چکانش دانش چندمرحله‌ای بر پایه بازنمایی‌های مبتنی بر زیرفضا

مجید سپهوند[1]

چکیده

چکانش دانش با هدف ساخت مدل‌های دانش‌آموز کم‌حجم تحت هدایت مدل‌های معلم بزرگ‌مقیاس به کار می‌رود و از این طریق امکان استفاده از شبکه‌های کارآمدتر را فراهم می‌سازد. با وجود این، فاصله عملکردی میان معلم و دانش‌آموز همچنان چشمگیر است، زیرا بخش مهمی از دانش موجود در معلم به‌طور کامل به دانش‌آموز منتقل نمی‌شود. برای حل این مشکل، در این مقاله مدل چکانش دانش پیشنهادی چندمرحله‌ای پیشنهاد شده که به‌صورت هم‌زمان دانش را از مسیر هم‌ترازی ویژگی‌ها و تطبیق لاجیت‌ها منتقل کرده و وابستگی‌های میان‌لایه‌ای شبکه را نیز مدل‌سازی می‌کند. این رویکرد سیگنال‌های نظارتی دقیق‌تری تولید کرده و دانش‌آموز را قادر می‌سازد بازنمایی‌های معلم را کامل‌تر فرا بگیرد. روش پیشنهادی از سه مؤلفه مکمل تشکیل شده است: ماژول توجه سه‌بعدی که نواحی مهم فضایی و کانالی را برجسته می‌کند؛ ماژول ماسک خصمانه که زیرفضاهای مفید و غیرمفید را به‌صورت تطبیقی جدا می‌سازد؛ و ماژول تنظیم فضای کروی که توزیع ویژگی‌های معلم و دانش‌آموز را روی ابرکره هم‌راستا می‌کند. ترکیب این سه ماژول باعث می‌شود دانش‌آموز بتواند فضای ویژگی و فضای خروجی معلم را دقیق‌تر کاوش کند و به بازنمایی‌های عام‌تر و پایدارتر دست یابد. آزمایش‌های گسترده روی CIFAR-100، STL-10 و TinyImageNet نشان می‌دهند که روش پیشنهادی در بیشتر پیکربندی‌ها عملکردی بهتر از روش‌های پیشرفته موجود ارائه می‌کند.

کلید واژه ها

چکانش دانش، بازنمایی آگاه از زیرفضا، چنددانه‌گی، توجه سه‌بعدی، ماسک خصمانه، تنظیم فضای کروی

۱- مقدمه

با افزایش سریع پیچیدگی وظایف و بزرگ‌تر شدن مجموعه‌های داده، بهبود عملکرد مدل‌های یادگیری عمیق بیشتر از گذشته به سمت استفاده از شبکه‌هایی با عمق و عرض بیشتر حرکت کرده است. این موضوع هرچند منجر به قدرت بالاتر می‌شود، اما هم‌زمان باعث رشد چشمگیر تعداد پارامترها و هزینه‌های محاسباتی نیز می‌شود.

مقاله در تاریخ ۱۴ آذر ماه ۱۴۰۴ دریافت شد.

[1] گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اراک
رایانه: m-sepahvand@araku.ac.ir

از این‌رو، نیاز به مدل‌های سبک و کارآمد افزایش یافته و فشرده‌سازی مدل‌ها به یکی از محورهای مهم تحقیقاتی تبدیل شده است، زیرا استقرار شبکه‌های بزرگ بر روی دستگاه‌های با منابع محدود همچنان چالش‌برانگیز است. برای مقابله با این چالش، هینتون و همکاران [۱] نخستین بار چکانش دانش[1] را مطرح کردند که طی آن دانش یک مدل معلم قدرتمند به یک مدل دانش‌آموز سبک منتقل می‌شود. پس از آن، نسخه‌های بهبودیافته و متنوعی از چکانش دانش ارائه شده‌اند [۲ـ۶]. چکانش دانش همچنین در حوزه‌های مختلفی از جمله طبقه‌بندی تصویر، شناسایی چهره [۷] و تحلیل تصاویر پزشکی [۸] کاربردهای گسترده‌ای پیدا کرده است. این روش با کوچک‌سازی مدل در حالی که بخش عمده‌ای از توان پیش‌بینی آن حفظ می‌شود، راهی عملی برای استفاده از شبکه‌های

[1] Knowledge Distillation

قدرتمند در محیط‌های کم‌منبع ارائه می‌دهد. روش‌های چکانش دانش معمولاً در سه دسته جای می‌گیرند: رویکردهای مبتنی بر لاجیت[1] [۹]، مبتنی بر ویژگی[2] [۱۰] و مبتنی بر شباهت [۱۱] که هرکدام روی جنبه خاصی از انتقال دانش تمرکز دارند.

روش‌های رایج چکانش دانش معمولا یا از طریق خروجی لاجیت‌ها یا با هم‌ترازسازی نمایش‌های ویژگی[3] در لایه‌های مشخص دانش را منتقل می‌کنند. رویکردهای مبتنی بر توجه مانند مدل AT [۱۲] و مدل SKD [۱۳] نیز با تقلید نقشه‌های توجه یا لاجیت‌های نرم‌شده، نواحی مهم را برجسته می‌کنند. با این حال، این روش‌ها غالباً در سطح محلی عمل کرده و وابستگی‌های ساختاری میان سطوح مختلف بازنمایی را مدل نمی‌کنند؛ در نتیجه انتقال دانش به‌صورت کامل و سلسله‌مراتبی انجام نمی‌شود.

رویکردهای مبتنی بر ماسک[4] برای حذف اطلاعات زائد پیشنهاد شده‌اند، اما با محدودیت‌هایی مواجه‌اند. برخی روش‌ها مانند MasKD [۱۴] به دقت مکانیزم انتخاب وابسته‌اند و ممکن است بخش‌های مفید را نیز حذف کنند. روش‌هایی نظیر CRLD [۱۵] و CRD [۱۶] با تحمیل سازگاری میان نماها یا هم‌ترازی کنتراستی، به کیفیت داده‌افزایی، اندازه دسته[5] یا تنظیم دما وابسته بوده و در مقیاس‌پذیری با چالش روبه‌رو هستند. همچنین SKD [۱۷] با نرمال‌سازی کروی لاجیت‌ها شکاف ظرفیت را کاهش می‌دهد، اما همچنان وابستگی‌های ساختاری عمیق در ویژگی‌های میانی را به‌طور کامل مدل نمی‌کند.

با وجود پیشرفت‌های اخیر، بسیاری از روش‌های چکانش دانش همچنان بر نظارت تک‌سطحی تکیه دارند؛ بدین معنا که یا لاجیت‌ها را تطبیق می‌دهند یا ویژگی‌های میانی را هم‌تراز می‌کنند، بدون آنکه تعامل میان سطوح مختلف بازنمایی را به‌صورت منسجم مدل کنند. در حالی‌که شبکه‌های عمیق ذاتاً ساختاری سلسله‌مراتبی دارند، انتقال دانش زمانی کامل‌تر خواهد بود که نشانه‌های محلی، تفکیک زیرفضایی و ساختار سراسری به‌طور هم‌زمان در نظر گرفته شوند. بر این اساس، در این پژوهش چارچوبی چندمرحله‌ای پیشنهاد می‌شود که سه مؤلفه شناخته‌شده در ادبیات، شامل توجه فضایی ـ کانالی[6]، تقسیم زیرفضاهای معنایی[7]، و تنظیم هندسی[8] در مقیاس سراسری[9] ارائه می‌دهد. ترکیب این مؤلفه‌ها امکان مدل‌سازی وابستگی‌های سلسله‌مراتبی میان لایه‌ها را فراهم ساخته و انتقال دانش را به‌صورت سازگارتر در سطوح مختلف هدایت می‌کند.

در ابتدا، ویژگی‌های استخراج‌شده از معلم و دانش‌آموز گردآوری می‌شوند. برای برجسته‌سازی نواحی مهم و افزایش توجه به ویژگی‌های مفید، یک ماژول توجه سه‌بعدی[10] روی این ویژگی‌ها اعمال می‌شود تا نقشه‌های توجه تولید شده و هم‌ترازی میان دو شبکه هدایت شود. پس از این مرحله، از ماژول ماسک خصمانه[11] استفاده می‌شود که ورودی آن ویژگی‌های دانش‌آموز و لاجیت‌های معلم هستند و یک ماسک نرم تولید می‌کند تا ویژگی‌های دانش‌آموز را به صورت تطبیقی به زیرفضاهای پر اهمیت و کم‌اهمیت تقسیم کند. برای هر زیرفضا، یک سر طبقه‌بندی[12] مستقل به کار گرفته شده و فرایند چکانش دانش در هر شاخه جداگانه انجام می‌شود. در ادامه، ما یک ماژول چکانش با تنظیم کروی[13] ارائه می‌کنیم که ویژگی‌های هر دو شبکه را روی یک ابرکره[14] نگاشت کرده، ماتریس‌های شباهت زوجی را محاسبه می‌کند و این توزیع‌ها را با یک تابع زیان سازگاری[15] با یکدیگر هم‌سو می‌سازد. این فرآیند قابلیت تشخیص میان نمونه‌ها را تقویت می‌کند و باعث می‌شود مدل دانش‌آموز ساختار کلی فضای ویژگی مدل معلم را بهتر تقلید کند.

از منظر هندسی، سه ماژول پیشنهادی در سطوح مکملی از فضای بازنمایی عمل می‌کنند. ماژول توجه سه‌بعدی با بازوزن‌دهی نورون‌سطحی، سیگنال‌های محلی و نواحی متمایزکننده را تقویت می‌کند؛ ماژول ماسک خصمانه با تفکیک تطبیقی زیرفضاها، ابعاد مؤثر و غیرمؤثر را از یکدیگر جدا می‌سازد؛ و ماژول تنظیم کروی با هم‌ترازسازی هندسه سراسری، ساختار روابط نمونه‌ها را تثبیت می‌کند. بدین ترتیب، چارچوب حاضر را می‌توان به‌عنوان فرآیند پالایش تدریجی دانش از سطح محلی به سطح سراسری تفسیر کرد، که هر مرحله خروجی مرحله پیشین را تکمیل و منظم‌تر می‌سازد.

به طور خلاصه، مهمترین نوآوری‌های این مقاله عبارت‌اند از:

- یک ماژول توجه سه‌بعدی معرفی شده که وزن‌های تطبیقی را برای تمام نورون‌ها در سطح لایه‌ها و کانال‌ها اختصاص می‌دهد؛ این امر موجب تقویت نواحی مهم و کاهش نویز شده و هم‌ترازسازی ویژگی‌ها میان معلم و دانش‌آموز را بهبود می‌دهد.
- ماژول ماسک خصمانه‌ای ارائه شده که ورودی آن با استفاده از Gumbel-Softmax به‌صورت قابل تفکیک نمونه‌برداری شده و به شکل تطبیقی به زیرفضاهای قوی‌تر و ضعیف‌تر تقسیم می‌شود. این ساختار زیرفضامحور ابعاد مفید را برجسته کرده و سیگنال‌های کم‌اهمیت را هنگام تطبیق لاجیت‌ها کاهش می‌دهد.

[1] Logit
[2] Feature
[3] Feature Representation
[4] Mask
[5] Batch
[6] Spatial–Channel Attention
[7] Semantic Subspaces
[8] Geometric Regularization
[9] Global Scale
[10] 3D Attention Module
[11] Adversarial Mask Module
[12] Classification Head
[13] Spherical Regularization
[14] Hypersphere
[15] Consistency Loss

- یک ماژول تنظیم کروی پیشنهاد شده که تفاوت‌های میان توزیع‌های ویژگی معلم و دانش‌آموز را روی ابرکره افزایش داده و هم‌زمان سازگاری هندسی را حفظ می‌کند؛ در نتیجه انتقال دانش مؤثرتر می‌شود.

ساختار مقاله در ادامه به این شرح است: در بخش ۲ کارهای مرتبط مرور شده است. در بخش ۳ مبانی اولیه تشریح شده سپس در بخش ۴ روش پیشنهادی ارائه شده است. در بخش ۵ نتایج آزمایشگاهی را تشریح کرده و نهایتا در بخش ۶ نتیجه گیری مقاله انجام شده است.

۲- کارهای مرتبط

۲-۱- چکانش دانش مبتنی بر لاجیت

چکانش مبتنی بر لاجیت یکی از مستقیم‌ترین منابع انتقال دانش در چکانش دانش محسوب می‌شود، زیرا لاجیت‌ها خروجی لایه نهایی مدل هستند. هینتون و همکاران [۱] نخستین بار از لاجیت‌ها برای چکانش دانش استفاده کردند و دانش پنهان مدل معلم را از طریق توزیع‌های نرم‌شده انتقال دادند برخی مطالعات با اصلاح توزیع لاجیت یا تنظیم دما تلاش کرده‌اند فرآیند چکانش را پایدارتر سازند، از جمله CTKD [۱۸]، NormKD [۱۹]، LSKD [۲۰] و TTM [۲۱] که با نرمال‌سازی، استانداردسازی یا تنظیم کلاس‌محور دما، اختلاف مقیاس و ناهمگنی توزیع را کاهش می‌دهند.

علاوه بر این، بازطراحی تابع زیان نیز مورد توجه قرار گرفته است؛ برای مثال DKD [۲, ۲۲] با تفکیک کلاس هدف و غیرهدف، DKL [۲۳] با بازفرمول‌بندی واگرایی KL divergence[1] و SDD [۲۴] با تجزیه لاجیت‌ها به مؤلفه‌های محلی و مکمل، تلاش کرده‌اند روابط میان‌کلاسی را دقیق‌تر مدل کنند.

در سطح ساختاری، برخی روش‌ها با معرفی معلم میانی یا سازوکارهای چندمنبعی مانند TAKD [۹]، DGKD [۲۵] و BAN [۲۶] شکاف ظرفیت را کاهش داده‌اند. همچنین رویکردهایی نظیر TeKAP [۲۷]، SKD [۱۷] و SSKD [۲۸] با بازسازی فضای ویژگی، هم‌ترازی زاویه‌ای یا ترکیب یادگیری خودنظارتی، به دنبال کاهش فاصله بازنمایی میان معلم و دانش‌آموز بوده‌اند.

۲-۲- دانش‌چکانش مبتنی بر ویژگی

در مقابل روش‌های مبتنی بر لاجیت، رویکردهای مبتنی بر ویژگی بر هم‌ترازی نمایش‌های میانی معلم و دانش‌آموز تمرکز دارند. مدل FitNets [۱۰] با معرفی «لایه راهنما» نخستین گام شاخص در این مسیر بود و AT [۱۲] با سازوکار توجه، نقشه‌های فضایی معلم را به دانش‌آموز منتقل کرد. در ادامه، پژوهش‌هایی مانند [۲۹] و [۳۰, ۳۱] چکانش را در امتداد بعد کانال و فضای زاویه‌ای توسعه داده و اهمیت روابط کانالی و هندسه زاویه‌ای را نشان دادند. همچنین روش‌هایی نظیر FAKD [۳۲]، FreeKD [۳۳]، NormKD [۳۴] و AlpKD [۳۵] با تقویت یا بازآرایی فضای ویژگی، ادغام چندلایه‌ای یا انتقال در دامنه‌های جایگزین، به دنبال غنی‌سازی سیگنال نظارتی بوده‌اند.

در سطح پیشرفته‌تر، رویکردهای سلسله‌مراتبی و رابطه‌ای مانند HMAT [۳۶]، MASCKD [۳۷] و MLKD [۳۸] با مدل‌سازی ناهمگنی میان لایه‌ها و بهره‌گیری از توجه چندسطحی تلاش کرده‌اند وابستگی‌های میان‌لایه‌ای را دقیق‌تر ثبت کنند. با این حال، هم‌ترازی مستقیم ویژگی‌ها همچنان با چالش‌هایی همراه است؛ زیرا نمایش‌های با ابعاد بالا معمولاً حاوی نویز و افزونگی‌اند و تطبیق یک‌به‌یک لایه‌ها قادر نیست تفاوت‌های معنایی و مقیاسی میان سطوح مختلف را به‌طور کامل منعکس کند. در نتیجه، ساختار سراسری[2] روابط نمونه‌ها اغلب به‌طور جامع مدل نمی‌شود و انتقال دانش می‌تواند سطحی یا حتی مستعد انتقال منفی باشد.

۲-۳- دانش‌چکانش مبتنی بر شباهت[3]

علاوه بر رویکردهای مبتنی بر لاجیت و ویژگی، روش‌های مبتنی بر شباهت با هدف انتقال روابط ساختاری استخراج‌شده توسط معلم توسعه یافته‌اند. این دسته از روش‌ها به‌جای تطبیق مستقیم خروجی، شباهت‌های زوجی یا ساختارهای گرافی را مدل می‌کنند تا اطلاعات مرتبه‌بالا، فشردگی درون‌کلاسی و جدایی میان‌کلاسی حفظ شود. از نخستین نمونه‌ها، بیم و همکاران [۱۱] ماتریس FSP روابط میان‌لایه‌ای را از طریق ضرب داخلی ویژگی‌ها هم‌تراز کرد. در ادامه، رویکردهایی مانند RelationKD [۳۹]، [۴۰]، [۴۱]، CIRKD [۴۲] و GIRD [۴۳] با استفاده از زیان‌های رابطه‌ای، مدل‌سازی گرافی و حافظه‌های کمکی، تلاش کرده‌اند هندسه فضای معلم را در سطح نمونه یا لایه بازسازی کنند. همچنین روش‌هایی نظیر DIST [۴۴]، IDD [۴۵] و AICSD [۴۶] و رویکردهای فضایی مانند PRRD [۴۷]، BCKD [۴۸] و LSCD [۴۹] ساختار میان‌کلاسی و وابستگی‌های چندمقیاسی را مورد توجه قرار داده‌اند. افزون بر این، CRLD [۱۵] سازگاری درون‌نما و بین‌نما را در فضای لاجیت اعمال می‌کند تا روابط پیش‌بینی در دیدهای مختلف حفظ شود.

در مجموع، روش‌های مبتنی بر شباهت توانایی انتقال دانش ساختاری را دارند؛ با این حال، وابستگی آن‌ها به ترکیب دسته، شیوه نمونه‌برداری و دقت ساختار هندسی معلم می‌تواند در دسته‌های کوچک یا حضور نویز منجر به ناپایداری شود و در برخی موارد تعمیم‌پذیری را محدود کند.

[1] Kullback–Leibler
[2] Global
[3] Similarity

۳- مبانی اولیه

۳-۱- آشنایی با چکانش دانش

در فرآیند چکانش دانش، هدف آن است که دانش موجود در یک شبکه معلم با ظرفیت بالا به یک شبکه دانش‌آموز سبک منتقل شود تا مدل دانش‌آموز بتواند رفتار پیش‌بینی و بازنمایی‌های درونی معلم را تقلید کند. در ابتدا، یک چارچوب استاندارد شبکه عصبی کانولوشنی (CNN[1]) در نظر گرفته شده است. شبکه معلم با (f^t, h^t) و شبکه دانش‌آموز با (f^s, h^s) نشان شده، که در آن هر شبکه شامل یک استخراج‌کننده ویژگی $f: I \to x$ و یک طبقه‌بند $h: x \to p$ است. برای یک تصویر ورودی $I \in \mathbb{R}^{H\times W\times 3}$، استخراج‌کننده ویژگی، بازنمایی $x \in \mathbb{R}^d$ با بعد d تولید می‌کند. طبقه‌بند h معمولا با یک ماتریس نگاشت خطی $W \in \mathbb{R}^{d\times K}$ پارامترسازی می‌شود که بردار ویژگی را به یک بردار لاجیت $z = W^\top x = [z_1, z_2, \dots, z_K] \in \mathbb{R}^K$ منتقل می‌کند؛ جایی‌که K تعداد کلاس‌هاست. در نهایت، لاجیت‌ها از طریق تابع softmax نرمال‌سازی می‌شوند و یک پارامتر دما τ برای تولید احتمال‌های نرم‌شده p_k^τ به کار می‌رود که در چکانش دانش نقش اهداف نرم آگاهانه را ایفا می‌کنند:

$$p_k^\tau = \frac{\exp(z_k/\tau)}{\sum_{j=1}^{K}\exp(z_j/\tau)}, 1 \le k \le K \qquad (۱)$$

تابع زیان استاندارد انتروپی متقاطع نسبت به برچسب حقیقی y_k به شکل زیر نوشته می‌شود:

$$L_{\mathrm{CE}} = -\sum_{k=1}^{K} y_k \log p_k \qquad (۲)$$

زیان چکانش معمولا به کمک واگرایی کولبک–لیبلر KL میان توزیع نرم معلم $P(i)$ و خروجی نرم دانش‌آموز $Q(i)$ تعریف می‌شود:

$$L_{\mathrm{KD}} = KL(P \parallel Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)} \qquad (۳)$$

ترکیب زیان انتروپی متقاطع و زیان چکانش دانش باعث می‌شود مدل دانش‌آموز هم از برچسب‌های حقیقی و هم از «دانش پنهان» معلم یاد بگیرد؛ دانشی که روابط میان‌کلاسی و میزان اطمینان نسبی معلم را در خود دارد.

۳-۲- نمای کلی فرآیند کار

روش‌های موجود چکانش دانش، اعم از مبتنی بر لاجیت، ویژگی یا شباهت، هر یک با محدودیت‌هایی مانند نظارت تک‌سطحی، انتقال اطلاعات زائد و نبود هم‌ترازی ساختاری کافی مواجه‌اند. برای رفع این چالش‌ها، چارچوبی چنددانه‌ای پیشنهاد می‌شود که به‌صورت هم‌زمان توجه محلی، تفکیک زیرفضایی و انسجام ساختار سراسری میان معلم و دانش‌آموز را مدل می‌کند. همان‌گونه که در شکل ۱ نشان داده شده، این چارچوب از سه ماژول مکمل شامل توجه سه‌بعدی، ماسک خصمانه و تنظیم کروی فضا تشکیل شده است.

برای هر ورودی، ابتدا شبکه‌های معلم و دانش‌آموز ویژگی‌های میانی را استخراج می‌کنند. ماژول توجه سه‌بعدی روی نقشه‌های ویژگی هر دو شبکه اعمال می‌شود تا نواحی اطلاعاتی برجسته و سیگنال‌های نویزی سرکوب شوند. هم‌ترازی دانش بر پایه نقشه‌های توجه تقویت‌شده انجام می‌شود تا تمرکز دانش‌آموز به بخش‌های اثرگذار در پیش‌بینی معلم هدایت گردد.

در مرحله بعد، تعبیه‌های[2] نهایی معلم و دانش‌آموز با یکدیگر ترکیب شده و به مولد ماسک داده می‌شوند. این مولد یک ماسک نرم تولید می‌کند که فضای ویژگی را به‌صورت تطبیقی به زیرفضاهای برتر و ضعیف‌تر تقسیم می‌کند. هر زیرفضا از طریق یک طبقه‌بند مستقل آموزش می‌بیند تا ابعاد اطلاعاتی تقویت و مؤلفه‌های نامرتبط تضعیف شوند.

در ادامه، ماژول تنظیم کروی فضا با نگاشت بازنمایی‌های دو شبکه روی یک ابرکره، هم‌ترازی هندسی سراسری را اعمال می‌کند. با تنظیم شباهت‌های زوجی میان نمونه‌ها، ساختار کلی فضای ویژگی معلم حفظ شده و بازنمایی‌هایی با فشردگی درون‌کلاسی و جدایش بین‌کلاسی بهتر حاصل می‌شود.

در مجموع، این خط لوله امکان انتقال هم‌زمان دانش محلی، زیرفضایی و ساختاری را فراهم کرده و به بهبود دقت و تعمیم‌پذیری دانش‌آموز منجر می‌شود.

۴- روش

در این بخش، چارچوب پیشنهادی ارائه شده است. ابتدا، سازوکارهای ماژول توجه سه‌بعدی، ماژول ماسک خصمانه، و ماژول تنظیم فضای کروی می‌شود. در نهایت، این مؤلفه‌ها را ترکیب کرده و تابع زیان کلی چارچوب را فرموله می‌شود.

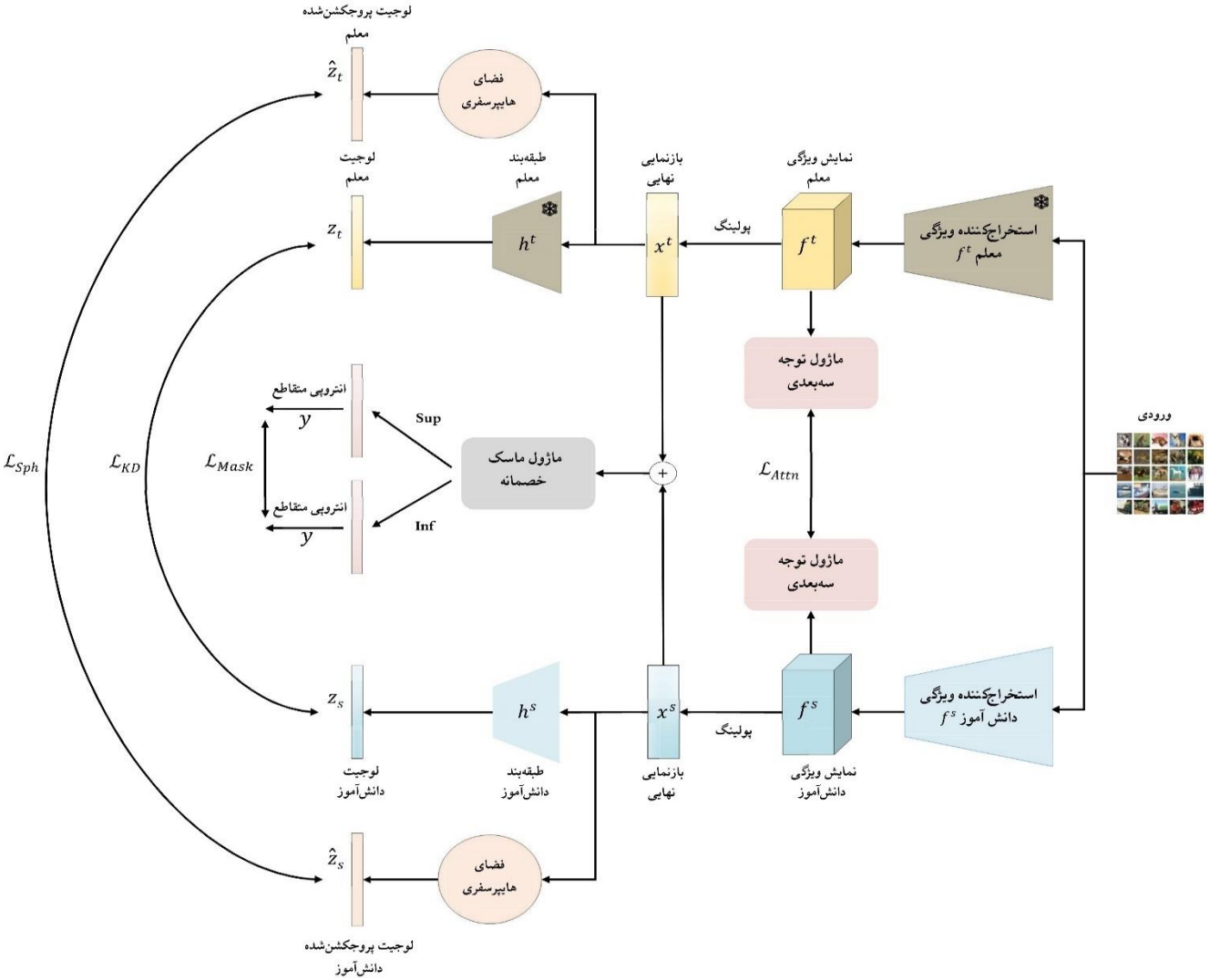
۴-۱- ماژول توجه سه‌بعدی

در چکانش دانش، هم‌ترازی ویژگی اغلب با تطبیق مستقیم بازنمایی‌های میانی شبکه معلم و دانش‌آموز انجام می‌شود. با این حال، چنین هم‌ترازی ساده‌ای اغلب قادر به ثبت کامل الگوهای توجه فضایی‌ـ کانالی معلم نیست و در نتیجه دانش‌آموز ممکن است ویژگی‌هایی را تقلید کند که از قدرت تمایزبخشی و تعمیم‌پذیری کافی برخوردار نیستند. برای افزایش توانایی دانش‌آموز در یادگیری ظرفیت بازنمایشی معلم طی فرآیند چکانش، یک سازوکار توجه کارآمد، در سطح نورون و بدون پارامتر معرفی شده که در مرحله هم‌ترازی ویژگی عمل کرده و برجستگی ویژگی‌های میانی را هنگام هم‌ترازسازی تقویت می‌کند. برخلاف روش‌های

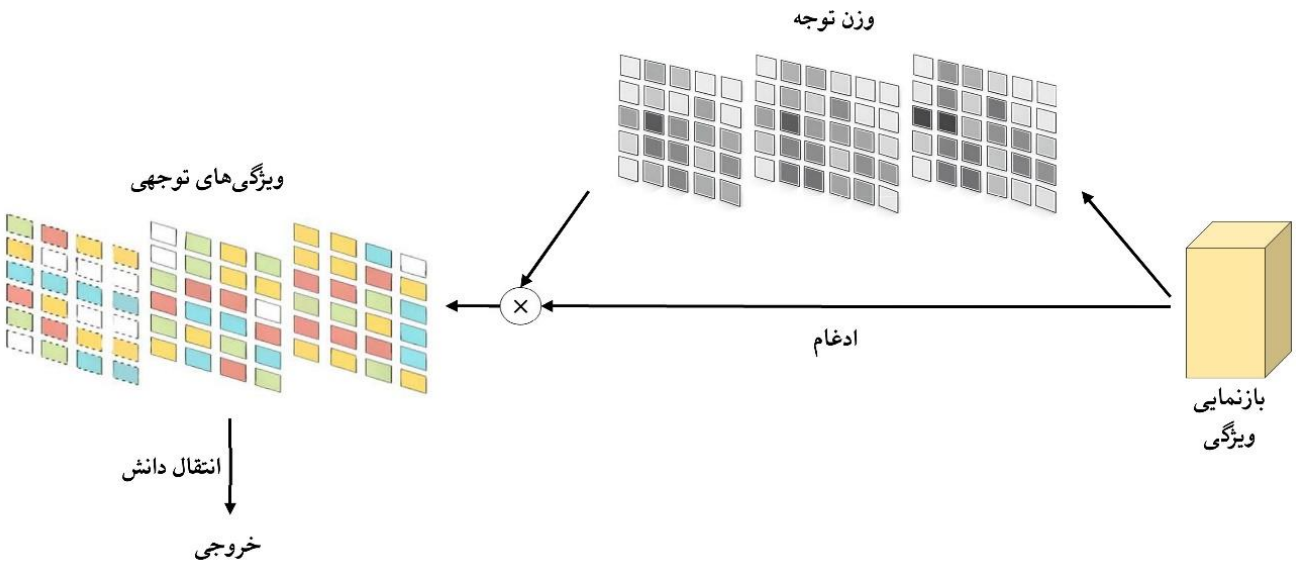
[1] Convolutional Neural Network (CNN)

[2] Embedding

سنتی که ویژگی‌های خام را مستقیما هم‌تراز می‌کنند، ایدهٔ اصلی این اســت که به هر نورون در نقشــهٔ ویژگی وزنی تطبیقی نســبت داده شـود، تا شـبکه بتواند به‌طور خودکار نواحی اطلاعاتی را برجسته و اطلاعات اضافی یا نویزی را سرکوب کند



شکل (۱) : نمای کلی مدل چکانش دانش پیشنهادی. این مدل از سه ماژول تشکیل شده است: (۱) ماژول توجه سه‌بعدی، (۲) ماژول ماسک خصمانه، و (۳) ماژول تنظیم‌سازی فضای کروی.



شکل (۲) : نمای کلی ماژول توجه سه‌بعدی پیشنهادی

به‌طور مشخص، اگر نقشه ویژگی میانی با F نمایش داده شود، که در آن B اندازه دسته، C تعداد کانال‌ها، و $H \times W$ ابعاد فضایی هستند. یک سازوکار توجه در سطح نورون طراحی شده که بر پایه یک تابع انرژی است. ایده اصلی این مکانیزم ارزیابی میزان انحراف هر نورون در نقشه ویژگی از نورون‌های دیگر در همان کانال با استفاده از یک تابع انرژی از پیش تعریف‌شده، و سپس تعیین وزن اهمیت تطبیقی بر اساس آن است. به این ترتیب، ویژگی‌های مهم تقویت شده و اطلاعات نامربوط سرکوب می‌شوند.

همان‌گونه که در شکل ۲ نشان‌داده شده است، برای هر نورون x_i در کانال c، میانگین فضایی و واریانس به‌شکل زیر محاسبه می‌شوند:

$$\mu_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} x_{c,i,j} \qquad (۴)$$

$$\sigma_c^2 = \frac{1}{HW-1}\sum_{i=1}^{H}\sum_{j=1}^{W}(x_{c,i,j}-\mu_c)^2 \qquad (۵)$$

بر اساس انحراف هر نورون از میانگین کانال، تابع انرژی بدون پارامتر زیر برای سنجش اهمیت هر نورون طراحی می‌شود:

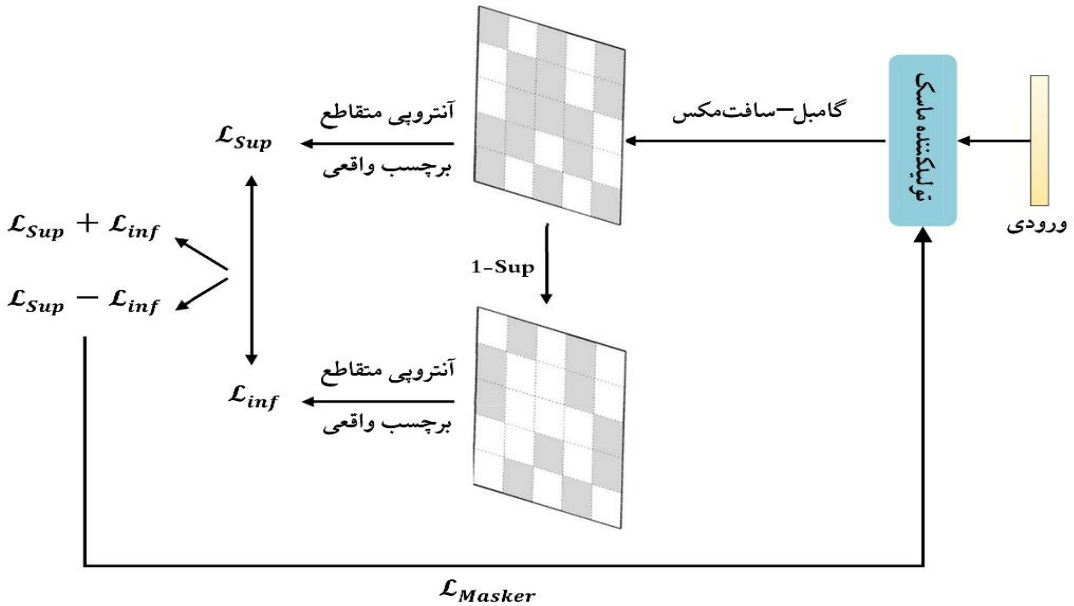
$$E_{c,i,j} = \frac{(x_{c,i,j}-\mu_c)^2}{4(\sigma_c^2+\lambda)} + 0.5 \qquad (۶)$$

که در آن λ عضوی برای منظم‌سازی است تا از تقسیم بر صفر جلوگیری شود. در نهایت، انرژی از طریق یک تابع سیگموید[1] نرمال می‌شود تا وزن نهایی توجه به‌دست آید:

$$\mathrm{attn}(t) = \sigma(E_{c,i,j}) = \frac{1}{1+\exp(-E_{c,i,j})} \qquad (۷)$$

در فرآیند چکانش، سازوکار توجه در سطح نورون بر نقشه‌های ویژگی هر دو شبکه معلم و دانش‌آموز اعمال می‌شود. با بهره‌گیری از این تابع انرژی بدون پارامتر، به هر نورون وزنی تطبیقی داده می‌شود. سپس خطای میانگین مربعات بین وزن‌های توجه معلم و دانش‌آموز کمینه می‌شود تا نقشه‌های توجهی آن‌ها در فضای ویژگی با یکدیگر هم‌راستا شوند. زیان چکانش توجه به‌شکل زیر تعریف می‌شود:

$$\mathcal{L}_{\mathrm{attn}} = \frac{1}{NCHW}\parallel \mathrm{attn}(f_t) - \mathrm{attn}(f_s) \parallel_2^2 \qquad (۸)$$



شکل (۳) : توضیح ماژول ماسک خصمانه. ویژگی دانش‌آموز f_s و ویژگی معلم f_t با یکدیگر ادغام شده و از طریق یک گزینش‌گر مبتنی بر Gumbel-Softmax عبور می‌کنند تا فضای ویژگی به زیرفضای برتر و زیرفضای ضعیف‌تر تفکیک شود.

۲-۴- ماژول ماسک خصمانه

برای تقویت نقش هدایت‌کننده شبکه معلم در فضای ویژگی دانش‌آموز طی فرآیند چکانش، یک ماژول ماسک‌گذاری خصمانه پیشنهاد شده است. این ماژول بر پایه چارچوب کلاسیک معلم–دانش‌آموز ساخته شده و شامل یک مولد ماسک است که با ترکیب ویژگی‌های دانش‌آموز و بازنمایی‌های معلم، ماسک‌های نرم تولید می‌کند. این ماسک‌ها ویژگی‌های دانش‌آموز را به زیرفضاهای برتر و پست‌تر تقسیم می‌کنند و چکانش دانش به‌طور مستقل در هر زیرفضا تحت نظارت خروجی‌های معلم انجام می‌شود. این طراحی امکان انتقال دانش دقیق‌تر و هدفمندتر را فراهم می‌سازد، در حالی‌که هم‌زمان اطلاعات اضافی، نادرست یا مزاحم در

[1] Sigmoid

بازنمایی‌های دانش‌آموز سرکوب می‌شود، و بدین ترتیب ظرفیت تعمیم و کیفیت بازنمایشی آن بهبود می‌یابد.

به‌طور مشخص، با داشتن یک تصویر ورودی، مدل دانش‌آموز بازنمایی ویژگی نهایی $f_s \in \mathbb{R}^d$ را تولید می‌کند، در حالی که مدل معلم بازنمایی ویژگی نهایی $f_t \in \mathbb{R}^C$ را خروجی می‌دهد، که در آن d بعد ویژگی و C تعداد کلاس‌ها است. این دو بازنمایی به یکدیگر الحاق شده و به مولد ماسک $\mathcal{W}(\cdot)$ داده می‌شوند که یک بردار ماسک نرم $m \in \mathbb{R}^d$ تولید می‌کند. بر اساس مقادیر ماسک، نسبت بالایی $\kappa \in (0,1)$ از ابعاد به زیرفضای برتر تخصیص داده می‌شود و ابعاد باقیمانده به زیرفضای ضعیف‌تر، همان‌گونه که در شکل ۳ نشان‌داده شده است. فرمول الحاق ویژگی‌ها به‌شکل زیر نوشته می‌شود:

$$f = [f_s, f_t] \in \mathbb{R}^{d+d_t}. \qquad (۹)$$

سپس:

$$m = \text{Gumbel-Softmax}(\mathcal{W}([f_s, f_t]), \kappa) \qquad (۱۰)$$

این ماسک با نمونه‌برداری k-hot قابل تفکیک تولید می‌شود. در پیاده‌سازی عملی، دمای Gumbel-Softmax برابر با ۱ تنظیم شد و در طول آموزش ثابت نگه داشته شد. بررسی تجربی نشان داد مدل نسبت به تغییرات کوچک این دما در بازه $[۰/۵، ۲]$ حساسیت بالایی ندارد. این مقدار با هدف ایجاد تعادل میان نمونه‌برداری نزدیک به گسسته و پایداری گرادیان انتخاب شد.

سپس ویژگی‌های الحاق‌شده h به دو زیرفضای sup و inf تقسیم می‌شوند:

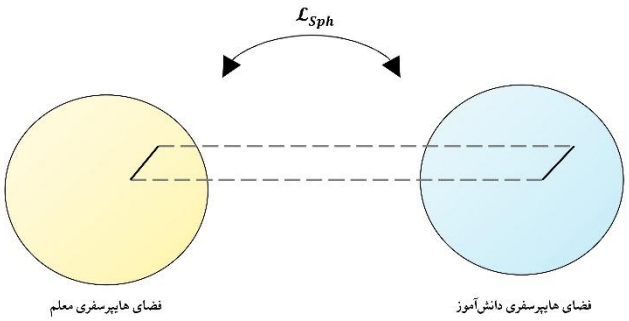
$$f^{\text{sup}} = f \odot m, f^{\text{inf}} = f \odot (1 - m)$$

که در آن $m \in (0,1)^{d+d_t}$ ماسک نرم و $\odot$ ضرب عضو به عضو است. هر زیرفضا از یک طبقه‌بند مستقل عبور کرده و خروجی‌های $h_{\text{sup}}(f^{\text{sup}})$ و $h_{\text{inf}}(f^{\text{inf}})$ را ایجاد می‌کند. هر دو شاخه با استفاده از واگرایی KL نسبت به لاجیت‌های معلم چکانش می‌شوند تا دانش‌آموز بتواند در هر زیرفضا مؤثرتر دانش معلم را فراگیرد. زیان‌های چکانش زیرفضای برتر و زیرفضای ضعیف‌تر به‌شکل زیر تعریف می‌شوند:

$$\mathcal{L}_{KD}^{\text{sup}} = -T^2 \, \mathbb{E}_{x \sim D_t} \sum_{i=1}^{C} p_i^t(x;T) \log p_i^{\text{sup}}(x;T) \qquad (۱۱)$$

$$\mathcal{L}_{KD}^{\text{inf}} = -T^2 \, \mathbb{E}_{x \sim D_t} \sum_{i=1}^{C} p_i^t(x;T) \log p_i^{\text{inf}}(x;T) \qquad (۱۲)$$

در مرحلهٔ آموزش، زیان $\mathcal{L}_{CE}^{\text{sup}}$ برای شاخه برتر کمینه شده و زیان $\mathcal{L}_{CE}^{\text{inf}}$ بیشینه می‌شود، تا مولد ماسک تشویق شود ویژگی‌های حساس به معلم را به زیرفضای برتر تخصیص دهد و ویژگی‌های زائد یا دشوار را به زیرفضای ضعیف‌تر منتقل کند:



شکل (۴) : توضیح ماژول تنظیم‌سازی فضای کروی. ویژگی‌های معلم و دانش‌آموز روی ابرکره‌ها فرافکنی می‌شوند و از طریق تابع زیان سازگاری کروی L_{sph} هم‌تراز می‌گردند؛ تابعی که شباهت ساختاری جهانی را اعمال و تضمین می‌کند.

$$\mathcal{L}_{\text{Masker}} = \min_{\mathcal{W}} 0.5 \cdot \mathcal{L}_{CE}^{\text{sup}} - 0.5 \cdot \mathcal{L}_{CE}^{\text{inf}} \qquad (۱۳)$$

در حالی‌که شبکهٔ اصلی شامل مولد ویژگی و دو طبقه‌بند با کمینه‌سازی مجموع زیان‌های دو زیرفضا به‌روزرسانی می‌شود:

$$\mathcal{L}_{\text{Mask}} = \min_{f, h_{\text{sup}}, h_{\text{inf}}} 0.5 \cdot \mathcal{L}_{CE}^{\text{sup}} + 0.5 \cdot \mathcal{L}_{CE}^{\text{inf}} \qquad (۱۴)$$

۳-۴- ماژول تنظیم فضای کروی

از آنجا که فضای ابرکره به‌طور طبیعی قیود زاویه‌ای و بزرگی نرمال‌شده فراهم می‌کند، می‌تواند مشکل فروپاشی ویژگی[1] را کاهش داده و تمایزبخشی بازنمایی‌ها را افزایش دهد. این مزیت اخیرا در یادگیری تقابلی و چکانش دانش مبتنی بر ویژگی توجه زیادی را به خود جلب کرده است. برای تقویت تمایزبخشی و قابلیت تعمیم دانش‌آموز، در این مقاله یک استراتژی جدید مبتنی بر سازگاری ساختاری کروی مطابق شکل ۴ معرفی شده است.

ویژگی‌های هر دو مدل معلم و دانش‌آموز روی ابرکره نگاشته شده و زیان سازگاری کروی $\mathcal{L}_{Sph}$ برای هم‌ترازسازی آن‌ها اعمال می‌شود. ایده اصلی ساخت ماتریس‌های احتمال شباهت زوجی روی ابرکره است تا ساختار هندسی کلی فضای ویژگی را توصیف کند. سپس یک زیان انتروپی متقاطع برای تضمین تطابق توزیع شباهت مدل دانش‌آموز با معلم به کار می‌رود.

نگاشت ویژگی‌ها روی ابرکره با نرمال‌سازی L2 باعث حذف اثر مقیاس و تمرکز بر روابط زاویه‌ای میان نمونه‌ها می‌شود. در این فضا، فاصله اقلیدسی میان بردارهای نرمال‌شده به‌طور مستقیم با شباهت کسینوسی مرتبط است، اما قید کروی از بزرگ‌شدن نامتوازن نرم بردارها و فروپاشی بازنمایی جلوگیری می‌کند. این ویژگی موجب می‌شود هم‌ترازی هندسی دانش‌آموز با معلم در سطح ساختار سراسری پایدارتر از تطبیق مستقیم اقلیدسی باشد.

ابتدا، خروجی‌های لاجیت هر دو مدل روی ابرکره نگاشته شده و ویژگی‌های h_i و g_i به‌صورت زیر نرمال می‌شوند:

[1] Feature Collapse

$$\hat{h}_i = \frac{h_i}{\| h_i \|_2}, \hat{g}_i = \frac{g_i}{\| g_i \|_2} \quad (۱۵)$$

سپس ماتریس شباهت زوجی روی ابرکره تعریف می‌شود:

$$Q_{ij}^{(h)} = \exp(-\| \hat{h}_i - \hat{h}_j \|^2), Q_{ij}^{(g)} = \exp(-\| \hat{g}_i - \hat{g}_j \|^2) \quad (۱۶)$$

هدف زیان کروی، کمینه کردن اختلاف میان این دو ماتریس شباهت است:

$$\mathcal{L}_{Sph} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[Q_{ij}^{(h)}\log(Q_{ij}^{(g)}) + (1 - Q_{ij}^{(h)})\log(1 - Q_{ij}^{(g)})] \quad (۱۷)$$

این ماژول بدون نیاز به داده‌افزایی یا ساخت زوج‌های مثبت/منفی، انسجام ساختار سراسری ویژگی‌ها را تضمین کرده و مرزبندی کلاسی را بهبود می‌دهد.

۴-۴- تابع هدف نهایی

برای بهینه‌سازی جامع آموزش، روش پیشنهادی سه نوع زیان چکانش را به‌طور مشترک با زیان طبقه‌بندی ترکیب می‌کند. هدف آموزش دانش‌آموز چنین است:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{\text{Attn}} + \beta\mathcal{L}_{\text{Mask}} + \gamma\mathcal{L}_{\text{Sph}} + \delta\mathcal{L}_{\text{KD}} \quad (۱۸)$$

که در آن $\alpha, \beta, \gamma, \delta$ وزن‌های تنظیم‌کننده زیان‌های کمکی هستند. اجزای زیان عبارتند از:

- $\mathcal{L}_{cls}$:زیان انتروپی متقاطع استاندارد، تضمین‌کننده توان طبقه‌بندی قوی.
- $\mathcal{L}_{\text{Attn}}$:هم‌ترازی توجه نورون‌سطحی میان معلم و دانش‌آموز.
- $\mathcal{L}_{\text{Mask}}$:ترکیب زیان‌های دو زیرفضا که تمایزپذیری و انسجام ویژگی‌ها را بهبود می‌دهد.
- $\mathcal{L}_{\text{Sph}}$:زیان سازگاری کروی که ساختار رابطه‌ای سراسری را حفظ می‌کند.
- $\mathcal{L}_{\text{KD}}$:زیان چکانش استاندارد لاجیت معلم (KL).

از منظر بهینه‌سازی، سه ماژول پیشنهادی را می‌توان به‌عنوان اعمال قیود مکمل بر فضای بازنمایی تفسیر کرد: ماژول توجه سه‌بعدی قیود محلی در سطح نورون، ماژول ماسک خصمانه قیود زیرفضایی تطبیقی، و ماژول تنظیم کروی قیود هندسی سراسری را اعمال می‌کند. ترکیب این قیود به‌صورت افزایشی موجب محدود شدن فضای جستجوی ناسازگار با ساختار معلم شده و فرآیند یادگیری دانش‌آموز را به سمت همگرایی پایدارتر و بازنمایی‌های سازگارتر هدایت می‌کند. نتایج مطالعه حذف مؤلفه‌ها نیز وجود این هم‌افزایی را به‌صورت تجربی تایید خواهد می‌کند.

۵- نتایج آزمایشگاهی

۵-۱- تنظیمات آزمایش

در این مقاله، از سه مجموعه‌داده رایج طبقه‌بندی تصویر برای ارزیابی استفاده شده است. به‌طور مشخص، CIFAR-100 [۵۰] برای آموزش اولیه و ارزیابی عملکرد پایه به‌کار می‌رود، درحالی‌که STL-10 [۵۱] و TinyImageNet [۵۲] برای ارزیابی قابلیت تعمیم و انتقال‌پذیری مدل‌های آموزش‌دیده در میان مجموعه‌داده‌های مختلف انتخاب شده‌اند. CIFAR-100 یک مجموعه‌داده طبقه‌بندی تصویر به‌طور گسترده استفاده شده است که شامل ۵۰,۰۰۰ تصویر آموزشی و ۱۰,۰۰۰ تصویر آزمایشی با اندازهٔ ۳۲×۳۲ می‌باشد. این مجموعه‌داده شامل ۱۰۰ کلاس است که برای هر کلاس ۵۰۰ تصویر آموزشی وجود دارد.

STL-10 شامل یک مجموعه آموزشی با ۵۰۰۰ تصویر برچسب‌خورده از ۱۰ کلاس و یک مجموعه آزمایشی شامل ۸۰۰۰ تصویر با وضوح ۹۶×۹۶ پیکسل است. TinyImageNet زیرمجموعه‌ای از ImageNet است که شامل ۲۰۰ کلاس می‌باشد و برای هر کلاس ۵۰۰ تصویر آموزشی و ۵۰ تصویر اعتبارسنجی دارد. در این مقاله آزمایش‌های گسترده‌ای را روی چندین زوج معلم_دانش آموز انجام شده است، از جمله ResNet32×4، WRN_40-2 و VGG13 به‌عنوان مدل‌های معلم، و ResNet8×4، WRN_40-1، WRN_16-2، VGG8، ResNet20، ShuffleNetV1 و ShuffleNetV2 به‌عنوان مدل‌های دانش آموز انتخاب شده است.

تمام مدل‌ها با استفاده از گرادیان کاهشی تصادفی (SGD[1]) با مومنتوم ۰/۹ و ضریب کاهش وزن ۵×۱۰$^{-۴}$ آموزش داده شدند. برای CIFAR-100، روش پیشنهادی به‌مدت ۲۴۰ دوره با اندازه دسته ۶۴ و نرخ یادگیری اولیه ۰/۰۵ آموزش داده شد. نرخ یادگیری در دوره‌های ۱۵۰، ۱۸۰ و ۲۱۰ با ضریب ۰/۱ کاهش می‌یابد.

همه آزمایش‌ها روی یک کامپیوتر شخصی مجهز به GPU مدل NVIDIA GeForce RTX4090 و پردازنده AMD EPYC 7542 با ۵۶ گیگابایت حافظه انجام شده است. سیستم‌عامل Ubuntu 22.04 LTS بوده و کد با استفاده از Python 3.11 و PyTorch 2.1.2 تحت CUDA 11.8 پیاده‌سازی شده است.

از نظر پیچیدگی محاسباتی، ماژول توجه سه‌بعدی فاقد پارامتر قابل‌آموزش بوده و تنها شامل محاسبات آماری ساده (میانگین و واریانس کانالی) است که سربار ناچیزی نسبت به شبکه پایه ایجاد می‌کند. ماژول تنظیم کروی صرفاً در مرحله آموزش فعال است و در زمان استنتاج حذف می‌شود، بنابراین اثری بر زمان inference ندارد. ماژول ماسک خصمانه نیز شامل یک ساختار سبک برای تولید ماسک است. اندازه‌گیری تجربی نشان داد افزایش FLOPs و زمان استنتاج کمتر از چند درصد نسبت به مدل پایه است، که در مقابل بهبود دقت حاصل‌شده قابل قبول است.

[1] Stochastic Gradient Descent (SGD)

۵-۲- تنظیمات ابرپارامتری

برای ماژول توجه سه‌بعدی، مقدار منظم‌ساز $\lambda = ۱۰^{-۴}$ در معادله (۶) تنظیم شده است. این مقدار پایداری عددی را تضمین می‌کند و مطابق با استانداردهای رایج است. برای ماژول ماسک خصمانه، ضریب کم‌تراکمی $\kappa = 0.5$ تنظیم شده است تا نسبت ابعاد ویژگی اختصاص‌یافته به زیرفضای برتر کنترل شود. برای تابع ضرر کلی در معادله (۱۸)، ضرایب وزنی $\alpha = 0.1$، $\beta = 0.1$، $\gamma = 0.1$ و $\delta = 0.9$ تنظیم شده‌اند. تمام آزمایش‌های چکانش دانش از دمای $T = 4$ استفاده می‌کنند. ضرایب وزنی α، β، γ و δ از طریق جستجوی شبکه‌ای (Grid Search) روی مجموعه اعتبارسنجی CIFAR-100 در بازه‌های {۰/۰۵، ۰/۱، ۰/۵، ۱/۰} تنظیم شدند و مقادیر نهایی بر اساس بهترین توازن میان دقت نهایی و پایداری آموزش انتخاب گردیدند.

بررسی تجربی نشان داد عملکرد مدل در یک بازه معقول از این ضرایب پایدار بوده و حساسیت شدیدی نسبت به تغییرات کوچک آن‌ها ندارد. مقدار دمای T نیز مطابق تنظیمات رایج در ادبیات چکانش دانش انتخاب شده است.

۵-۳- مقایسه با روش‌های پیشرفته موجود

آزمایش‌های گسترده‌ای روی مجموعه‌داده CIFAR-100 با استفاده از انواع مختلف زوج‌های معلم_دانش آموز انجام دادیم که هم ترکیب‌های معماری مشابه و هم مختلف را پوشش می‌دهد. همان‌طور که در جدول ۱ تا جدول ۳ نشان داده شده است، مدل چکانش دانش پیشنهادی به‌طور پیوسته بهترین عملکرد را در میان روش‌های مقایسه‌شده هم در تنظیمات مبتنی بر لاجیت و هم مبتنی بر ویژگی به‌دست می‌آورد.

برای مثال، در زوج ResNet8×4 / ResNet32×4، مدل چکانش دانش پیشنهادی به دقت ۷۵/۲۷٪ می‌رسد که از KD، CTKD، DKD، SRD و TeKAP به‌طور قابل توجهی بهتر است. این نتیجه نشان می‌دهد که طراحی چندانه‌ای به دانش آموز کمک می‌کند تا دانش پنهان معلم را بهتر استخراج و منظم‌سازی کند، که منجر به روابط بین‌کلاسی پایدارتر و کاهش بیش‌اعتمادی می‌شود، قابلیتی که روش‌های مبتنی بر تطبیق لاجیت به‌تنهایی قادر به ارائهٔ آن نیستند. همچنین در زوج WRN_40-2 / WRN_40-1 و WRN_40-2 / WRN_16-2 نیز مدل چکانش دانش پیشنهادی بهترین عملکرد را نسبت به روش‌های پایه در جدول ۱ به‌دست می‌آورد.

این نتایج نشان می‌دهند که مدل چکانش دانش پیشنهادی با ادغام نشانه‌های معنایی غنی‌تر و منظم‌سازی ساختاری قوی‌تر، به‌طور مؤثر نظارت سطح لاجیت را بهبود می‌دهد. زمانی که معلم و دانش آموز شباهت معماری دارند، این بهبودها بیشتر نمایان شده و نشان می‌دهند که مدل چکانش دانش پیشنهادی توانایی انتقال بازنمایی‌های سطح بالا و تقویت هم‌ترازی تبعیضی در فضای کلاس‌ها را افزایش می‌دهد. همان‌طور که جدول ۲ و جدول ۴ نشان می‌دهند، در تنظیمات با معماری‌های مختلف، این بهبودها حتی چشمگیرتر هستند. برای نمونه، در ResNet32×4 / ShuffleNetV1، مدل چکانش دانش پیشنهادی به دقت ۷۵/۳۲٪ در حالت مبتنی بر لاجیت و ۷۵/۸۴٪ در حالت مبتنی بر ویژگی می‌رسد که هر دو بهتر ازTeKAP و CRD هستند. در ResNet32×4 / ShuffleNetV2، دقت مدل چکانش دانش پیشنهادی به ۷۶/۲۶٪ افزایش می‌یابد که حتی بیشتر از معلم با دقت ۷۴/۶۴٪ است. در WRN_40-2 / ShuffleNetV1، مدل چکانش دانش پیشنهادی دقت ۷۶/۷۹٪ را به‌دست آورده و از تمام رقبا بهتر عمل می‌کند، ضمن این‌که فاصله با دقت معلم به حداقل می‌رسد که نشان‌دهنده توان مدل چکانش دانش پیشنهادی در پل‌زدن فاصله معماری‌ها است. این نتایج نشان می‌دهند که مدل چکانش دانش پیشنهادی طیف وسیعی از تنظیمات را پوشش می‌دهد، هم در روش‌های سنتی و هم در روش‌های پیشرفته اخیر بهبودهای قابل‌توجهی ایجاد می‌کند و نسبت به شباهت یا تفاوت معماری نیز حساس نیست. عملکرد بهتر روی مدل‌های دانش آموزی سبک مانند ShuffleNetV1 نشان‌دهنده توانایی مدل چکانش دانش پیشنهادی در افزایش ظرفیت بازنمایی مدل‌های فشرده است، به‌گونه‌ای که می‌توانند با معلم‌های خود رقابت کرده یا حتی از آن‌ها پیشی بگیرند.

جدول (۱) : مقایسه روش‌های چکانش مبتنی بر لاجیت در معماری‌های مشابه معلم_دانش‌آموز بر روی CIFAR-100

	معماری مشابه				
معلم **دانش آموز**	ResNet32x4 ResNet8x4	WRN_40_2 WRN_40_1	WRN_40_2 WRN_40_1	VGG13 VGG8	ResNet56 ResNet20
معلم	۷۹/۴۲	۷۵/۶۱	۷۵/۶۱	۷۴/۶۴	۷۲/۳۴
دانش آموز	۷۲/۵۰	۷۱/۹۸	۷۳/۲۶	۷۰/۳۶	۶۹/۰۶
KD [۱]	۷۳/۳۳	۷۳/۱۹	۷۴/۹۲	۷۲/۹۸	۷۰/۶۶
CTKD [۱۸]	۷۴/۲۴	۷۳/۴۳	۷۴/۸۹	۷۳/۵۲	۷۱/۱۹
DKD [۲]	۷۴/۱۳	۷۲/۳۸	۷۴/۵۳	۷۳/۲۸	۷۰/۸۲
CRD [۵۳]	۷۳/۸۹	۷۲/۸۶	۷۴/۵۶	۷۳/۷۳	۷۱/۳۲
TeKAP [۲۷]	۷۴/۷۹	۷۳/۸۰	۷۵/۲۱	۷۴/۰۰	۷۱/۳۲
مدل چکانش دانش پیشنهادی	**۷۵/۲۷**	**۷۴/۳۲**	**۷۵/۶۷**	**۷۴/۷۶**	**۷۲/۰۴**

جدول (۲) : مقایسه روش‌های چکانش مبتنی بر لاجیت در معماری‌های متفاوت معلم و دانش‌آموز بر روی CIFAR-100

	معماری متفاوت		
معلم دانش آموز	ResNet32x4 ShuffleNetV1	ResNet32x4 ShuffleNetV2	WRN_40_2 ShuffleNetV1
معلم	۷۹/۴۲	۷۴/۶۴	۷۵/۶۱
دانش آموز	۷۰/۵۰	۷۰/۳۶	۷۰/۵۰
KD [۱]	۷۴/۰۷	۷۲/۹۸	۷۴/۸۳
CTKD [۱۸]	۷۳/۹۶	۷۴/۳۲	۷۴/۸۶
DKD [۲]	۷۴/۳۲	۷۴/۸۸	۷۵/۵۸
CRD [۵۳]	۷۴/۵۹	۷۳/۸۶	۷۵/۷۹
TeKAP [۲۷]	۷۴/۹۲	۷۴/۴۳	۷۶/۷۵
مدل چکانش دانش پیشنهادی	**۷۵/۳۲**	**۷۵/۱۳**	**۷۷/۱۲**

جدول (۳) : مقایسه روش‌های چکانش مبتنی بر ویژگی در معماری‌های مشابه معلم و دانش‌آموز (CIFAR-100)

	معماری مشابه				
معلم دانش آموز	ResNet32x4 ResNet8x4	WRN_40_2 WRN_40_1	WRN_40_2 WRN_40_1	VGG13 VGG8	ResNet56 ResNet20
معلم	۷۹/۴۲	۷۵/۶۱	۷۵/۶۱	۷۴/۶۴	۷۲/۳۴
دانش آموز	۷۲/۵۰	۷۳/۲۶	۷۱/۹۸	۷۰/۳۶	۶۹/۰۶
KD [۱]	۷۴/۵۶	۷۵/۳۳	۷۴/۲۶	۷۳/۱۲	۷۱/۵۸
CTKD [۱۸]	۷۴/۲۷	۷۴/۸۲	۷۳/۶۹	۷۳/۵۰	۷۱/۵۷
DKD [۲]	۷۵/۱۶	۷۵/۹۵	۷۳/۸۹	۷۳/۴۴	۷۱/۲۴
CRD [۵۳]	۷۵/۵۱	۷۵/۴۸	۷۴/۱۴	۷۳/۹۴	۷۱/۱۶
TeKAP [۲۷]	۷۵/۶۵	۷۵/۸۳	۷۴/۲۱	۷۴/۱۰	۷۱/۷۱
مدل چکانش دانش پیشنهادی	**۷۵/۷۱**	**۷۵/۹۸**	**۷۴/۶۲**	**۷۴/۸۶**	**۷۱/۹۲**

جدول (۴) : مقایسه روش‌های چکانش مبتنی بر ویژگی در معماری‌های متفاوت معلم و دانش‌آموز (CIFAR-100)

	معماری متفاوت		
معلم دانش آموز	ResNet32x4 ShuffleNetV1	ResNet32x4 ShuffleNetV2	WRN_40_2 ShuffleNetV1
معلم	۷۹/۴۲	۷۴/۶۴	۷۵/۶۱
دانش آموز	۷۰/۵۰	۷۰/۳۶	۷۰/۵۰
KD [1]	۷۴/۹۹	۷۵/۸۴	۷۶/۰۸
CTKD [۱۸]	۷۴/۸۷	۷۵/۹۲	۷۵/۵۱
DKD [۲]	۷۵/۲۳	۷۵/۸۸	۷۶/۱۲
CRD [۵۳]	۷۵/۱۱	۷۵/۶۵	۷۶/۰۵
TeKAP [۲۷]	۷۵/۵۵	۷۵/۲۳	۷۶/۶۰
مدل چکانش دانش پیشنهادی	**۷۵/۸۴**	**۷۶/۲۶**	**۷۶/۷۹**

۴-۵- تحلیل سربار محاسباتی و زمان اجرا

به‌منظور ارزیابی سربار محاسباتی چارچوب پیشنهادی، تعداد پارامترها، عملیات محاسباتی (FLOPs)، زمان آموزش در هر دوره (epoch) و زمان استنتاج مورد بررسی قرار گرفت. این اندازه‌گیری‌ها در تنظیمات CIFAR-100 و بر روی زوج نماینده ResNet32×4 (معلم) و ResNet8×4 (دانش‌آموز) انجام شد. اندازه دسته ۶۴ بوده و تمامی آزمایش‌ها بر روی GPU مدل NVIDIA RTX4090 اجرا شده‌اند. نتایج در جدول ۵ ارائه شده است. همان‌طور که مشاهده می‌شود، افزایش تعداد پارامترها در روش پیشنهادی کمتر از ۲٪ و افزایش FLOPs حدود ۶٪ نسبت به KD است. زمان آموزش در هر دوره تقریباً ۱۴٪ بیشتر از KD و حدود ۶٪ بیشتر از DKD گزارش شد که ناشی از محاسبات اضافی در ماژول ماسک خصمانه است. با این حال، ماژول تنظیم کروی صرفاً در مرحله آموزش فعال بوده و در زمان استنتاج حذف می‌شود؛ ازاین‌رو زمان inference تقریباً برابر با روش‌های پایه باقی می‌ماند (افزایش کمتر از ۵٪). نتایج نشان می‌دهد که بهبود دقت حاصل‌شده (حدود ۰/۵ الی ۱٪) با افزایش محدود سربار محاسباتی همراه است و از منظر عملی و کاربردی، این مبادله هزینه فایده قابل قبول ارزیابی می‌شود. همچنین روند افزایش سربار در سایر تنظیمات آزمایشی مشابه بوده و تغییر قابل توجهی مشاهده نشد.

جدول (۵) : مقایسه سربار محاسباتی و عملکرد روش‌های مختلف چکانش دانش (CIFAR-100)

روش	عملیات محاسباتی (گیگا FLOPs)	زمان آموزش هر دوره (ثانیه)	زمان استنتاج هر تصویر (میلی‌ثانیه)
KD [1]	۱/۸۲	۴۲	۳/۸
CTKD [۱۸]	۱/۸۷	۴۰	۳/۷
DKD [۲]	۱/۸۵	۴۵	۳/۹
CRD [۵۳]	۱/۸۹	۴۱	۳/۶
TeKAP [۲۷]	۱/۸۵	۴۶	۳/۹
روش پیشنهادی	**۱/۹۳**	۴۸	**۴/۰**

۵-۵- ارزیابی Few-Shot در شرایط کم‌نمونه

در این بخش توانایی تعمیم مدل چکانش دانش پیشنهادی را در سناریوهای با داده محدود (کم‌نمونه) ارزیابی شده است. برای این منظور، آزمایش‌های few-shot روی CIFAR-100 انجام شده، به‌طوری‌که تنها ٪۲۵، ٪۵۰ و ٪۷۵ داده‌های آموزشی حفظ شدند و مجموعهٔ آزمایشی ثابت باقی ماند. زوج معلم‌ـ دانش آموز ResNet56 / ResNet20 برای ارزیابی انتخاب شد.

طبق نتایج ارائه‌شده در جدول ۶، مدل چکانش دانش پیشنهادی به‌طور پیوسته تمام روش‌های مقایسه‌شده را در تمام نسبت‌های کم‌نمونه شکست می‌دهد. به‌طور خاص، مدل چکانش دانش پیشنهادی در نسبت‌های ٪۲۵، ٪۵۰ و ٪۷۵ دقت‌های ٪۶۴/۸۹، ٪۶۹/۱۳ و ٪۷۰/۶۹ را کسب می‌کند. این نتایج نشان می‌دهند مدل چکانش دانش پیشنهادی در شرایط کمبود داده نیز بسیار مقاوم و قابل‌اتکا است، زیرا قادر است اطلاعات چنددانه‌ای معلم را به‌طور مؤثر استخراج و منتقل کند و درعین‌حال انسجام درون‌کلاسی و جدایش بین‌کلاسی را حتی در نظارت محدود حفظ نماید.

لازم به ذکر است که ارزیابی few-shot در این مطالعه بر روی یک زوج معلم—دانش‌آموز انجام شده است. گسترش این تحلیل به معماری‌های دیگر می‌تواند تصویر جامع‌تری از رفتار چارچوب در شرایط کم‌نمونه ارائه دهد و به‌عنوان مسیر پژوهشی آینده در نظر گرفته می‌شود.

۶-۵- قابلیت انتقال به مجموعه‌داده‌های دیگر (Transferability)

دانش‌آموزی که با مدل چکانش دانش پیشنهادی آموزش دیده است، در مقایسه با روش پایه، توانایی انتقال‌پذیری بسیار بهتری به مجموعه‌داده‌های مختلف نشان می‌دهد. برای ارزیابی این موضوع، آزمایش‌های انتقال روی دو وظیفه از CIFAR-100 به STL10 و ازCIFAR-100 بهTinyImageNet انجام شد. پس از اتمام مرحله چکانش روی CIFAR-100 و با استفاده از زوج شبکه WRN_40-2 به WRN_40-1 به عنوان معلم‌ـ دانش‌آموز، مدل معلم کنار گذاشته می‌شود و تنها مدل دانش‌آموز چکانش شده برای انتقال مورد استفاده قرار می‌گیرد.

برای سازگار کردن مدل دانش‌آموز آموزش‌دیده روی CIFAR-100 با مجموعه‌داده هدف، لایهٔ طبقه‌بندی ۱۰۰ کلاسی اصلی حذف شده و با یک لایه خطی تازه مقداردهی‌شده جایگزین می‌شود؛ تعداد خروجی‌های این لایه مطابق تعداد کلاس‌های مجموعه‌داده هدف تنظیم می‌گردد. در طول آزمایش انتقال، بدنه شبکه ثابت نگه داشته می‌شود و تنها طبقه‌بند جدید آموزش داده می‌شود. این پروتکل باعث می‌شود کیفیت بازنمایی به‌دست‌آمده در مرحله چکانش به‌طور جداگانه سنجیده شود و امکان مقایسه‌ای منصفانه میان تمام روش‌های پایه فراهم گردد؛ زیرا همه این روش‌ها از تنظیمات یکسان یادگیری انتقال بهره می‌برند.

سر طبقه‌بندی با استفاده از SGD برای ۱۰۰ دوره و با نرخ یادگیری ۰٫۰۱، مومنتوم ۰٫۹، وزن‌کاهی ۵×۱۰^{-۴}، و اندازه دسته ۱۲۸ آموزش داده می‌شود. در این مرحله هیچ‌گونه زیان چکانشی اعمال نمی‌گردد و تمام روش‌های مقایسه‌شده از پروتکل یکسان استفاده می‌کنند. بنابراین تفاوت‌های مشاهده‌شده در عملکرد انتقال، مستقیما کیفیت بازنمایی آموخته‌شده در مرحله چکانش CIFAR-100 را منعکس می‌کند و مبنای مقایسه‌ای عادلانه را فراهم می‌سازد. نتایج تجربی ارائه‌شده در جدول ۷ نشان می‌دهد که مدل چکانش دانش پیشنهادی به ترتیب دقت‌های ٪۶۴/۰۳ و ٪۷۹/۲۸ را به دست آورده است که بهبودهایی برابر با ٪۱/۰۸ و ٪۹۱/۰ نسبت به چکانش استاندارد ایجاد می‌کند و از تمام روش‌های مقایسه‌شده دیگر نیز بهتر عمل می‌کند.

روی مجموعه‌داده STL10، ویژگی‌هایی که توسط مدل چکانش دانش پیشنهادی آموخته شده‌اند انتقال مؤثرتری دارند و مرزهای طبقه‌بندی واضح‌تر و دقت تشخیص بالاتری تولید می‌کنند. روی TinyImageNet که با تعداد کلاس‌های بیشتر و تنوع درون‌کلاسی بالا چالش‌برانگیزتر است، مدل چکانش دانش پیشنهادی همچنان بهترین عملکرد را ارائه می‌دهد. این نتایج نشان می‌دهد که بازنمایی‌هایی که مدل چکانش دانش پیشنهادی می‌کند نه‌تنها قدرت تمایز بالایی دارند، بلکه در میان مجموعه‌داده‌های مختلف نیز پایدار باقی می‌مانند.

این بهبودها ناشی از توانایی دانش‌آموز در دریافت و سازمان‌دهی چندسطحی اطلاعات معلم است؛ این امر باعث تشکیل خوشه‌های فشرده‌تر و جدایش واضح‌تر بین کلاس‌ها در فضای ویژگی می‌شود و در نهایت مرزهای تصمیم دقیق‌تر و قابل‌اعتمادتر را فراهم می‌آورد. در نتیجه، دانش‌آموز ویژگی‌هایی قابل‌انتقال‌تر و تعمیم‌پذیرتر می‌آموزد، که سازگاری آن را با مجموعه‌داده‌های متفاوت افزایش داده و گستره کاربردی وسیع مدل چکانش دانش پیشنهادی را تأیید می‌کند

برای مدل دانش آموز ResNet20، مدل چکانش دانش پیشنهادی کمترین mCE برابر با ٪۶۷/۲۴ را به دست می‌آورد که بهتر از روش‌های KD با ٪۷۰/۰۲، CRD با ٪۶۷/۵۰، و TeKAP با ٪۶۹/۶۱ است. در مقایسه با روش‌های پایه، مدل چکانش دانش پیشنهادی کاهش خطای پایداری را در تقریبا تمامی انواع اغتشاشات نشان می‌دهد. به‌طور خاص، تحت اغتشاشات نویزی مانند Gaussian و Shot noise، مدل چکانش دانش پیشنهادی خطای بسیار پایین‌تری ایجاد می‌کند و در برابر اغتشاشات دیجیتال مانند Pixel و Elastic نیز پیش‌بینی‌های پایدارتر و قابل اتکاتری نسبت به روش‌های جایگزین ارائه می‌دهد.

جدول (۶) : ارزیابی Few-shot روی CIFAR-100 با درصدهای مختلف داده، با استفاده از زوج معلم_دانش‌آموز / ResNet56 ResNet20

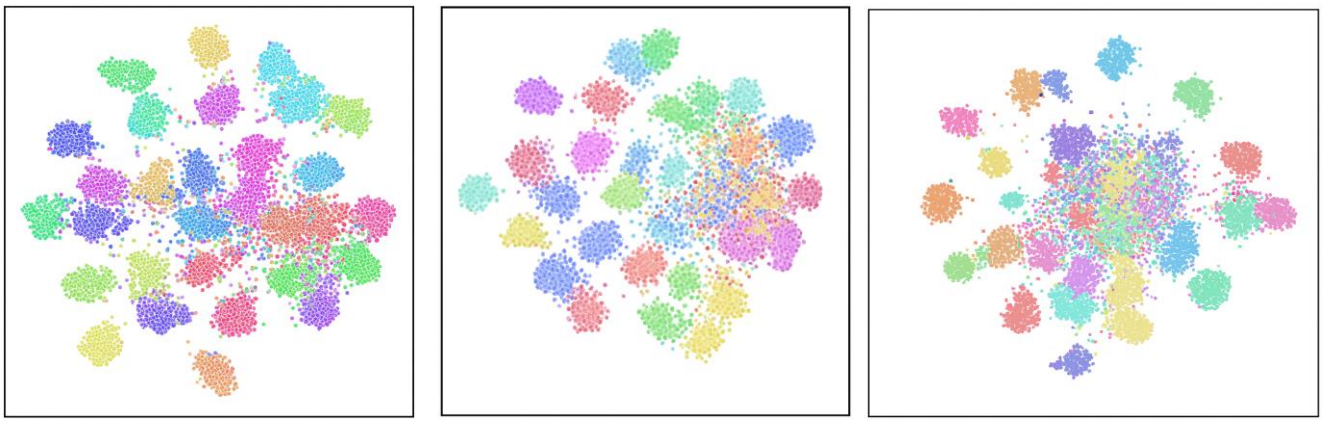
درصد	KD [۱]	FitNet [۱۰]	AT [۱۲]	CTKD [۱۸]	RKD [۴۰]	DKD [۲]	CRD [۵۳]	TeKAP [۲۷]	مدل چکانش دانش پیشنهادی
٪۲۵	۶۳/۱۶	۶۴/۲۸	۶۳/۹۷	۶۴/۵۶	۶۴/۵۴	۶۳/۸۹	۶۲/۷۲	۶۴/۳۹	۶۴/۸۹
٪۵۰	۶۷/۹۴	۶۸/۷۱	۶۸/۱۵	۶۸/۵۰	۶۷/۹۵	۶۸/۵۷	۶۶/۵۳	۶۸/۷۲	۶۹/۱۳
٪۷۵	۷۰/۰۷	۷۰/۱۵	۷۰/۰۸	۷۰/۲۸	۶۹/۸۴	۷۰/۶۶	۶۹/۴۸	۷۰/۳۵	۷۰/۶۹

جدول (۷) : عملکرد روش‌های چکانش دانش در وظایف انتقالی از CIFAR-100 به STL10 و ازCIFAR-100 به TinyImageNet

انتقال به مجموعه‌داده	KD [۱]	FitNet [۱۰]	AT [۱۲]	CTKD [۱۸]	RKD [۴۰]	DKD [۲]	CRD [۵۳]	TeKAP [۲۷]	مدل چکانش دانش پیشنهادی
از CIFAR-100 به STL10	۶۲/۳۳	۶۳/۲۰	۶۳/۵۳	۶۲/۷۳	۶۳/۹۸	۶۳/۴۳	۶۳/۵۳	۶۲/۹۵	۶۴/۰۳
از CIFAR-100 به TinyImageNet	۲۷/۸۸	۲۰/۱۳	۲۸/۶۸	۲۸/۵۷	۲۸/۵۸	۲۸/۱۲	۲۸/۲۷	۲۸/۳۶	۲۸/۷۹

جدول (۸) : ارزیابی پایداری مدل‌های دانشجوی ResNet-20 و ShuffleNetV1 در سناریوی از CIFAR-100 به CIFAR-100-C

روش	Noise			Blur				Weather			Digital					mCE↓
	Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	
ResNet20																
[1] KD	۸۹/۰۸	۸۳/۱۳	۸۳/۶۱	۶۱/۰۸	۸۷/۱۳	۶۹/۳۴	۶۸/۶۹	۶۲/۰۵	۶۸/۳۰	۶۶/۶۱	۴۰/۹۲	۷۶/۷۳	۶۱/۰۰	۶۶/۲۸	۶۶/۰۷	۷۰/۰۲
[۵۳] CRD	۸۶/۲۳	۸۰/۶۲	۷۷/۸۶	۵۸/۳۵	۸۶/۱۴	۶۵/۰۹	۶۴/۹۸	۶۰/۶۱	۶۷/۴۱	۶۳/۵۷	۳۹/۹۵	۷۰/۷۵	۶۱/۰۰	۶۵/۳۸	۶۳/۷۶	۶۷/۵۰
[۲۷] TeKAP	۸۸/۹۳	۸۲/۳۳	۸۱/۵۲	۶۱/۶۶	۸۷/۷۸	۶۹/۳۰	۷۰/۳۱	۶۲/۱۴	۶۳/۱۴	۶۵/۵۴	۴۰/۹۱	۷۶/۲۱	۶۱/۰۰	۶۳/۵۴	۶۴/۴۸	۶۹/۶۱
مدل چکانش دانش پیشنهادی	۸۶/۱۹	۸۰/۵۹	۷۷/۹۰	۵۸/۱۶	۸۵/۵۸	۶۵/۱۲	۶۴/۷۷	۶۱/۰۲	۶۸/۴۴	۶۱/۷۶	۴۱/۰۷	۷۲/۶۴	۵۹/۰۰	۶۳/۷۵	۶۲/۴۵	۶۷/۲۴
ShuffleNetV1																
[۱] KD	۸۶/۴۶	۷۹/۷۰	۷۳/۷۲	۵۵/۷۳	۸۳/۹۷	۶۵/۵۲	۶۲/۸۰	۵۶/۷۹	۶۰/۵۲	۶۰/۵۷	۳۷/۸۶	۷۲/۴۰	۵۵/۸۶	۵۳/۸۶	۵۶/۷۹	۶۴/۱۷
[۵۳] CRD	۸۵/۲۳	۷۹/۰۰	۷۸/۶۳	۵۴/۵۷	۸۶/۵۴	۶۲/۷۳	۶۲/۳۰	۵۸/۶۷	۶۶/۶۱	۵۸/۰۱	۳۷/۱۷	۶۸/۹۹	۵۶/۱۷	۶۳/۱۶	۶۲/۴۷	۶۴/۴۷
[۲۷] TeKAP	۸۶/۱۵	۷۸/۵۹	۷۳/۵۸	۵۵/۶۵	۸۱/۹۲	۶۳/۲۸	۶۳/۵۹	۵۲/۹۵	۵۷/۵۹	۵۸/۸۷	۳۵/۹۴	۷۱/۷۰	۵۴/۹۴	۵۱/۴۱	۵۴/۲۳	۶۲/۶۷
مدل چکانش دانش پیشنهادی	۸۴/۶۲	۷۷/۲۵	۷۲/۵۸	۵۲/۷۶	۸۱/۴۰	۶۰/۴۸	۵۹/۸۳	۵۳/۰۰	۵۷/۳۳	۵۶/۲۳	۳۵/۳۶	۶۸/۹۱	۵۳/۳۶	۵۲/۲۹	۵۵/۴۱	۶۱/۳۸



[۱] KD　　[۲۷] TeKAP　　مدل چکانش دانش پیشنهادی

شکل (۵) : بصری‌سازی t-SNE از بازنمایی‌های آموخته‌شده برای روش‌های مختلف چکانش دانش.

۷-۵- مقاومت در برابر حملات و اغتشاش‌ها (Adversarial Robustness)

برای ارزیابی بیشتر میزان مقاومت روش پیشنهادی، آزمایش‌هایی روی مجموعه‌داده CIFAR-100-C به CIFAR-100 انجام شده؛ مجموعه‌ای که شامل ۱۵ نوع اغتشاش رایج است و در چهار دسته Noise، Blur، Weather، و Digital طبقه‌بندی می‌شود.

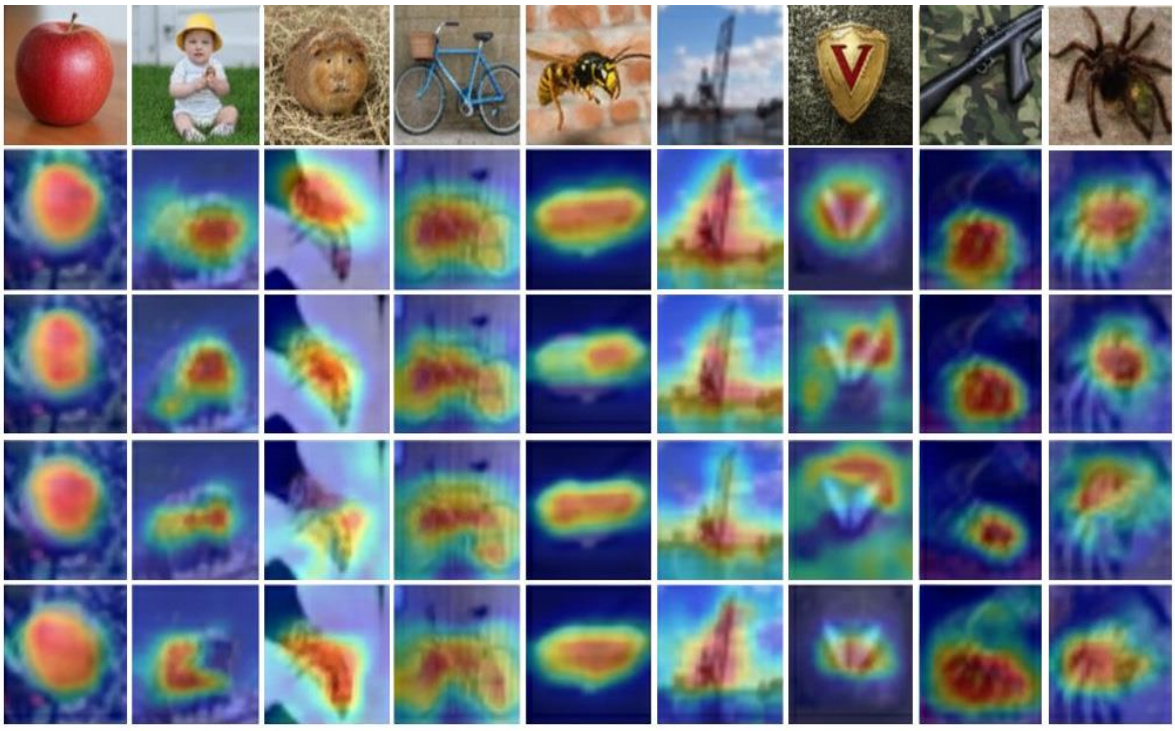
در این جا تعدادی از جفت‌مدل‌های جدول ۱ انتخاب شده که سه نوع نماینده دانش آموز شامل: ResNet20 / ResNet56، ShuffleNetV1 / ResNet32×4 هستند. با افزوده‌شدن نویز نشان می‌دهد که مدل چکانش دانش پیشنهادی در حفظ دانش حتی تحت اغتشاشات شدید ورودی مؤثرتر بوده و هم‌زمان انسجام درون‌کلاسی و جدایی میان‌کلاسی را حفظ می‌کند. اغتشاشی، عملکرد طبقه‌بندی تمام مدل‌های دانش آموزی پیش‌تمرین‌شده

به‌صورت ناگزیر تا حدی کاهش می‌یابد. میانگین خطای تست مدل‌های دانش آموز در مواجهه با این اغتشاشات و با استفاده از روش‌های مختلف چکانش گزارش شده و میزان خطای میانگین اغتشاش (mCE) در جدول ۸ خلاصه شده است. به‌طور مشابه، برای دانش آموزی ShuffleNetV1، مدل چکانش دانش پیشنهادی کمترین mCE برابر با ٪۶۱/۳۸ را تولید می‌کند که به‌طور محسوسی بهتر از KD با ٪۶۴/۱۷، CRD با ٪۶۴/۴۷، و TeKAP با ٪۶۲/۶۷ است. این بهبود در تمام چهار گروه اغتشاش مشاهده می‌شود و بزرگ‌ترین افزایش عملکرد در دسته‌های Noise و Digital دیده می‌شود. این نتایج نشان می‌دهد که مدل چکانش دانش پیشنهادی به‌ویژه در مقاوم‌سازی معماری‌های سبک‌وزن مؤثر است. در مجموع، این یافته‌ها نشان می‌دهند که مدل چکانش دانش پیشنهادی نه‌تنها دقت مدل را در حالت بدون اغتشاش افزایش می‌دهد، بلکه تاب‌آوری در برابر شرایط متخاصم را نیز بهبود می‌بخشد. مدل چکانش دانش پیشنهادی با هم‌ترازسازی دانش ساختاری در چندین دیدگاه مختلف، حساسیت مدل دانش آموز را نسبت به انواع اغتشاشات کاهش می‌دهد و در نتیجه توانایی تعمیم مدل را در شرایط واقعی تقویت می‌کند.

اگرچه ارزیابی کامل خارج از توزیع (Out-of-Distribution) در دامنه این مطالعه قرار نداشته است، نتایج روی CIFAR-100-C نشان می‌دهد چارچوب پیشنهادی در حضور انواع اغتشاش‌های نویزی، دیجیتال و جوی عملکرد پایدارتری نسبت به روش‌های پایه دارد. این یافته‌ها بیانگر آن است که منظم‌سازی چندسطحی پیشنهادی می‌تواند به بهبود تعمیم در شرایط تغییر توزیع ورودی کمک کند. بررسی سناریوهای OOD گسترده‌تر می‌تواند به‌عنوان مسیر پژوهشی آینده دنبال شود.

۸-۵- تحلیل‌های تفسیرپذیری t-SNE

از الگوریتم همسایگی تصادفی t-SNE [۵۴] استفاده شده که یک روش کاهش ابعاد غیرخطی که به‌طور گسترده برای نمایش داده‌های با ابعاد بالا به‌کار می‌رود. بصری‌سازی‌های t-SNE در شکل ۵ توزیع ویژگی‌های مدل معلم، KD، TeKAP و مدل چکانش دانش پیشنهادی را روی مجموعه‌داده CIFAR-100 نشان می‌دهد. برای یک مقایسه منصفانه، ویژگی‌های لایه ماقبل‌آخر استخراج شده و آن‌ها را با استفاده از همان ابرپارامترهای t-SNE به دو بعد نگاشت شده؛ ۳۰ کلاس نیز به‌صورت تصادفی برای نمایش انتخاب می‌شود. مدل KD، ویژگی‌های دانش‌آموز خوشه‌هایی نسبتا نامنظم و پراکنده تشکیل می‌دهند که دارای هم‌پوشانی‌های زیاد و مرزهای نامشخص هستند، و این امر باعث جدایش بین‌کلاسی ضعیف می‌شود. همچنین فاصله بین نمونه‌های هم‌کلاس بزرگ‌تر است و مقداری تداخل میان آنها دیده شده که موجب کاهش انسجام درون‌کلاسی و افت قابلیت تشخیص می‌گردد. مدل TeKAP تا حدی این مشکل را با ارائه راهنمای چندنمایی از سوی معلم کاهش می‌دهد و موجب ایجاد خوشه‌هایی فشرده‌تر و ساختارمندتر می‌شود؛ اما همچنان هم‌پوشانی و ابهام مرزی باقی است. در مقابل، مدل چکانش دانش پیشنهادی نمایش‌هایی با انسجام درون‌کلاسی بسیار بیشتر و حاشیه‌های بین‌کلاسی بزرگ‌تر ایجاد می‌کند؛ به‌طوری‌که خوشه‌ها جدا و منظم هستند، نمونه‌های پرت کاهش می‌یابند و مرزهای بین کلاس‌ها واضح‌تر و مطابق‌تر با توزیع ویژگی‌های معلم می‌شوند. این نتایج نشان می‌دهند که روش پیشنهادی به‌طور مؤثر هم چسبندگی محلی و هم جداسازی سراسری را در فضای ویژگی تقویت کرده و دانش‌آموز را قادر می‌سازد نمایش‌هایی متمایزتر و بهتر خلق کند.



شکل (۶) : نقشه‌های CAM برای زوج‌های مدل معلم و دانش‌آموز تحت نظارت روش‌های مختلف چکانش. تصاویر ورودی در ردیف اول نمایش داده شده‌اند وCAM های مدل معلم و روش‌های گوناگون در ردیف‌های بعدی نشان داده می‌شوند.

در ادامه نقشه‌های فعال‌سازی کلاس (CAMs) [۵۵] مدل معلم و مدل‌های دانش‌آموز روی مجموعه‌داده CIFAR-100 برای بررسی کیفیت توجه و نشان دادن تاثیر مدل چکانش دانش پیشنهادی بصری‌سازی شده است. همان‌گونه که در شکل ۶ نشان داده شده، ردیف اول تصاویر ورودی و ردیف‌های بعدی CAMهای مدل معلم، مدل KD، مدل TeKAP و مدل چکانش دانش پیشنهادی هستند.

در مقایسه با مدل KD و مدل TeKAP، مدل چکانش دانش پیشنهادی نواحی مهم‌تر و دقیق‌تری را فعال می‌کند که با نواحی مورد توجه معلم سازگاری بیشتری دارند. برای نمونه، در ستون دوم، مدلKD تنها بخش‌هایی پراکنده را فعال می‌کند و مدل TeKAPبخش‌هایی از پس‌زمینه را فعال می‌کند، در حالی که مدل چکانش دانش پیشنهادی به‌طور کامل بدن را برجسته می‌کند. در ستون چهارم، روش‌های پایه در تشخیص ساختار دوچرخه ناکام‌اند و روی نواحی نامرتبط تمرکز می‌کنند، در حالی‌که مدل چکانش دانش پیشنهادی هر دو بال را با ساختار کامل مشخص می‌سازد. در ستون پنجم، مدل KD و TeKAP توجه پراکنده‌ای روی پس‌زمینه نشان می‌دهند، اما مدل چکانش دانش پیشنهادی دقیقا ناحیه تنه زنبور را با مرزبندی واضح برافراز می‌کند. در ستون هشتم، توجه مدل KD و TeKAP از جسم اصلی منحرف شده و روی ماسه‌های اطراف پخش می‌شود، اما مدل چکانش دانش پیشنهادی کل بدن عقرب را به‌درستی بازیابی می‌کند و شباهت زیادی به توجه معلم دارد. با انتقال دانش ساختاری ریزدانه از معلم، مدل چکانش دانش پیشنهادی نه‌تنها تفسیرپذیری پیش‌بینی‌های مدل دانش‌آموز را ارتقا می‌دهد بلکه عملکرد آن را نیز روی دسته‌های گوناگون تقویت می‌کند.

۹-۵- مطالعه حذف مؤلفه‌ها

برای ارزیابی سهم هر مؤلفه در مدل چکانش دانش پیشنهادی، مطالعات حذف مؤلفه‌ها بر روی مجموعه‌داده CIFAR-100 و با استفاده از دو پیکربندی مختلف معلم- دانش آموز: یک جفت هم‌معماری (ResNet32×4 / ShuffleNetV2) و یک جفت نامتجانس (ResNet32×4 / ResNet8×4) انجام شد. در آزمایش اثر هر بلوک، سه ماژول پیشنهادی شامل ماژول چکانش توجه سه‌بعدی، ماژول ماسک خصمانه، و ماژول منظم‌سازی فضای کروی به‌طور جداگانه به چارچوب دانش‌چکانش پایه افزوده شدند. همان‌طور که در جدول ۹ نشان داده شده، اضافه کردن ماژول توجه سه‌بعدی باعث بهبود ٪۰/۷۷ و ٪۱/۴۹ در دو جفت دانش‌آموز- معلم ResNet32×4/ResNet8×4 و ResNet32×4/ShuffleNetV2 می‌شود. این نتایج نشان می‌دهد که به‌کارگیری توجه در سطح نورون در هنگام هم‌ترازسازی ویژگی‌ها، به دانش آموز امکان می‌دهد تمرکز تمایزبخش معلم را بهتر استخراج کرده و سیگنال‌های زائد را سرکوب کند. ماژول ماسک خصمانه نیز بهبودهای ٪۰/۶۵ و ٪۱/۳۷ را در این دو سناریو ایجاد می‌کند. این ماژول با تقسیم خصمانه ویژگی‌های دانش آموز به زیرفضاهای برتر و ضعیف‌تر، انتقال دانش ریزدانه‌تر را تسهیل کرده و اثر اطلاعات نامربوط را کاهش می‌دهد. ماژول منظم‌سازی فضای کروی نیز بهبودهای ٪۰/۲۶ و ٪۱/۵۶ را فراهم می‌کند که نشان‌دهنده نقش هم‌ترازی توزیع‌های ویژگی در فضای ابرکروی برای تقویت قابلیت تمایز و تعمیم مدل است.

آزمایش بعدی مربوط به اثر مکمل سه ماژول است. تأثیر مکمل سه ماژول زمانی آشکارتر می‌شود که ترکیب‌های دوتایی آن‌ها بررسی شود. همان‌طور که در جدول ۹ گزارش شده، ترکیب هر دو ماژول، همواره عملکرد بهتری نسبت به استفادهٔ مستقل از هر ماژول دارد. به‌عنوان مثال، ترکیب ماژول توجه سه‌بعدی با ماژول ماسک خصمانه (نسخه D) بهبود ٪۱/۳۹ و ٪۱/۶۴ را نسبت به خط پایه نشان می‌دهد که تأییدی بر تقویت متقابل نظارت نورونی محلی و جداسازی خصمانهٔ زیرفضا است. ترکیب توجه سه‌بعدی با ماژول منظم‌سازی فضای کروی (نسخه E) نیز بهبود ٪۱/۰۵ و ٪۱/۷۲ را کسب می‌کند که نشان‌دهنده هم‌افزایی میان توجه ریزدانه و هم‌ترازی هندسی سراسری است. در ترکیب سوم، ماژول ماسک خصمانه، و ماژول منظم‌سازی فضای کروی (نسخه F)، بیشترین بهبود یعنی ٪۱/۰۸ و ٪۱/۷۵ به دست می‌آید، که نشان می‌دهد هر دو ماژول با هم می‌توانند سازگاری ساختاری و هم‌ترازی سراسری را توأمان تقویت کنند.

جدول (۹) : مطالعه حذف مؤلفه‌ها با ماتریس انتخاب ماژول؛ علامت (✓) نشان می‌دهد که ماژول فعال است.

ماژول توجه	ماژول ماسک خصمانه	ماژول منظم‌سازی فضای کروی	ترکیب	ResNet32×4 ResNet8×4	ResNet32×4 ShuffleV2
-	-	-	مدل پایه	۷۳/۹۵	۷۲/۹۸
✓	-	-	نسخه A	۷۴/۱۰	۷۴/۴۷
-	✓	-	نسخه B	۷۳/۹۸	۷۴/۳۵
-	-	✓	نسخه C	۷۳/۵۹	۷۴/۵۴
✓	✓	-	نسخه D	۷۴/۷۲	۷۴/۶۲
✓	-	✓	نسخه E	۷۴/۳۸	۷۴/۷۰
-	✓	✓	نسخه F	۷۴/۴۱	۷۴/۷۳
✓	✓	✓	نسخه G	**۷۵/۲۷**	**۷۵/۱۳**

در مجموع، هر سه ماژول به‌صورت مکمل عمل کرده و توجه محلی، انتخاب پیکسل‌ها و هم‌ترازی ساختاری سراسری را به‌طور مشترک تقویت می‌کنند که نهایتا منجر به انتقال دانش ریزدانه‌تر و مقاوم‌تر، و افزایش توان یادگیری و عملکرد نهایی دانش‌آموز می‌شود.

۶- نتیجه‌گیری

در این مقاله، یک چارچوب چنددانه‌ای برای چکانش دانش ارائه شد که با ترکیب سه ماژول مکمل شامل توجه سه‌بعدی، ماسک خصمانه و تنظیم کروی فضا، انتقال هم‌زمان دانش محلی، زیرفضایی و ساختار سراسری معلم را امکان‌پذیر می‌سازد. این طراحی موجب شد دانش‌آموز علاوه بر تقلید ویژگی‌های تمایزبخش، ساختار هندسی کلی فضای بازنمایی معلم را نیز فراگیرد. آزمایش‌ها روی CIFAR-100، STL10 و TinyImageNet نشان داد روش پیشنهادی به‌طور مداوم عملکردی بهتر از روش‌های پیشرفته ارائه می‌دهد و در برابر اغتشاشات مختلف در CIFAR-100-C مقاوم‌تر است. تحلیل‌های t-SNE، CAM و مطالعات حذف مؤلفه‌ها نیز نشان دادند هر ماژول به‌صورت مستقل مؤثر است، اما ترکیب آن‌ها بیشترین بهبود و هم‌افزایی را ایجاد می‌کند. به‌طور کلی، نتایج نشان می‌دهد انتقال دانش چندسطحی می‌تواند دقت، پایداری و تعمیم‌پذیری مدل‌های سبک را بهبود دهد و مسیر توسعه آن در کاربردهای گسترده‌تر آینده قابل پیگیری است.

با وجود بهبودهای مشاهده‌شده، روش پیشنهادی در برخی سناریوها با محدودیت‌هایی مواجه است. به‌طور خاص، در کلاس‌هایی با شباهت بصری بسیار بالا یا در حضور نویز شدید، فاصله عملکرد نسبت به برخی روش‌های رابطه‌ای کاهش می‌یابد. همچنین در تنظیمات با ظرفیت بسیار محدود دانش‌آموز، بهره‌گیری کامل از نظارت چندسطحی ممکن است دشوار باشد. تحلیل کیفی نشان می‌دهد در چنین مواردی، تمایز میان نمونه‌های مرزی همچنان چالش‌برانگیز باقی می‌ماند.

مراجع

[۱] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531,* 2015.

[۲] M. Sepahvand, F. Abdali-Mohammadi, and A. Taherkordi, "Teacher–student knowledge distillation based on decomposed deep feature representation for intelligent mobile applications," *Expert Systems with Applications,* vol. 202, p. 117474, 2022.

[۳] M. Sepahvand, F. Abdali-Mohammadi, and A. Taherkordi, "An adaptive teacher–student learning algorithm with decomposed knowledge distillation for on-edge intelligence," *Engineering Applications of Artificial Intelligence,* vol. 117, p. 105560, 2023.

[۴] M. Sepahvand and F. Abdali-Mohammadi, "Joint learning method with teacher–student knowledge distillation for on-device breast cancer image classification," *Computers in Biology and Medicine,* vol. 155, p. 106476, 2023.

[۵] M. Sepahvand and F. Abdali-Mohammadi, "A novel method for reducing arrhythmia classification from 12-lead ECG signals to single-lead ECG with minimal loss of accuracy through teacher-student knowledge distillation," *Information Sciences,* vol. 593, pp. 64-77, 2022.

[۶] M. Mardanpour, M. Sepahvand, F. Abdali-Mohammadi, M. Nikouei, and H. Sarabi, "Human activity recognition based on multiple inertial sensors through feature-based knowledge distillation paradigm," *Information Sciences,* vol. 640, p. 119073, 2023.

[۷] Y. Li, Y. Wang, and D. Li, "Privacy-preserving lightweight face recognition," *Neurocomputing,* vol. 363, pp. 212-222, 2019.

[۸] S. W. Lim, C. S. Chan, E. R. M. Faizal, and K. H. Ewe, "Progressive expansion: Cost-efficient medical image analysis model with reversed once-for-all network training paradigm," *Neurocomputing,* vol. 581, p. 127512, 2024.

[۹] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," *In Proceedings of the AAAI conference on artificial intelligence*, 2020, vol. 34, 04 ed., pp. 5191-5198.

[۱۰] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," presented at the International Conference Learning Representation (ICLR), 2014.

[۱۱] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133-4141.

[۱۲] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928,* 2016.

[۱۳] M. Yuan, B. Lang, and F. Quan, "Student-friendly knowledge distillation," *Knowledge-Based Systems,* vol. 296, p. 111915, 2024.

[۱۴] T. Huang *et al.*, "Masked distillation with receptive tokens," *arXiv preprint arXiv:2205.14589,* 2022.

[۱۵] W. Zhang, D. Liu, W. Cai, and C. Ma, "Cross-view consistency regularisation for knowledge distillation," *Association for Computing Machinery*, pp. 2011-2020, 2024.

[۱۶] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699,* 2019.

[۱۷] J. Guo, M. Chen, Y. Hu, C. Zhu, X. He, and D. Cai, "Reducing the teacher-student gap via spherical knowledge disitllation," *arXiv preprint arXiv:2010.07485,* 2020.

[۱۸] Z. Li *et al.*, "Curriculum temperature for knowledge distillation," 2023, vol. 37, 2 ed., pp. 1504-1512 .

[۱۹] Z. Chi *et al.*, "Normkd: Normalized logits for knowledge distillation," *arXiv preprint arXiv:2308.00520,* 2023.

[۲۰] S. Sun, W. Ren, J. Li, R. Wang, and X. Cao, "Logit standardization in knowledge distillation," *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15731-15740. 2024.

[۲۱] K. Zheng and E.-H. Yang, "Knowledge distillation based on transformed teacher matching," *arXiv preprint arXiv:2402.11148,* 2024.

[۲۲] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," *In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953-11962, 2022.

[۲۳] J. Cui, Z. Tian, Z. Zhong, X. Qi, B. Yu, and H. Zhang, "Decoupled kullback-leibler divergence loss," *Advances in Neural Information Processing Systems,* vol. 37, pp. 74461-74486, 2024.

[۲۴] S. Wei, C. Luo, and Y. Luo, "Scaled decoupled distillation," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15975-15983 .

[۲۵] W. Son, J. Na, J. Choi, and W. Hwang, "Densely guided knowledge distillation using multiple teacher assistants," *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9395-9404 .

[۲۶] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," *PMLR*, 2018, pp. 1607-1616 .

[۲۷] M. I. Hossain, S. Akhter, C. S. Hong, and E.-N. Huh, "Single teacher, multiple perspectives: Teacher knowledge augmentation for enhanced knowledge distillation," *In The Thirteenth International Conference on Learning Representations*, 2025 .

[۲۸] G. Xu, Z. Liu, X. Li, and C. C. Loy, "Knowledge distillation meets self-supervision," 2020: Springer, pp. 588-604 .

[۲۹] C. Shu, Y. Liu, J .Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," *In Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5311-5320 .

[۳۰] Z. Liu, Y. Wang, X. Chu, N. Dong, S. Qi, and H. Ling, "A simple and generic framework for feature distillation via channel-wise transformation," *In Proceedings of the IEEE/CVF International Conference on Computer Vision* , 2023, pp. 1129-1138 .

[۳۱] T. Liu, C. Chen, X. Yang, and W. Tan, "Rethinking knowledge distillation with raw features for semantic segmentation," *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1155-1164 .

[۳۲] J. Yuan, M. H. Phan, L. Liu, and Y. Liu, "Fakd: Feature augmented knowledge distillation for semantic segmentation," *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* , 2024, pp. 595-605 .

[۳۳] Y. Zhang, T. Huang, J. Liu, T. Jiang, K. Cheng, and S. Zhang, "Freekd: Knowledge distillation via semantic frequency prompt," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* , 2024, pp. 15931-15940 .

[۳۴] X. Liu, L. Li, C. Li, and A. Yao" ,Norm: Knowledge distillation via n-to-one representation matching," *arXiv preprint arXiv:2305.13803,* 2023.

[۳۵] P. Passban, Y. Wu, M. Rezagholizadeh, and Q. Liu, "Alp-kd: Attention-based layer projection for knowledge distillation," *In Proceedings of the AAAI Conference on artificial intelligence* , 2021, vol. 35, 15 ed., pp. 13657-13665 .

[۳۶] J. Gou, L. Sun, B. Yu, S. Wan, and D. Tao, "Hierarchical multi-attention transfer for knowledge distillation," *ACM Transactions on Multimedia Computing, Communications and Applications,* vol. 20, no. 2, pp. 1-20, 2023.

[۳۷] J. Gou, L. Sun, B. Yu, S. Wan, W. Ou, and Z. Yi, "Multilevel attention-based sample correlations for knowledge distillation," *IEEE Transactions on Industrial Informatics,* vol. 19, no. 5, pp. 7099-7109, 2022.

[۳۸] Z. Tao, H. Li, J. Zhang, and S. Zhang, "Multi-level knowledge distillation via dynamic decision boundaries exploration and exploitation," *Information Fusion,* vol. 112, p. 102586, 2024/12/01/ 2024, doi: https://doi.org/10.1016/j.inffus.2024.102586.

[۳۹] S. You, C. Xu, C. Xu, and D. Tao, "Learning from multiple teacher networks," *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1285-1294 .

[۴۰] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967-3976 .

[۴۱] Y. Liu *et al.*, "Knowledge distillation via instance relationship graph," *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7096-7104 .

[۴۲] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," *Proceedings of the IEEE/CVF international conference on computer vision*, 2022, pp. 12319-12328 .

[۴۳] H. Hu, H. Zeng, Y. Xie, Y. Shi, J. Zhu, and J. Chen, "Global Instance Relation Distillation for convolutional neural network compression," *Neural Computing and Applications,* vol. 36, no. 18, pp. 10941-10953, 2024.

[۴۴] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," *Advances in Neural Information Processing Systems,* vol. 35, pp. 33716-33727, 2022.

[۴۵] Z. Zhang, C. Zhou, and Z. Tu, "Distilling inter-class distance for semantic segmentation," *arXiv preprint arXiv:2205.03650,* 2022.

[۴۶] A. M. Mansourian, R. Ahamdi, and S. Kasaei, "Aicsd: adaptive inter-class similarity distillation for semantic segmentation," *Multimedia Tools and Applications,* pp. 1-20, 2025.

[۴۷] C. Wang *et al.*, "Prrd: Pixel-region relation distillation for efficient semantic segmentation," *2023: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, pp. 1-5.

[۴۸] Q. Wang, L .Liu, W. Yu, S. Chen, J. Gong, and P. Chen, "BCKD: block-correlation knowledge distillation," *In 2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3225-3229 .

[۴۹] C. Wang, J. Zhong, Q. Dai, R. Li, Q. Yu, and B. Fang, "Local structure consistency and pixel-correlation

distillation for compact semantic segmentation," *Applied Intelligence,* vol. 53, no. 6, pp. 6307-6323, 2023.

[۵۰] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images.(2009)," ed, 2009.

[۵۱] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," 2011: J*MLR Workshop and Conference Proceedings*, pp. 215-223 .

[۵۲] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*, pp. 248-255 .

[۵۳] J. Yang, X. Zhu, A. Bulat, B. Martinez, and G. Tzimiropoulos, "Knowledge distillation meets open-set semi-supervised learning," *International Journal of Computer Vision,* vol. 133, no. 1, pp. 315-334, 2025.

[۵۴] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems,* vol. 15, 2002.

[۵۵] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921-2929

مجید سپهوند استادیار گروه مهندسی کامپیوتر دانشگاه اراک است. ایشان دارای مدرک دکترای مهندسی کامپیوتر با تمرکز بر پردازش داده‌های زیست‌پزشکی و هوش مصنوعی هستند. وی به عنوان پژوهشگر با دانشگاه ملاردالن سوئد (Mälardalen University) همکاری داشته است. زمینه‌های پژوهشی ایشان شامل پردازش داده‌های زیست‌پزشکی، هوش مصنوعی در سلامت دیجیتال، تحلیل رفتار و الگوهای انسانی، یادگیری عمیق و توسعه مدل‌های هوشمند برای تحلیل سیگنال‌های فیزیولوژیک و تشخیص‌های پزشکی است.