۱	**۳/۴۵**	**۸۶۶/۹۷**	**۸۵/۳۳**	**۸۳/۹۲**
PDD		-	۹۲/۶۸±۰/۵۶	–	-۲/۶۱	۲/۳۶	۴۱۴/۷۳	۸۱/۷۸	۸۲/۵۹
KDFS	deepfake and real images	۰/۸	۸۸±۰/۳۵	۰/۰۲۸	۲/۳۶	۳/۳۲	۸۷۴/۱۷	۸۵/۸۹	۸۳/۷۸
پیشنهادی		-	۸۷/۴۱±۰/۷۵		۱/۷۷	**۳/۳۶**	**۸۶۶/۱۱**	۸۵/۷۲	**۸۳/۹۳**
PDD		-	۸۸/۴۹±۲/۵۶	–	۵/۷	۱/۹۸	۴۰۰/۰۵	۸۲/۲۸	۸۳/۲
KDFS	GRAVEX-200K	۰/۷	۹۱/۲±۰/۰۹	< ۰/۰۰۱	-۲/۲	۳/۷۲	۱۰۰۱/۵	۸۴/۱۹	۸۱/۴۲
پیشنهادی		-	۸۸/۸۴±۰/۰۴		**-۴/۵۶**	**۲/۹۵**	**۷۱۱/۰۲**	**۸۷/۴۴**	**۸۶/۸۱**
PDD		-	۸۲/۱۷±۰/۱۱	–	-۱۱/۹	۲/۰۴	۴۱۳/۹۶	۸۱/۷۲	۸۲/۶۲
KDFS	Deepfake-dataset	۰/۸	۹۲/۷±۰/۱۵	< ۰/۰۰۱	-۱/۲۴	۳/۰۳	۷۷۲/۷۶	۸۷/۱۲	۸۵/۶۶
پیشنهادی		-	۹۰/۷۹±۰/۱۸		-۳/۱۵	۳/۳۱	۸۹۰/۴۹	۸۵/۹	۸۳/۴۸
PDD		-	۸۶/۶۷±۰/۲	–	-۷/۴	۱/۹۷	۴۰۲/۵۱	۸۲/۳	۸۳/۱

[1] Trade-off

[2] Area Under Curve (AUC)

در روش PDD برای هر لایه پیچشی یک ماسک با اندازه برابر تعداد کانال‌های آن لایه طراحی می‌شود. هر عنصر ماسک با استفاده از تابع ApproxSign باینری می‌شود و به یک کانال خاص نگاشت می‌شود. در هر مرحله از فرآیند آموزش این ماسک‌ها روی تمام کانال‌ها اعمال می‌شوند. پس پیچیدگی زمانی آن برابر با $O(L.D)$ است. همچنین پیچیدگی فضایی آن برابر با $O(L.D)$ می‌باشد. که ناشی از ذخیره پارامترهای ماسک به ازای هر کانال است.

بنابراین از نظر تئوری روش پیشنهادی پرهزینه است. البته با استفاده از روش‌های بهینه‌سازی، زمان آموزش روش پیشنهادی تا حدی کاهش یافت.



شکل (۲): منحنی ROC به همراه مقادیر سطح زیر منحنی برای دسته real در سه روش ذکر شده برای مجموعه‌داده Real vs Fake Faces-10k

با توجه به جدول ٦، روش PDD زمان آموزش بسیار کمتری دارد اما نتایج دقت آن در جدول ۵ مطلوب نیست. روش پیشنهادی به طور متوسط حدود ۱۲ درصد افزایش زمان آموزش نسبت به روش KDFS دارد. در مجموع، اگرچه روش پیشنهادی زمان آموزش طولانی‌تری دارد، اما با توجه به مزایای قابل توجه آن از جمله کاهش بیشتر پارامترها (به‌طور متوسط ۲ تا ۳ درصد)، کاهش بیشتر محاسبات (به‌طور متوسط ۲ تا ۵ درصد) و پایداری خوب، این هزینه اضافی از نظر عملکردی به‌عنوان تعادلی مناسب است.

شکل ۳ درصد فیلترهای فعال در هرلایه را برای دو روش ذکر شده نشان می‌دهد. به طور متوسط، مقدار ماسک برای روش ذکر شده نشان می‌دهد. به طور متوسط، مقدار ماسک برای روش پیشنهادی ۰/۱۷ است، در حالی که این مقدار برای روش KDFS به ۰/۲۹ می‌رسد. این اختلاف نشان می‌دهد که روش پیشنهادی شبکه‌ای فشرده‌تر تولید می‌کند. در لایه‌های ابتدایی روش KDFS به صورت غیریکنواخت هرس انجام می‌شود در حالی که روش پیشنهادی رویکردی پایدارتر و محافظه‌کارانه‌تر دارد که با اهمیت حفظ ویژگی‌های سطح پایین در این لایه‌ها همخوانی دارد.

در ادامه به منظور بررسی تاثیر ضریب λ_3 و تحلیل حساسیت روش پیشنهادی، مطالعه فرسایش[1] انجام شد. با حذف ضریب λ_3 عملکرد شبکه به شدت کاهش می‌یابد و کاهش پارامتر و هزینه محاسباتی حدود ۶۰ درصد می‌شود. این نتیجه نشان دهنده آن است که ضریب λ_3 و به عبارتی زیان همبستگی نقش مهمی در فرایند هرس دارند و بدون آن نتیجه مطلوب در هرس حاصل نمی‌شود.

جدول (٦): مقایسه زمان آموزش روش‌های پیشنهادی و KDFS وPDD

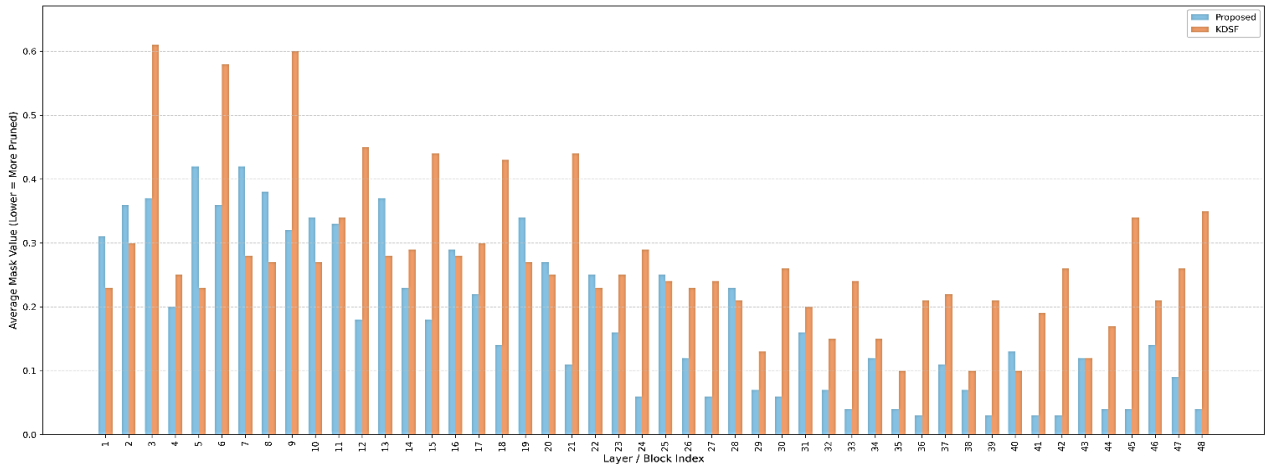
زمان آموزش	مجموعه داده	روش
۴ ساعت و ۲۲ دقیقه	Real vs Fake Faces-10k	KDFS
۴ ساعت و ۴۹ دقیقه		پیشنهادی
۱ ساعت و ۸ دقیقه		PDD
۱۱ ساعت و ۸ دقیقه	140k Real and Fake Faces	KDFS
۱۲ ساعت و ۳۸ دقیقه		پیشنهادی
۷ ساعت و ۴۵ دقیقه		PDD
۶ ساعت و ۵۷ دقیقه	deepfake and real images	KDFS
۸ ساعت و ۳۸ دقیقه		پیشنهادی
۸ ساعت و ۲۸ دقیقه		PDD
۱۵ ساعت و ۲۴ دقیقه	GRAVEX-200K	KDFS
۱۵ ساعت و ۴۵ دقیقه		پیشنهادی
۱۰ ساعت و۳۴ دقیقه		PDD
۱۱ ساعت و۵۴ دقیقه	Deepfake-dataset	KDFS
۱۳ ساعت و ۱۰ دقیقه		پیشنهادی
۱۱ ساعت و۲ دقیقه		PDD

همچنین ضریب λ_3 با مقدار ۵، یعنی ۱۰ برابر مقدار فعلی، اعمال شد. مشاهده شد که دقت آزمون و میزان هرس حاصل شده کمی پایین‌تر از مقدار فعلی λ_3 است. مقدار λ_3=۰/۵ نقطه بهینه‌ای است که بالاترین میزان کاهش پارامتر و محاسبات را با دقتی بالا نتیجه می‌دهد. تفاوت اندک در میزان هرس نشان می‌دهد که روش پیشنهادی در برابر تغییرات ضریب λ_3 در این بازه به‌نسبت پایدار است. نتایج این بخش در جدول ۷ آمده است.

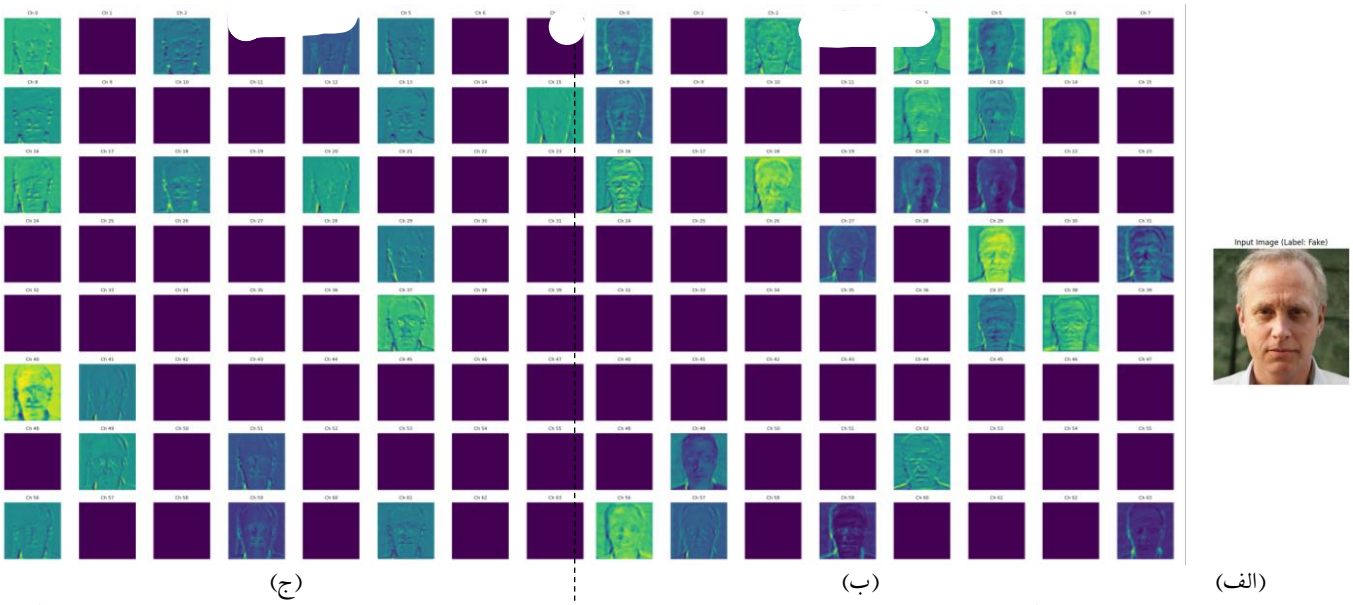
شکل ۴ نقشه‌های ویژگی لایه دوم در بلوک اول را نشان می‌دهد.خانه‌های بنفش رنگ به معنی کانال‌های غیرفعال و رنگ‌های روشن به معنی شدت فعال بودن کانال‌ها است. همانطورکه مشاهده می‌شود در روش KDFS نقشه‌های ویژگی شباهت بالایی به یکدیگر دارند، وضوح مناسبی ندارند و در بیشر موارد کلیات چهره را استخراج می‌کنند. این حجم از افزونگی کارایی شبکه را کاهش می‌دهد.در مقابل، روش پیشنهادی تنوع بیشتری در نقشه‌های ویژگی دارد و اجزای چهره مانند دهان، بینی و چشم‌ها را با وضوح خوبی مشخص می‌کند. یعنی ضریب همبستگی پیرسون توانسته است به خوبی فیلترهای زائد و مشابه را حذف کند و فیلترهایی که ویژگی‌های متنوعی را استخراج می‌کنند را نگه دارد.

جدول۸ نتایج قدرت تعمیم شبکه‌های فشرده شده (دانش‌آموز) روی دیگر مجموعه‌داده‌ها را نشان می‌دهد. نام‌گذاری این شبکه‌ها براساس روش و مجموعه‌داده‌هایی که با آن‌ها آموزش دیده‌اند انجام شده است.

[1] Ablation Study



شکل (۳): درصد فیلترهای فعال در هر لایه در مجموعه‌داده Real vs Fake Faces-10k . این مقادیر در بهترین دور از مرحله اعتبارسنجی به دست آمدند.



(ج) (ب) (الف)

شکل(۴) : مقایسه نقشه‌های ویژگی در لایه دوم از بلوک اول در مجموعه‌داده Real vs Fake Faces-10k . الف)تصویر ورودی ب)نقشه‌های ویژگی شبکه فشرده شده حاصل از روش پیشنهادی ج)نقشه‌های ویژگی شبکه فشرده شده حاصل از روشKDFS

به منظور کاهش تعداد اجراها و صرفه جویی در زمان، از آزمون مک‌نمار برای بررسی تفاوت معنادار روش‌ها استفاده شد. از آنجایی که روش پیشنهادی و KDFS دقت‌های مشابه در این جدول دارند این آزمون بر روی این دو روش اعمال شد. در ادامه به بررسی جدول (۸) می‌پردازیم.

شبکه Proposed-compressed-10K در تمامی مجموعه‌داده‌ها قدرت تعمیم بالاتری نسبت به شبکه KDFS-compressed-10K از خود نشان می‌دهد. نمونه‌ای از جدول آزمون مک‌نمار بر روی مجموعه‌داده Deepfake-dataset در جدول ۹ آمده است. براساس مقادیر جدول، مقدار احتمال به دست آمده کمتر از ۰/۰۵ شد که نشان دهنده تفاوت معنادار است.

نتایج ماتریس توافقی نشان می‌دهد که اگرچه هر دو شبکه در بخش عمده‌ای از داده‌ها (۲۶۸۳۰ نمونه صحیح و ۱۶۴۳ نمونه غلط) رفتار یکسانی داشته‌اند، اما تحلیل موارد اختلاف بیانگر برتری شبکه پیشنهادی است؛ به‌طوری‌که شبکه پیشنهادی موفق شده است ۱۲۹۰ نمونه را که توسط KDFS به اشتباه طبقه‌بندی شده بود، به درستی تشخیص دهد، در حالی که شبکه پایه در ۱۱۴۲ مورد نسبت به شبکه پیشنهادی عملکرد بهتری داشته است. این برتری در تعداد اصلاحات نشان‌دهنده ارتقای دقت کلی در روش پیشنهادی نسبت به روش KDFS است.

در سایر مجموعه‌داده‌ها، هرچند در برخی موارد دقت روش پیشنهادی از KDFS بالاتر است، اما این تفاوت از نظر آماری معنادار نیست. شبکه‌های فشرده‌شده حاصل از روش KDFS در اکثر موارد دقت بالاتری نسبت به روش پیشنهادی دارند که این تفاوت نیز براساس نتایج مک‌نمار معنادار است.

در نهایت، روش PDD در هیچ یک از موارد نتوانست از نظر قدرت تعمیم با دو روش دیگر رقابت کند و عملکرد ضعیف‌تری را به ثبت رساند.

جدول (۷): مطالعه فرسایش روش پیشنهادی با دو مقدار λ_3

λ_3	دقت آزمون (%)	افزایش دقت(%)	تعداد پارامتر (میلیون)	فلاپس (مگا فلاپس)	کاهش پارامتر (%)	کاهش FLOPs (%)
۰	۷۸/۲	ـ ۱۷/۸	۹/۳۱	۲۱۷۰/۲۹	۶۰/۴۱	۵۹/۷۳
۵	۹۳/۶	ـ ۲/۴	۳/۹	۱۱۰۲/۳۷	۸۳/۴	۷۹/۵۵

جدول (۸): مقایسه قدرت تعمیم روش پیشنهادی، KDFS و PDD روی مجموعه‌داده‌های متفاوت. اعداد پررنگ به معنی برتری روش پیشنهادی است.

شبکه هرس و تقطیر شده	Real vs Fake Faces-10k	140k Real and Fake Faces	deepfake and real images	GRAVEX-200K	Deepfake-dataset
KDFS-compressed-10K		۹۵/۴۱	۸۰/۶۵	۸۸/۲۴	۹۰/۵۱
Proposed -compressed-10K		**۹۵/۶۸**	**۸۲/۸۲**	**۸۸/۳۲**	**۹۰/۹۹**
PDD-compressed-10k		۸۹/۹۳	۷۹/۱۴	۸۵/۰۰	۸۸/۷۴
KDFS-compressed-140K	۹۹/۷۳		۸۰/۸۳	۹۱/۴۲	۹۳/۷۴
Proposed -compressed-140K	۹۹/۴۷		۸۳/۸۱	۹۱/۱۲	۹۳/۴
PDD-compressed-140k	۶۸/۵۳		۷۰/۵۹	۷۴/۷۲	۷۳/۴۱
KDFS-compressed-190K	۸۱/۲۷	۹۱/۸۹		۸۶/۴۵	۹۰/۸۶
Proposed -compressed-190K	۸۰/۰۷	۹۲/۰۱		۸۶/۱۰	۹۰/۷۸
PDD-compressed-190k	۶۷/۸۷	۶۶/۰۲		۷۱/۳۹	۶۸/۲۸
KDFS-compressed-200K	۹۴/۷	۹۷/۷۲	۸۲/۳۰		۹۱/۸۴
Proposed -compressed-200K	۹۳/۵۳	۹۴/۴۲	۸۰/۱۲		۸۸/۸۰
PDD-compressed-200k	۶۷/۴۷	۷۹/۳۲	۷۸/۸۹		۸۰/۱۵
KDFS-compressed-330K	۹۵/۵۳	۹۶/۶۴	۸۹/۰۸	۸۸/۱۸	
Proposed -compressed-330K	۹۳/۱۳	۹۵/۸۸	۸۸/۴۰	۸۸/۷۷	
PDD-compressed-330k	۷۵/۵۶	۷۳/۳۰	۶۹/۳۴	۷۳/۰۶	

جدول (۹): نتایج آزمون مک‌نمار برای مقایسه شبکه‌های KDFS-compressed-10K و Proposed-compressed-10K روی مجموعه‌داده Deepfake-dataset.

	Proposed Correct	Proposed Wrong
KDFS Correct	۲۶۸۳۰	۱۱۴۲
KDFS Wrong	۱۲۹۰	۱۶۴۳

۵- جمع‌بندی

این پژوهش رویکردی برای فشرده‌سازی شبکه‌های عصبی پیچشی در تشخیص جعل عمیق چهره ارائه داد و با بهره‌گیری از تقطیر دانش و هرس فیلترها، کارایی محاسباتی بهبود یافت. ایده اصلی، استفاده از ضریب همبستگی پیرسون است که فیلترهای زائد را با دقت بالا شناسایی و حذف می‌کند. با استفاده از احتمال هرسِ هر فیلتر و امتیاز همبستگی بین فیلترها، زیان همبستگی روی تمام لایه‌ها به دست می‌آید که این زیان همراه با سایر زیان‌ها به روز رسانی می‌شود تا کمینه شود. این روش نه‌تنها تعداد پارامترها و هزینه محاسباتی را کاهش داد، بلکه در یکی از مجموعه‌داده‌ها دقت روش پایه را به‌طور کامل حفظ کرد و در سایر موارد تنها کاهش اندکی در دقت نشان داد. از نظر قدرت تعمیم نیز در یک مورد بر روش پایه برتری داشت و در بقیه موارد،دقت بسیار نزدیک به روش پایه بود.

این رویکرد امکان استفاده از شبکه‌های سبک را در محیط‌های محدود از نظر منابع محاسباتی ممکن می‌سازد. شبکه‌های فشرده‌شده حاصل از این روش می‌تواند در سامانه‌های تشخیص هویت بیومتریک، نظارت امنیتی و برنامه‌های موبایلی که نیاز به پردازش سریع و کم‌مصرف دارند به کار گرفته شود.

مراجع

[1] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A.N. Khan, “DeepFake detection for human face images and videos: A survey,”in IEEE Access, vol. 10, pp. 18757–18775, 2022.

[2] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network”, Mar. 2015, arxiv.org/abs/1503.02531.

[3] L. Wang and K. -J. Yoon, "Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3048-3068,June 2022.

[4] J. Park and A. No, “Prune Your Model Before Distill It,” Jul.2022, arxiv.org/abs/2109.14960.

[5] S. Lin et al., “Filter Pruning for Efficient CNNs via Knowledge- driven Differential Filter Sampler,” Jul.2023,arxiv.org/abs/2307.00198.

[6] M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," IEEE Access, vol. 10, pp. 25494-25513, 2022.

[7] Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” Jun 2018,arxiv.org/abs/1806.02877.

[8] T. Jung, S. Kim, and K. Kim, “DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern,” IEEE Access, vol. 8, pp. 83144–83154, 2020.

[9] X. Yang, Y. Li, and S. Lyu, “Exposing Deep Fakes Using Inconsistent Head Poses,” Nov.2018, arxiv.org/abs/1811.00661.

[10] T. Wang, X. Liao, K. P. Chow, X. Lin, and Y. Wang, "Deepfake detection: A comprehensive survey from the reliability perspective," ACM Computing Surveys, vol. 57, no. 3, Mar. 2025.

[11] M. Gachchannavar, J. R. Naveenkumar, and R. Velangi, "A survey an optimized dense CNN model for recognizing deepfake images," International Journal for Multidisciplinary Research, vol. 6, no. 4, Aug. 2024.

[12] M. Kim, S. Tariq, and S. S. Woo, "FReTAL: Generalizing deepfake detection using knowledge distillation and representation learning," in Proc. of IEEE/CVF Computer Vision and Pattern Recognition Workshop (CVPRW), Nashville, TN, USA, pp. 1001–1012, 2021.

[13] M. Kim, S. Tariq, and S. S. Woo, "CoReD: Generalizing fake media detection with continual representation using distillation," in Proc. 29th ACM International Conference Multimedia (MM), Chengdu, China, pp. 337–346 , Oct. 2021.

[14] C. Zhou et al., “Two-in-one Knowledge Distillation for Efficient Facial Forgery Detection,” Feb.2023, arxiv.org/abs/2302.10437.

[15] Y. Lin, H. Chen, B. Li, and J. Wu, "Towards generalizable deepfake face forgery detection with semi-supervised learning and knowledge distillation," in Proc. IEEE International Conference on Image Processing (ICIP), Bordeaux, France, pp. 576–580, Oct. 2022.

[16] C. Wang, L. Meng, Z. Xia, N. Ren, and B. Ma, "Cross-domain deepfake detection based on latent domain knowledge distillation," IEEE Signal Process. Lett., vol. 32, pp. 896–900, 2025.

[17] B. M. Le and S. S. Woo, “ADD: Frequency Attention and Multi-View based Knowledge Distillation to Detect Low- Quality Compressed Deepfake Images,”Dec.2021, arxiv.org/abs/2112.03553.

[18] L. Al Amin, Md. I. Hossain, T. T. Nguyen, T. Jahan, M. Islam, and F. Quader, “Uncovering Critical Features Deepfake Detection through the Lottery Ticket Hypothesis,”,Jul. 2025 ,arxiv.org/abs/2507.15636.

[19] L. Chen, Y. Chen, J. Xi, and X. Le, "Knowledge from the original network: Restore a better pruned network with knowledge distillation," Complex Intell. Syst., vol. 8, no. 2, pp. 709–718, Apr. 2022.

[20] C. Deng, D. Jing, Z. Ding, and Y. Han, "Sparse channel pruning and assistant distillation for faster aerial object detection," Remote Sens., vol. 14, no. 21, p. 5347, 2022.

[21] N. Aghli and E. Ribeiro, "Combining weight pruning and knowledge distillation for CNN compression," in Proc.IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Nashville, TN, USA, pp. 3185–3192, 2021.

[22] D. Chen et al., “EPSD: Early Pruning with Self-Distillation for Efficient Model Compression,” Jan.2024, arxiv.org/abs/2402.00084.

[23] Y. Liu, K. Fan, and W. Zhou, "Iterative filter pruning with combined feature maps and knowledge distillation," International Journal of Machine Learning and Cybernetics, vol. 16, no. 3, pp. 1955–1969, Mar. 2025.

[24] X. Dan, Y. Zhang, L. Li, and H. Wang, "PDD: Pruning neural networks during knowledge distillation," Cognitive Computation, vol. 16, no. 6, pp. 3457–3467, Nov. 2024.

[25] Kumar, A., Yin, B., Shaikh, A. M., et al., “CorrNet: Pearson Correlation Based Pruning for Efficient Convolutional Neural Networks,” *International Journal of Machine Learning and Cybernetics*, vol. 13, pp. 3773–3783, Springer, 2022.

[26] Singh, P., Verma, V., Rai, P., Namboodiri, V., “Leveraging Filter Correlations for Deep Model Compression,” *Proceedings of the Conference|Journal*, pp. 824–833, 2020.

[27] Xu, J., Liu, C., Qian, H., Zhu, Q., Chen, J., “Pruning of the Object Detection Model Based on Multi-level Feature De-correlation,” *Proceedings of the 2nd International Conference on Artificial Intelligence of Things and Computing (AITC '25)*, pp. 245–250, New York, USA, 2025.

[28] Li, G., Shao, H., Deng, X., Jiang, Y., “Adaptive Convolutional Network Pruning through Pixel-Level Cross-Correlation and Channel Independence for Enhanced Model Compression,” *Engineering Applications of Artificial Intelligence*, vol. 154, 2025.

[29] Wang, W., Fu, C., Guo, J., Cai, D., He, X., "COP: Customized Deep Model Compression via Regularized Correlation-Based Filter-Level Pruning," *Proceedings of the International Joint Conference on Artificial Intelligence* (*IJCAI 2019*), pp. 3785–3791, 2019.

[30] He, J., Chen, B., Ding, Y., Li, D., "Feature Variance Ratio-Guided Channel Pruning for Deep Convolutional Network Acceleration," *Computer Vision – ACCV 2020*, Lecture Notes in Computer Science, vol. 12625, Springer, 2021.

[31] M. Nadeem, R. Imam, R. Al-Refai, M. Chkir, M. Hoda, and A. El Saddik, "EVOKE: Emotion Enabled Virtual Avatar Mapping Using Optimized Knowledge Distillation," Jan. 2024, arxiv.org/abs/2401.06957.

[32] https://www.kaggle.com/datasets/sachchitkunichetty/rvf10k

[33] https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces/data

[34] https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images/data

[35] https://www.kaggle.com/datasets/muhammadbilal6305/200k-real-vs-ai-visuals-by-mbilal

[36] https://www.kaggle.com/datasets/tusharpadhy/deepfake-dataset

سارا عسکری همت در سال ۱۴۰۱ مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم افزار از دانشگاه شهید باهنر کرمان دریافت کرد. او در حال حاضر دانشجوی کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز در دانشگاه شهید باهنر کرمان است. زمینه‌های پژوهشی مورد علاقه او یادگیری عمیق و پردازش تصویر است.

مهدی افتخاری مدرک کارشناسی خود را در سال ۱۳۷۹ در رشته مهندسی کامپیوتر گرایش سخت افزار از دانشگاه شیراز و مدرک کارشناسی ارشد و دکترای خود را به ترتیب در سال‌های ۱۳۸۲ و ۱۳۸۶ در رشته مهندسی کامپیوتر گرایش هوش مصنوعی از همان دانشگاه اخذ کرد. وی از سال ۱۳۸۶ تا کنون عضو هیأت علمی بخش مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه شهید باهنر کرمان است و در سال ۱۴۰۰ به مرتبه استادی ارتقاء پیدا کرده است. حوزه‌های تخصصی پژوهش ایشان یادگیری ماشین، یادگیری عمیق و مجموعه‌ها و سیستم‌های فازی است. وی تا کنون بیش از ۱۶۰ مقاله علمی در نشریات و کنفرانس‌های معتبر داخلی و خارجی به چاپ رسانیده است.