دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی‌ارشد

گرایش هوش مصنوعی و رباتیکز

شناسایی اعمال با روش یادگیری مستمر

نگارش

پریسا ملاحسینی

استاد راهنما

جناب آقای دکتر محمد رحمتی

مرداد ۱۴۰۴


بسم الله الرحمن الرحيم

صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تایید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

نکات مهم:

- نگارش پایان نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت پشت و رو(دورو) بلامانع است و انجام آن توصیه می شود.

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

به نام خدا

تعهدنامه اصالت اثر

تاریخ: مرداد ۱۴۰۴

اینجانب **پریسا ملاحسینی** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

پریسا ملاحسینی

پریسا ملاحسینی

مرداد ۱۴۰۴

نویسنده پایان نامه، در صورت تمایل میتواند برای سپاسگزاری پایان نامه خود را به شخص یا اشخاص و یا ارگان خاصی تقدیم نماید.

سپاس‌گزاری

نویسنده پایان‌نامه می‌تواند مراتب امتنان خود را نسبت به استاد راهنما و استاد مشاور و یا دیگر افرادی که طی انجام پایان‌نامه به نحوی او را یاری و یا با او همکاری نموده‌اند ابراز دارد.

پریسا ملاحسینی

مرداد ۱۴۰۴

چکیده

در سال‌های اخیر، یادگیری ماشین و به‌ویژه شبکه‌های عصبی عمیق پیشرفت چشمگیری را تجربه کرده‌اند و توانسته‌اند در حوزه‌هایی همچون بینایی ماشین، پردازش زبان طبیعی و تشخیص گفتار، به سطح عملکردی نزدیک به انسان دست یابند. دستیابی به چنین عملکردی معمولاً نیازمند پیش‌آموزش این شبکه‌ها بر روی مجموعه‌داده‌های بزرگ است، فرآیندی که امکان بهره‌برداری مجدد از مدل‌های آموزش‌دیده را در وظایف گوناگون فراهم می‌سازد. بنابراین، در بسیاری از کاربردهای واقعی، داده‌ها به‌صورت تدریجی و در قالب وظایف متوالی در دسترس قرار می‌گیرند که لزوم استفاده از رویکردهای یادگیری پیوسته را پررنگ می‌سازد. یکی از چالش‌های اساسی این حوزه، فراموشی فاجعه‌بار است که موجب افت شدید عملکرد مدل روی وظایف گذشته، پس از یادگیری وظایف جدید، می‌شود. اخیرا با وجود توسعه‌ی روش‌های مبتنی‌بر مدل‌های زبانی بزرگ و مدل‌های بینایی-زبان، محدودیت‌هایی مانند مصرف بالای حافظه همچنان پابرجاست. در این پژوهش، روشی با عنوان ProActionCLIP برای یادگیری پیوسته در داده‌های ویدیویی ارائه شده است که ترکیبی از قابلیت‌های مدل بینایی-زبان Open-VCLIP در استخراج ویژگی‌های ویدیو و سازوکار پرامپت‌های یادگیرنده در مدل L2P را به‌کار می‌گیرد. این ترکیب، بدون نیاز به تغییر پارامترهای اصلی مدل پایه‌ی Open-VCLIP، امکان انطباق با وظایف متوالی را فراهم کرده و با بهینه‌سازی انتخاب پرامپت‌ها، از فراموشی فاجعه‌بار جلوگیری می‌کند. نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی علاوه بر حفظ دانش پیشین و یادگیری مؤثر دانش جدید، از نظر مصرف حافظه و منابع محاسباتی نیز کارایی بالایی دارد و می‌تواند به‌عنوان راهکاری مؤثر برای یادگیری پیوسته در حوزه‌ی تشخیص حرکت انسان در ویدیو مورد استفاده قرار گیرد.

واژه‌های کلیدی:

یادگیری پیوسته، مدل بینایی-زبان، یادگیری پرامپت، فراموشی فاجعه‌بار

فهرست مطالب

عنوان صفحه

فهرست شکل‌ها

شکل صفحه

فهرست جدول‌ها

جدول صفحه

فهرست نمادها

نماد	مفهوم
E	ماتریس کلید سازوکار توجه
V	ماتریس مقدار سازوکار توجه
q	بردار پرسمان
d	عامل نرمال‌سازی
y'	خروجی توجه
θ	وزن‌های مدل
λ	ضریب منظم‌سازی
α	ضریب ترکیب وزن‌های دو مدل
β	ضریب تنظیم‌کننده برای درون‌یابی
D	داده
d_{in}	ابعاد ویژگی تعبیه‌ی ورودی
d_k	ابعاد کلید
J	تعداد گام آموزش برای میانگین‌گیری
M	اندازه‌ی استخر پرامپت
P	پرامپت
L_p	تعداد نشانه‌ی پرامپت
R	اعداد حقیقی
K	مجموعه‌ی کلیدها
K_x	بهترین کلیدهای انتخاب‌شده برای ورودی موردنظر
z	استخراج‌کننده‌ی ویژگی برای مقایسه با کلیدها
f	استخراج‌کننده‌ی ویژگی با مدل پیش‌آموزش
γ	تابع برای سنجش تطابق بین بردار ویژگی ورودی و کلید پرامپت
N	تعداد کلید انتخابی

f_r	تابع بخش باقی‌مانده از مدل پیش‌آموزش‌دیده
x_p	بردار حاصل از اتصال پرامپت به بردار تعبیه‌ی ورودی
x	ورودی
w	وزن اهمیت معیار نزدیکی کلید و ویژگی
g_ϕ	خروجی دسته‌بند
T	قاب‌های ویدیو
t	یک قاب ویدیو
a	لایه‌ی ترنسفورمر
τ	ضریب مقیاس‌دهی
y	برچسب واقعی داده

فصل اول

مقدمه

یادگیری ماشین و به‌ویژه شبکه‌های عصبی عمیق[1] در سال‌های اخیر پیشرفت چشمگیری داشته‌اند و توانسته‌اند در حوزه‌های مختلفی مانند بینایی ماشین، پردازش زبان طبیعی و تشخیص گفتار، به عملکردی نزدیک به انسان دست یابند [۱، ۲، ۳، ۴]. برای اینکه این شبکه‌ها بتوانند به عملکرد چشم‌گیری دست پیدا کنند، معمولاً به پیش‌آموزش در مقیاس بزرگ نیازمند هستند [۱، ۵]. این پیش‌آموزش امکان استفاده‌ی دوباره‌ی این شبکه‌ها را در وظایف[2] بعدی فراهم می‌کند [۶]. با وجود این پیشرفت‌ها، روش‌های متداول، برای آموزش مدل‌ها از تمام داده‌ها به‌صورت هم‌زمان استفاده می‌کنند [۷]. این فرض در بسیاری از کاربردهای واقعی برقرار نیست. در واقع بسیاری از سناریوها وجود دارند که در آن داده‌ها به مرور زمان و در قالب وظایف متفاوت در دسترس قرار می‌گیرند. اخیرا دسته‌ای از روش‌های یادگیری ارائه شده اند که برای مواجهه با این سناریوها بکار گرفته می‌شوند. این روش‌های یادگیری با عنوان یادگیری پیوسته[3] شناخته می‌شوند [۸، ۹]. در این نوع یادگیری، مدل باید بتواند دانش جدید را بیاموزد، بدون آنکه دانش پیشین خود را از دست بدهد. یکی از چالش‌های اساسی در این زمینه، فراموشی فاجعه‌بار[4] [۱۰] است که باعث می‌شود مدل پس از یادگیری وظایف جدید، عملکرد خود را روی وظایف قبلی به‌طور چشمگیری از دست بدهد. براساس [۸]، رویکردهای یادگیری پیوسته به‌طور کلی به چهار دسته تقسیم می‌شوند: رویکردهای مبتنی بر تنظیم که با منظم‌سازی وزن‌ها تعادل بین یادگیری وظایف جدید و حفظ وظایف قبلی را برقرار می‌کنند؛ رویکردهای مبتنی بر بازپخش که با ذخیره یا بازتولید داده‌ها و ویژگی‌های وظایف گذشته و ترکیب آن‌ها با داده‌های جدید از فراموشی جلوگیری می‌کنند؛ رویکردهای مبتنی بر بهینه‌سازی که با اصلاح فرآیند به‌روزرسانی پارامترها مانع تغییرات ناسازگار با وظایف قبلی می‌شوند؛ و رویکردهای مبتنی بر معماری که با طراحی یا اختصاص بخش‌های خاص مدل برای هر وظیفه، تداخل وظایف را کاهش داده و حفظ دانش پیشین را تسهیل می‌کنند. این روش‌ها با وجود آن‌که توانسته اند به نتایج قابل قبولی در حوزه‌ی یادگیری پیوسته دست یابند، با چالش‌هایی نیز مواجه هستند. به عنوان مثال، در روش مبتنی بر بازپخش، حافظه‌ی مصرفی با زیادشدن تعداد وظایف، افزایش می‌یابد. این محدودیت حافظه، سبب می‌شود که تعداد وظایف نتواند از حد خاصی بیشتر شود. در روش مبتنی بر معماری نیز محدویت حافظه می‌تواند به‌طور مشابه رخ دهد. به این ترتیب که با اضافه شدن وظایف جدید، بایستی لایه‌های جدیدی به شبکه اضافه شود که این امر نیازمند تخصیص حافظه می‌باشد. در روش‌های مبتنی بر تنظیم نیز، محدودیت ظرفیت شبکه سبب محدود شدن تعداد وظایفی می‌شود که شبکه می‌تواند آن

[1] Deep Nueral Networks (DNNs)
[2] Tasks
[3] Continual learning
[4] Catastrophic forgetting

را یاد بگیرد [۸، ۹].

اخیرا، معرفی مدل‌های مبتنی بر سازوکار توجه [5] [۴، ۱]، مدل‌های زبانی بزرگ [6] [۱۱، ۱۲] و مدل‌های بینایی-زبان [7] [۱۳، ۱۴]، زمینه را برای ارائه‌ی روش‌های یادگیری پیوسته فراهم آوردند[۱۵]. مدل‌های زبانی بزرگ و مدل‌های بینایی-زبان، نوعی شبکه‌ی عصبی مبتنی بر سازوکار توجه هستند که با آموزش بر روی حجم بسیار زیادی از داده، توانایی تولید، درک و تحلیل زبان طبیعی [8] و تصویری را به‌دست آورده‌اند. در پی این موفقیت‌ها، تحقیقات متعددی به بهره‌گیری از آن‌ها در یادگیری پیوسته، نیز پرداخته‌اند [۱۵، ۷، ۱۶، ۱۷، ۱۸، ۱۹]. با وجود آنکه استفاده از مدل‌های زبانی بزرگ و مدل‌های بینایی-زبان، سبب افزایش صحت در وظایف یادگیری شده، چالش‌های روش‌های پیشین یادگیری پیوسته را برطرف نکرده است. در واقع به علت حجیم بودن این مدل‌ها، مشکلاتی مانند محدودیت حافظه پررنگ‌تر نیز شده است. از این رو برای مواجهه با این محدودیت، تمرکز محققین به سمت روش‌های مبتنی بر معماری رفته است [۱۵]. ایده‌ی اصلی این روش‌ها آن است که با ایجاد تغییر در ساختار مدل، از فراموشی فاجعه‌بار جلوگیری کنند. این تغییر در ساختار مدل‌های زبانی بزرگ شامل تنظیم پرامپت [9]، تنظیم پیشوند [10]، سازگاری رتبه پایین [11]، وفق‌دهنده [12]، مخلوط خبره‌ها [13] می‌شود [۱۵]. در این میان، مدل L2P[۷] با بهره‌گیری از تنظیم پرامپت، توانسته است به نتایج برجسته‌ای نسبت به سایر روش‌های مشابه دست یابد. این مدل، روشی برای یادگیری پیوسته معرفی می‌کند که به جای تغییر پارامترهای اصلی مدل CLIP، از گروهی پرامپت قابل‌آموزش (استخر پرامپت [14]) استفاده می‌کند. به طوری‌که در هر گام از آموزش، پرامپت‌های مشابه داده‌های جدید، انتخاب و به‌روزرسانی می‌شوند تا مدل بتواند دانش تازه را بیاموزد؛ بدون آنکه دانش قبلی را فراموش کند. این رویکرد، تعادلی مؤثر بین حفظ دانش گذشته و یادگیری وظایف جدید ایجاد کرده است. با وجود آن‌که این مدل عملکرد خوب به همراه رفع چالش محدودیت حافظه در یادگیری پیوسته، کسب کرده است، تنها برای داده‌های تصویری قابل استفاده است. در زمینه‌ی استفاده از مدل‌های بینایی-زبان برای داده‌های ویدیویی نیز مطالعات متعددی انجام شده

[5] Attention mechanism
[6] Large Language Models (LLMs)
[7] Vision–Language Models (VLMs)
[8] Natural language processing
[9] Prompt tuning
[10] Prefix tuning
[11] low-rank adaptation (LoRA)
[12] Adapter
[13] mixture of experts (MoE)
[14] Prompt pool

است که در این میان، می‌توان به مدل Open-VCLIP[۲۰] به‌عنوان یکی از برترین روش‌ها از نظر عملکرد و بهینه‌بودن حجم مدل در زمینه‌ی درک داده‌ی ویدیویی اشاره کرد. مدل Open-VCLIP با توسعه‌ی معماری CLIP، امکان تحلیل ویدیو را فراهم می‌کند و با بهره‌گیری از تکنیک‌هایی، ضمن یادگیری دانش جدید حاصل از داده‌های ویدیویی، از فراموشی دانش مدل CLIP جلوگیری کرده و تعمیم‌پذیری مدل را بهبود می‌بخشد. هدف اصلی این تحقیق، ارائه‌ی رویکردی کارآمد برای یادگیری پیوسته در داده‌های ویدیویی است که بتواند با حداقل منابع محاسباتی، هم دانش پیشین را حفظ کند و هم دانش جدید را بیاموزد. برای تحقق این امر، روش پیشنهادی ProActionCLIP با ترکیب قابلیت‌های Open-VCLIP در استخراج ویژگی‌های داده‌ی ویدیویی و سازوکار پرامپت‌های یادگیرنده در L2P، توسعه یافته است. این ترکیب، بدون نیاز به تغییر مستقیم پارامترهای مدل Open-VCLIP، امکان تطبیق مدل با وظایف پیوسته را فراهم کرده، مشکل فراموشی فاجعه‌بار را کاهش می‌دهد، نیاز حافظه را به حداقل می‌رساند. نتایج آزمایش‌های تجربی بر روی مجموعه‌داده‌های UCF101[۲۱] و HMDB1[۲۲] نشان می‌دهد که مدل ProActionCLIP توانسته است میانگین صحت و میزان فراموشی را نسبت به سایر روش‌های مشابه، بهبود ببخشد.

ساختار نگارش این تحقیق به این صورت است که در فصل ۲، کارهای پیشین در حوزه یادگیری پیوسته مرور و رویکردهای موجود دسته‌بندی می‌شوند؛ در فصل ۳ روش پیشنهادی معرفی و جزئیات فنی آن بررسی می‌شود؛ در فصل ۴ نتایج آزمایش‌های تجربی ارائه و عملکرد روش پیشنهادی با روش‌های موجود مقایسه می‌شود و در فصل ۵ جمع‌بندی نتایج انجام و پیشنهادهایی برای کارهای آینده مطرح خواهد شد.

فصل دوم

مرور کارهای پیشین

۲-۱ مقدمه

در این فصل به معرفی یادگیری پیوسته در دو حوزه‌ی تصویر و ویدیو و مدل‌های بینایی-زبان در این حوزه‌ها می‌پردازیم. روش‌های یادگیری پیوسته به بخش‌های مبتنی بر تنظیم، حافظه، بهینه‌سازی و معماری تقسیم می‌شوند که به مرور هرکدام از دسته‌های فوق می‌پردازیم.

۲-۲ یادگیری پیوسته

یادگیری پیوسته به توانایی یک سامانه‌ی هوشمند برای کسب، به‌روزرسانی، جمع‌آوری و بهره‌برداری از دانش در طول عمر آن اشاره دارد. این، شامل یادگیری یک دنباله از مطالب یا وظایف یکی پس از دیگری و سازگاری با اطلاعات جدید بدون فراموشی دانش قبلاً آموخته شده است. به علت تدریجی اضافه شدن مطالب، یادگیری مداوم به عنوان یادگیری افزایشی[۱] ، یادگیری مستمر[۲] و یادگیری مادام‌العمر[۳] نیز شناخته می‌شود. هدف یادگیری پیوسته حفظ تعادل بین یادگیری اطلاعات جدید و حفظ دانش قبلاً کسب شده، با غلبه بر چالش فراموشی فاجعه‌بار است [۸، ۹]. در ادامه به معرفی چالش ها، سناریوها، رویکردها و کاربردهای این حوزه پرداخته می‌شود.

چالش اصلی که در این نوع یادگیری به وجود می‌آید، چالشی با نام فراموشی فاجعه بار است. علت این امر این است که وقتی داده های جدید برای یادگیری اضافه می‌شوند، سامانه مجبور می‌شود اطلاعات قبل را تا حدی به فراموشی بسپرد و شرایط جدید را در نظر بگیرد. برای همین به نوعی یک تعادل بین حالتی که حافظه ثابت است و حالتی که یادگیری انعطاف‌پذیر است، نیاز می‌باشد. در واقع هدف بر این است که یادگیری مستمر تعمیم‌پذیری خوبی برای تطبیق در شرایط مختلف با داده‌های جدید و با توزیع‌های جدید داشته باشد، همچنین کارایی منابع را نیز تضمین کند [۸، ۹] .

هم‌چنین، بر اساس نوع داده‌ها و وظیفه‌هایی که باید انجام شود، سناریوهای مختلفی برای یادگیری پیوسته ارائه شده است [۸]:

[1] Incremental learning
[2] Continuous learning
[3] Lifelong learning

- **یادگیری نمونه‌ای افزاینده**

در یادگیری نمونه‌ای-افزاینده [۴] نمونه های داده ی جدید به طور پیوسته به مدل معرفی می شوند که هر یک نقطه داده جدید را نشان می دهد. این مدل باید با توزیع داده‌های در حال تکامل سازگار شود و در عین حال نمونه‌های جدید را نیز در نظر بگیرد.

- **یادگیری افزایشی دامنه**

در یادگیری افزایشی دامنه [۵] ، دامنه تغییر می‌کند و به معنای افزایش یا تغییر در توزیع داده‌ها، ویژگی‌ها یا محیط‌ها در مسئله‌ای که مدل در حال یادگیری آن است، می‌باشد. این تغییر می‌تواند به صورت افزودن داده‌های جدید از دامنه‌های جدید، تغییر در ویژگی‌ها، یا حتی تغییر مفهوم برخی اجزاء از داده‌ها (مثلاً تغییر تعبیر برچسب‌ها) اتفاق بیافتد. چالش این است که اطمینان حاصل شود که مدل می‌تواند با این حوزه‌های جدید سازگار شود، بدون اینکه عملکرد آن در حوزه‌هایی که قبلا دیده شده‌اند، کاهش یابد [۹، ۲۳، ۲۴].

- **یادگیری افزایشی وظیفه**

در یادگیری افزایشی وظیفه [۶] وظایف جدید در طول زمان معرفی می‌شوند و مدل باید یاد بگیرد که در هر وظیفه ی جدید به خوبی عمل کند و در عین حال عملکرد خود را در وظایفی که قبلاً آموخته‌است حفظ کند. در این نوع یادگیری، معمولا شناسه‌ی وظیفه [۷] قبل از ورود داده به مدل مشخص می‌شود. به عنوان مثال، مشخص می‌شود که کدام وظیفه از مدل، باید با داده‌ی فعلی پیش‌بینی انجام دهد. [۲۳].

- **یادگیری افزایشی دسته‌ای**

دسته های جدید به تدریج به داده‌های آموزشی مدل اضافه می شوند. مدل باید بیاموزد که دسته های جدید را تشخیص دهد و بین آنها تمایز قائل شود بدون اینکه دانش خود را در مورد دسته های آموخته شده قبلی فراموش کند [۹، ۲۳، ۲۵، ۲۶].

[4] Instance incremental learning
[5] Domain incremental learning
[6] Task incremental learning
[7] Task identity

همانطور که پیش‌تر گفته شد، رویکردهای متفاوتی برای یادگیری پیوسته ارائه شده است. ونگ و همکاران [۸] ، به صورت زیر رویکردها را تقسیم‌بندی کرده‌اند:

۲-۲-۱ رویکرد مبتنی بر تنظیم

این رویکرد بر اساس این است که از تکنیک های مبتنی بر تنظیم [8] برای ایجاد تعادل در وظایف قدیم و جدید استفاده کند. در حالت کلی نیز تکنیک‌های منظم سازی با ایجاد تغییراتی در محاسبات باعث جلوگیری از بیش برازش مدل می شوند و در این جا نیز هدف این است که بین تاثیر مدل های قبلی و جدید تعادل ایجاد کند. برای رسیدن به این هدف نیز نیاز است که یک کپی از مدل های قبلی داشته باشد تا باعث فراموشی نشود. بر اساس اینکه چه نوع روشی استفاده شود، منظم سازی به دو دسته تقسیم می شود: منظم سازی وزن (که وزن هایی که اهمیت بیشتری در نتیجه دارند را بیشتر نگه می دارد و وزن های بی اهمیت به مدل جدید با ضریب خیلی کم یا صفر انتقال می یابند). و منظم سازی تابع (مدل جدید که با نام شاگرد از آن یاد می کند، تلاش می کند از اطلاعات خروجی مدل قبلی که با نام معلم از آن یاد می شود، برای ایجاد خروجی وظیفه ی خودش راهنمایی بگیرد) [۲۷، ۲۸].

۲-۲-۲ رویکرد مبتنی بر حافظه

هدف این رویکرد در واقع این است که بخشی از داده‌های قبلی را در حافظه ذخیره کرده و آن ها را در وظیفه‌ی جدید با داده‌های جدید آموزش دهد و به این صورت هم مدل جدید آموزش داده می شود و هم اطلاعات قبلی فراموش نمی شوند. این رویکرد نیز به سه دسته‌ی بازپخش تجربه [9] (انتقال بخشی از داده‌های قبلی به حافظه)، بازپخش مولد [10] (ایجاد یک مدل مولد اضافی برای بازپخش داده‌های تولید شده)، بازپخش ویژگی [11] (انتقال ویژگی های مهم داده‌های قبلی به وظیفه ی جدید و جلوگیری از فراموشی فاجعه بار). به عنوان مثال شین و همکاران [۲۹] ، مدلی به نام مدل مولد عمیق ارائه داده اند که چارچوبی است از یک معماری مدل دوگانه تعاونی، با یک مدل مولد عمیق (مولد) و یک مدل حل وظیفه (حل‌کننده) تشکیل شده است. مولد، مسئول تولید ورودی های جعلی است که شبیه داده های گذشته است، در حالی که حل کننده برای حل وظیفه فعلی آموزش دیده است. با درهم آمیختن

[8] Regularization based
[9] Experience replay
[10] Generative replay
[11] Feature replay

داده‌های آموزشی برای وظایف قبلی با داده‌های مربوط به وظیفه جدید، مدل می تواند بدون فراموش کردن دانش وظایف قدیمی، وظایف جدید را یاد بگیرد. این رویکرد از ماهیت مولد هیپوکامپ، یک سامانه حافظه کوتاه مدت در مغز پستانداران الهام گرفته شده است.

۲-۲-۳ رویکرد مبتنی بر بهینه‌سازی

رویکرد مبتنی بر بهینه‌سازی [۱۲] به جای تغییر در تابع خطا، به دستکاری برنامه های بهینه سازی می پردازد مثلا به روزرسانی پارامترها [۱۳] به گونه ای صورت گیرد که با جهت هایی مانند فضای ورودی قبلی عمود یا هم تراز باشند تا اطلاعات وظایف قبلی نیز حفظ شود.

۲-۲-۴ رویکرد مبتنی بر معماری

رویکرد مبتنی بر معماری [۱۴] به ساخت پارامترهای خاص هر وظیفه برای جلوگیری از دخالت وظیفه ها در یکدیگر می پردازد و در واقع هدف بر کاهش محدودیت استفاده از پارامترهای مشترک وظیفه‌هاست زیرا با وجود این محدودیت‌ها، اطلاعات کمتری از وظیفه‌ی قبل به بعد منتقل می شود و فراموشی فاجعه‌بار با احتمال بیشتری رخ می دهد. در این رویکرد نیز تکنیک‌های مختلفی ارائه شده است مانند تخصیص پارامتر، تجزیه مدل و شبکه پیمانه‌ای.

۲-۲-۵ رویکرد ترکیب رویکردها و سناریوها

یکی از راه حل‌ها این است که از ترکیب رویکردهای ذکر شده با یکدیگر یا با دیگر رویکردهای شبکه عصبی استفاده شود. مثلا استفاده از ترکیب رویکردهای مبتنی بر منظم سازی و مبتنی بر حافظه.

۲-۲-۶ کاربردها

یادگیری پیوسته در حوزه بینایی ماشین کاربردهای گسترده‌ای دارد که از مهم‌ترین آن‌ها می‌توان به تشخیص چهره، دسته‌بندی تصویر، تشخیص اعمال در ویدیو، بخش‌بندی معنایی و تلفیق زبان و بینایی اشاره کرد. این قابلیت‌ها امکان آموزش پیوسته مدل‌ها را بدون فراموشی اطلاعات قبلی فراهم می‌سازند و

[12] Optimization based
[13] Parameters
[14] Architecture based

باعث می‌شوند سامانه‌های هوشمند در مواجه شدن با داده‌های جدید، ضمن حفظ دانش گذشته، عملکرد خود را ارتقا دهند.

۲-۳ یادگیری پیوسته در بینایی کامپیوتر

یادگیری پیوسته در زمینه‌های مختلفی مورد استفاده قرار گرفته است. یکی از این زمینه‌ها بینایی کامپیوتر است که دو حوزه پرکاربرد به نام دسته‌بندی تصویر و تشخیص عمل را شامل شده و در ادامه به بررسی مقالات مطرح درباره ی آن ها می پردازیم.

۲-۳-۱ دسته‌بندی تصویر

مای و همکاران [۹] ، مقایسه ای بین رویکردهای مطرح ارائه شده برای دسته بندی تصویر انجام داده اند که در ادامه به معرفی مختصر رویکردها و مقایسه آن ها می پردازیم.

روش تثبیت وزن کشسان

جیمز و همکاران [۲۷] مقاله‌ای در رابطه با روش جدیدی در زمینه‌ی منظم‌سازی ارائه داده‌اند. تثبیت وزن کشسان [۱۵] یک الگوریتم است که به شبکه‌های عصبی عمیق اجازه می‌دهد تا مجموعه‌ای از وظایف پیچیده را بدون فراموش کردن فاجعه‌آمیز یاد بگیرند. این کار را با کاهش انتخابی انعطاف‌پذیری وزن انجام می‌دهد. این روش با محدود کردن هر وزن با یک جریمه درجه دوم کار می‌کند که آن را به مقداری متناسب با اهمیت آن برای عملکرد در وظایفی که قبلاً یاد گرفته شده، به سمت مقادیر قدیمی خود می‌کشاند. در واقع از ماتریسی به نام ماتریس اطلاعات فیشر برای محاسبه‌ی اهمیت وزن‌ها در مدل قبلی استفاده می‌کند.

روش یادگیری بدون فراموشی

لی و همکاران [۲۸] ، الگوریتم یادگیری بدون فراموشی [۱۶] را ارائه داده اند. این الگوریتم برای یادگیری بدون فراموشی یک الگوریتم یادگیری پیوسته است که هدف آن یادگیری وظایف جدید بدون فراموشی دانش وظایف قبلی می باشد. با آموزش یک مدل اولیه بر روی یک مجموعه وظایف اولیه کار می‌کند.

[15] Elastic weight consolidation (EWC)

[16] Learning without forgetting (LWF)

این مدل اولیه به عنوان مدل استاد نامیده می‌شود. هنگامی که با یک وظیفه جدید مواجه می‌شود، یک مدل جدید با نام مدل شاگرد، با استفاده از داده‌های وظیفه جدید، آموزش می‌بینند. مدل شاگرد همچنین با استفاده از داده‌های وظیفه‌های قدیمی، آموزش می‌بینند تا از مدل استاد پیروی کند. این به صورت محاسبه خروجی مدل استاد برای داده‌های جدید و سپس استفاده از آن خروجی به عنوان هدف برای آموزش مدل شاگرد انجام می‌شود. این روند به مدل شاگرد کمک می‌کند تا دانش وظایف قدیمی را حفظ کند، در حالی که در عین حال برای وظیفه جدید نیز بهینه می‌شود. از این روند به عنوان تقطیر دانش [17] نیز یاد می کنند. عملکرد آن می‌تواند در صورتی که وظیفه جدید بسیار متفاوت از وظایف قدیمی باشد، کاهش یابد.

روش میانگین حافظه رخدادی گرادیان

چودری و همکاران [۳۰] ، الگوریتمی به نام الگوریتم میانگین حافظه رخدادی گرادیان [18] ارائه کردند که با استفاده از یک حافظه ذخیره شده برای ذخیره اطلاعات مربوط به وظایف قبلی، از فراموشی فاجعه بار جلوگیری می کند. در هر مرحله از یادگیری، مدل از حافظه ذخیره شده برای تولید یک زیرمجموعه تصادفی از تجارب استفاده می کند. این تجربیات سپس برای محاسبه گام های گرادیان استفاده می شوند. گام های گرادیان فعلی سپس با گام های گرادیان زیرمجموعه تصادفی از حافظه ذخیره شده میانگین گیری می شوند. این میانگین گیری به مدل کمک می کند تا یک دیدگاه کلی تر از وظیفه فعلی به دست آورد. این دیدگاه کلی تر به مدل کمک می کند تا دانش وظایف قبلی خود را حفظ کند، حتی زمانی که در حال یادگیری یک وظیفه جدید است.

روش طبقه بندی افزایشی و یادگیری بازنمایی

ربوفی و همکاران [۳۱] ، روش طبقه بندی افزایشی و یادگیری بازنمایی [19] در یادگیری پیوسته ارائه دادند که مدل با استفاده از نمونه هایی از همه دسته ها، از جمله دسته های جدید و قدیمی، آموزش داده می شود. تابع ضرر شامل یک ضرر طبقه بندی برای تشویق مدل به پیش بینی برچسب های صحیح برای دسته های جدید و یک ضرر تقطیر دانش برای تشویق مدل به بازتولید خروجی های مدل قبلی برای دسته های قدیمی است. همچنین یک روش به روزرسانی حافظه را پیشنهاد می دهد که بر اساس فاصله در فضای ویژگی های نهفته است. این روش برای انتخاب زیرمجموعه ای از نمونه ها از هر دسته

[17] Knowledge distillation (KD)

[18] Averaged Gradient Episodic Memory (A-GEM)

[19] Incremental Classifier and Representation Learning (iCaRL)

استفاده می شود که میانگین ویژگی های نهفته آنها به میانگین همه نمونه ها در این دسته نزدیک ترین است.

روش حداکثر تداخل ارزیابی

الجاندی و همکاران [۳۲] ، روشی به نام حداکثر تداخل ارزیابی [۲۰] را ارائه داده اند که یک روش مبتنی بر بازپخش است که اخیراً با هدف بهبود راهبرد بازیابی حافظه پیشنهاد شده است. این روش، نمونه‌های بازپخش را با توجه به افزایش ضرر با داشتن به به‌روزرسانی پارامتر تخمین زده شده بر اساس دسته های کوچک ورودی انتخاب می‌کند. نمونه‌های حافظه را که بیشترین تداخل (افزایش ضرر) را با به‌روزرسانی پارامتر با دسته ورودی جدید دارند، انتخاب می‌کند. همچنین نمونه‌گیری مخزن را در به‌روزرسانی حافظه اعمال می‌کند و نمونه‌های حافظه انتخابی را با نمونه‌های جدید در به‌روزرسانی مدل دوباره پخش می‌کند.

۲-۳-۲ تشخیص عمل

تشخیص عمل یکی از مباحث پیشرفته امروزی است که به علت اضافه شدن پارامتر زمان، پیچیدگی‌های بیشتری نسبت به تصویر پیدا کرده است. همچنین به علت حجیم بودن داده‌ها در این مسائل، یادگیری مدوام ضرورت پیدا می‌کند. زیرا در طول زمان مثلاً دسته‌ها و داده‌های بیشتری به مدل اضافه می‌شوند و مدل نمی‌تواند دوباره از ابتدا این حجم داده را آموزش دهد. پس چند سال اخیر، مطالعه‌هایی نیز در این زمینه شده و روش‌های جدید با مجموعه داده‌های مختلف ارائه شده است. که به چندین مقاله در ادامه می‌پردازیم.

روش مین هاس

مین هاس و همکاران [۳۳] ، یک روش یادگیری پیوسته برای تشخیص اعمال انسان ارائه داده‌اند. این روش بر پایه یک چارچوب است که دو تکنیک، یعنی تقریب شکل و یادگیری تحلیلی، را با هم ترکیب می‌کند. تقریب شکل برای ثبت شکل بازیگر در ویدیو استفاده می‌شود. به این صورت که با تغییراتی که به شکل می‌دهیم، از تصاویر شدت نور بهره‌برداری می‌کنیم تا ویژگی‌های مربوط به جهت گرادیان‌ها را استخراج کنیم. هنگامی که حرکت در ویدیو پیش می‌رود، شکل با تغییر و تنظیم چندین قطعه کوچک داخل یک پنجره‌ی پیگیری، به‌طور دقیق به تغییرات مرزها پیگیری می‌کند. به منظور یادگیری دینامیک‌های غیرخطی حرکت‌ها، از یادگیری تحلیلی استفاده می‌شود. این فرآیند یادگیری به شکل

[20] Maximally Interfered Retrieval (MIR)

بازگشتی انجام می‌شود و از طریق آن آموزش به نمایش خطی ساده تبدیل می‌شود. این روش دو مزیت دارد: کمینه کردن خطاها و کاهش قابل توجه زمان محاسباتی، و از بین بردن محدودیت‌های آموزش به صورت دسته‌ای برای تشخیص اعمال. این روش یادگیری پیوسته اجازه می‌دهد که مدل به تدریج با ورودی داده‌های جدید به‌روز شود. این روش مقابل یادگیری دسته‌ای است که در آن برای آموزش دسته‌بند تمام مجموعه داده آموزشی استفاده می‌شود.

روش الزنت

لی و همکاران [۳۴] روشی به نام الزنت را ارائه کردند که با انتخاب و به‌روزرسانی پویاترین بلوک‌های یادگیری از فراموشی فاجعه‌بار در شناسایی عمل جلوگیری می‌کند. هنگام یادگیری اعمال جدید، الزنت به دنبال بلوک‌های یادگیری می‌گردد که بیشترین ارتباط را با عمل فعلی دارند و پارامترهای آن‌ها را به‌روزرسانی می‌کند، در حالی که پارامترهای بلوک‌های غیرانتخابی را حفظ می‌کند. این راهبرد به‌روزرسانی انتخابی به حفظ دانش حرکات قبلاً یادگیری شده کمک می‌کند و مشکل فراموشی را کاهش می‌دهد. با به‌روزرسانی فقط بلوک‌های مرتبط، از وارد کردن نویز و اختلالات غیرمرتبط به دانش قبلی جلوگیری می‌کند، که منجر به عملکرد بهتر در یادگیری اعمال جدید می‌شود.

روش تعبیه همسایه تی تصادفی موقت تحت نظارت

چنگ و همکاران [۳۵] یک روش برای تشخیص حرکات انسان با استفاده از تعبیه‌سازی همسایگی تصادفی زمانی نظارت شده و یادگیری پیوسته ارائه کرده اند. الگوریتم برای یادگیری ارتباط بین قاب‌های عمل به‌کار می‌رود و اطلاعات دسته و زمانی را تلفیق می‌کند. یادگیری پیوسته برای تعبیه‌سازی کم‌بعدی داده‌های جدید با استفاده از رویکردهایی نظیر تعبیه خطی محلی و پیش بینی حفظ محلی استفاده می‌شود. همچنین سه روش برای یادگیری پیوسته در زمینه تشخیص حرکات انسان توصیف می‌کند.

روش پاریزی

پاریزی و همکاران [۳۶] رویکردی را ارائه کرده‌اند که باعث جلوگیری از فراموشی دانش با شبکه‌ی سلسله مراتبی خودسازمانده می‌شود. در این شبکه‌ی سلسله مراتبی هر لایه به صورت شبکه رشد هنگام نیاز است به این صورت که نورون‌های جدید را تخصیص می‌دهد یا نورون‌های موجود را بر اساس اختلاف بین توزیع ورودی و وزن‌های نورون‌های نمونه‌ای به‌روز می‌کند.

۲-۴ مدل‌های بینایی-زبان

مدل‌های زبانی بزرگ، شبکه‌های عصبی ترنسفورمر مقیاس‌پذیری هستند که با آموزش بر حجم عظیمی از داده‌های متنی، قادرند زبان طبیعی را تولید، درک و تحلیل کنند. این مدل‌ها به‌دلیل ظرفیت بالای خود در یادگیری، نقش اساسی در پیشرفت‌های اخیر پردازش زبان طبیعی ایفا کرده‌اند. به دنبال این پیشرفت‌ها پژوهش‌های زیادی انجام شده است که از این مدل‌ها در کاربردهای بینایی ماشین نیز استفاده کرده‌اند [۱۴]. مدل‌های بینایی-زبان دسته‌ای از مدل‌های هوش مصنوعی هستند که به‌طور هم‌زمان قادر به تحلیل و درک داده‌های بصری (تصویر یا ویدیو) و زبانی (متن) می‌باشند. این مدل‌ها با استفاده از حجم انبوهی از داده‌های تصویر-متن که به‌صورت گسترده در وب موجود است، آموزش می‌بینند. ایده اصلی پشت این مدل‌ها، یادگیری هم‌بستگی میان نمایش‌های تصویری و متنی در یک فضای مشترک نهفته [۲۱] است. به‌عنوان نمونه، مدل CLIP [۲۲] [۱۳] که توسط OpenAI ارائه شده است، با بهره‌گیری از صدها میلیون جفت تصویر و متن، توانسته است عملکرد قابل قبولی در وظایف مختلف بینایی و زبانی ارائه دهد. مدل‌های بینایی-زبان به دلایل متعددی مورد توجه پژوهشگران قرار گرفته‌اند که به برخی از آن‌ها در ادامه اشاره میکنیم [۱۴]:

- توانایی پیش‌بینی در حالت یادگیری بدون نمونه: این مدل‌ها قادرند وظایف جدید را بدون نیاز به بازآموزی [۲۳]، انجام دهند. به این حالت، یادگیری بدون نمونه [۲۴] گفته می‌شود.
- چندکاربردی بودن: یک مدل واحد می‌تواند در وظایف متنوعی همچون دسته‌بندی تصویر، تشخیص اشیا، بازیابی تصویر بر اساس متن و تولید توضیح برای تصویر به‌کار گرفته شود.
- قابلیت مقیاس‌پذیری بالا: امکان آموزش بر روی میلیاردها جفت تصویر-متن و دستیابی به تعمیم‌پذیری قابل توجه در دامنه‌های گوناگون را دارد.

آموزش این مدل‌ها هزینه‌ی محاسباتی بالایی دارد اما طریقه‌ی استفاده از آن‌ها به صورتی است که این چالش تعدیل شود. استفاده از این مدل‌ها به سه مرحله‌ی اصلی تقسیم می‌شود: پیش آموزش، یادگیری انتقالی و تقطیر دانش که در ادامه بررسی می‌گردد.

[21] Embedding space
[22] Contrastive Language-Image Pre-training (CLIP)
[23] Retraining
[24] Zero-shot learning

۲-۴-۱ پیش آموزش مدل‌های بینایی-زبان

در مرحله‌ی پیش آموزش مدل‌های بینایی-زبان [۲۵] مدل با بهره‌گیری از حجم انبوهی از داده‌های تصویر-متن بدون برچسب، به‌گونه‌ای آموزش می‌بینند که توانایی درک هم‌زمان مفاهیم زبانی و تصویری را کسب کند. سه نوع هدف آموزشی عمده در این بخش عبارت‌اند از:

اهداف تقابلی

در روش اهداف تقابلی [۲۶] مدل یاد می‌گیرد تا نمایش جفت‌های صحیح تصویر-متن را به یکدیگر نزدیک و جفت‌های نادرست را از هم دور کند. به عنوان مثال مدل CLIP که با هدف تقابلی و داده‌های وب‌مقیاس آموزش داده شد و توانست در بیش از ۳۰ وظیفه نمونه-صفر عملکرد موفقی ارائه دهد.

اهداف مولد

در اهداف مولد [۲۷] مدل به بازسازی بخش‌های حذف‌شده از تصویر یا متن می‌پردازد یا توصیف متنی برای تصویر تولید می‌کند. به عنوان نمونه، مدل FLAVA [۳۷] با بهره‌گیری هم‌زمان از ماسک‌کردن تصویر و زبان، دانش چندحالته‌ای را در یک مدل واحد می‌آموزد.

اهداف هم‌ترازی

اهداف هم‌ترازی [۲۸] بر هم‌خوانی معنایی میان تصویر و متن، به صورت کلی (تطابق تصویر-متن [۲۹]) یا حتی به صورت محلی (تطابق واژه-ناحیه [۳۰]) تمرکز دارند. مدل GLIP [۳۸] با هم‌ترازی زبان-ناحیه توانست به شناسایی اشیای واژگان-باز [۳۱] دست یابد.

۲-۴-۲ یادگیری انتقالی مدل‌های بینایی-زبان

برای استفاده از مدل‌های بینایی-زبان در وظایف خاص مانند دسته‌بندی تصویر، تشخیص اشیا یا بازیابی تصویر، لازم است که مدل با روش‌هایی کم‌هزینه و تطبیقی انتقال یابد. مهم‌ترین روش‌ها عبارت‌اند از:

[25] Vision-language model pre-training
[26] Contrastive objectives
[27] Generative objectives
[28] Alignment objectives
[29] image-text matching
[30] region-word matching
[31] Open-vocabulary

تنظیم پرامپت

در روش تنظیم پرامپت، به‌جای تغییر ساختار داخلی مدل یا بازآموزی کامل آن، تلاش می‌شود تا ورودی‌های متنی (و در برخی موارد تصویری) به‌گونه‌ای هوشمندانه طراحی یا بهینه شوند که مدل بتواند عملکرد بهتری در وظیفهٔ موردنظر ارائه دهد. در واقع، مدل اصلی ثابت می‌ماند و تنها شکل ورودی‌هایی که به آن داده می‌شود، به کمک الگوریتم‌هایی قابل یادگیری، تغییر می‌کند. این رویکرد به‌ویژه برای وظایف با یادگیری محدود بسیار کارآمد است؛ زیرا به مدل اجازه می‌دهد با استفاده از اطلاعات آموخته‌شده قبلی، خود را با وظیفه‌ی جدید تطبیق دهد بدون آنکه پارامترهایش خیلی تغییر کند. لیو و همکاران [۳۹]، تشخیص حرکت را به مسئله‌ی تطبیق ویدیو-متن تبدیل کرده‌اند تا از قدرت نمایش‌های زبانی بهره ببرند. پرامپت‌سازی نقش کلیدی در نزدیک‌سازی وظیفه‌ی هدف به ساختار داده‌های پیش‌تمرین‌شده دارد و موجب بهبود عملکرد مدل در شرایط یادگیری بدون نمونه می‌شود. در مدل CoOp [۱۷] به‌جای استفاده از پرامپت‌های متنی دستی مانند "a photo of a [class name]" پرامپت‌هایی از کلمات قابل یادگیری طراحی شدند. این کلمات به‌صورت بردارهایی آموزش‌پذیر به مدل داده می‌شوند و نقش آنها تقویت معنای دسته‌بندی برای مدل است. این روش، که در بخش بعدی بیشتر به آن پرداخته می‌شود، باعث شد صحت مدل CLIP در دسته‌بندی چنددسته به‌ویژه در شرایط یادگیری محدود به‌طور قابل توجهی بهبود یابد.

وفق دهنده‌ی ویژگی

وفق‌دهنده‌ی ویژگی یکی از روش‌های مؤثر برای انتقال مدل‌های بینایی-زبان به وظایف جدید بدون نیاز به بازآموزی کامل شبکه است. در این رویکرد، به‌جای تغییر پارامترهای اصلی مدل، لایه‌هایی سبک و کم‌پارامتر به انتهای یا میانه‌ی شبکه اضافه می‌شود تا ویژگی‌های استخراج‌شده از تصویر یا متن را با نیازهای وظیفه‌ی خاص منطبق کند. این تطبیق‌دهنده‌ها می‌توانند به‌صورت افزونه‌هایی جدا از معماری اصلی عمل کنند، بنابراین هسته‌ی مدل بدون تغییر باقی می‌ماند. در روش CLIP-Adapter [۴۰] مجموعه‌ای از لایه‌های سبک‌وزن به مدل CLIP افزوده شد تا ویژگی‌های استخراج‌شده از تصویر و متن پیش از تصمیم‌گیری نهایی پردازش و تطبیق یابند. این کار برای وظایف یادگیری کم‌نمونه نیز مؤثر بود، زیرا بدون نیاز به تغییر در مدل پایه، عملکرد بسیار مناسبی حاصل شد. این روش انتقالی به دلیل کم‌هزینه بودن و عدم نیاز به تنظیم مجدد کل مدل، برای بسیاری از کاربردهای عملی مناسب است.

سایر روش‌ها

در کنار تنظیم پرامپت و تطبیق‌دهنده‌های ویژگی، برخی روش‌ها نیز با تغییر مستقیم در پارامترهای مدل، در بهبود عملکرد مدل برای وظایف خاص، نقش دارند. این روش‌ها معمولاً شامل تنظیم دقیق کامل یا تلفیق مدل‌های یادگرفته‌شده با مدل اولیه هستند. در روش Wise-FT [۴۱] ، یک رویکرد ساده اما مؤثر ارائه شده است که در آن وزن‌های مدل پایه و مدل تنظیم دقیق شده به‌صورت میانگین‌گیری وزنی ترکیب می‌شوند. این تکنیک باعث می‌شود که مدل هم از تعمیم‌پذیری مدل اولیه بهره ببرد و هم بتواند دانش خاص وظیفه‌ی جدید را بیاموزد، بدون آنکه دچار بیش‌برازش شود. در توسعه‌ی این روش، ونگ و همکاران [۲۰]، روش Open-VCLIP را برای تطبیق مدل CLIP برای داده‌های ویدیویی ارائه دادند به صورتی که دانش مدل CLIP نیز حفظ شود. این مدل در فصل بعد به صورت مفصل‌تری توضیح داده خواهد شد.

۲-۴-۳ تقطیر دانش

در تقطیر دانش، دانش مدل بینایی-زبان به یک مدل سبک‌تر منتقل می‌شود تا بتوان از آن در کاربردهای خاص و با منابع محدود استفاده کرد. دو کاربرد اصلی عبارت‌اند از:

تقطیر دانش برای تشخیص شیء

در حوزه‌ی بینایی کامپیوتر، یکی از چالش‌های مهم، شناسایی اشیائی است که در داده‌های آموزش مدل پایه وجود نداشته‌اند. روش‌های متداول تشخیص شیء نیازمند برچسب‌گذاری دقیق و پرهزینه‌ی داده‌ها برای هر دسته هستند. در این میان، مدل‌های بینایی-زبان مانند CLIP که از داده‌های وب‌مقیاس و متن‌های توصیفی متنوع آموزش دیده‌اند، دارای دانش گسترده‌ای درباره‌ی مفاهیم بصری و زبانی هستند که می‌توان از آن‌ها برای توسعه‌ی مدل‌های تشخیص شیء استفاده کرد. در مدل VILD [۴۲] نمونه‌ای برجسته از این رویکرد است. این مدل با استفاده از تقطیر دانش از CLIP یک آشکارساز دو مرحله‌ای توسعه داده است که می‌تواند اشیاء خارج از مجموعه‌ی برچسب‌گذاری‌شده‌ی اولیه را شناسایی کند؛ به این صورت که ویژگی‌های بصری استخراج‌شده از تصاویر با تعبیه های متنی مدل CLIP مقایسه می‌شوند تا به جای اتکا به دسته‌های از پیش تعریف‌شده، اشیاء جدید نیز قابل شناسایی باشند. این روش نوعی تشخیص شیء واژگان-باز را ممکن می‌سازد که در بسیاری از کاربردهای دنیای واقعی اهمیت بالایی دارد.

تقطیر دانش برای بخش‌بندی معنایی

بخش‌بندی معنایی به معنای اختصاص یک برچسب معنایی به هر پیکسل از تصویر است و از وظایف کلیدی در درک صحنه محسوب می‌شود. پیاده‌سازی موفق این وظیفه معمولاً نیازمند مجموعه داده‌های پرحجم و برچسب‌خورده در سطح پیکسل است، که تولید آن‌ها بسیار پرهزینه و زمان‌بر است. با این حال، مدل‌های بینایی-زبان که از داده‌های ضعیف‌برچسب‌خورده یا بدون برچسب بهره می‌برند، می‌توانند دانش انتزاعی خود را به مدل‌های سبک‌تر انتقال دهند تا نیاز به برچسب‌گذاری کاهش یابد. در CLIPSeg [۴۳] از ویژگی‌های استخراج‌شده توسط CLIP برای هر تصویر استفاده می‌کند و با افزودن یک رمزگشای سبک [۳۲] امکان پیش‌بینی نقشه‌های بخش‌بندی معنایی را تنها بر اساس توصیف متنی (prompt) فراهم می‌کند؛ برای مثال، با دادن جمله‌ای مانند «گربه در تصویر کجاست؟»، مدل قادر به تولید نقشه‌ای است که نواحی مربوط به گربه را برجسته کند. نکته قابل توجه این است که که این مدل به یادگیری بدون نمونه دست یافته و برای انجام این کار نیازی به آموزش مجدد بر روی داده‌های هدف ندارد، که آن را برای کاربردهای در دنیای واقعی بسیار کارآمد و مقیاس‌پذیر می‌سازد.

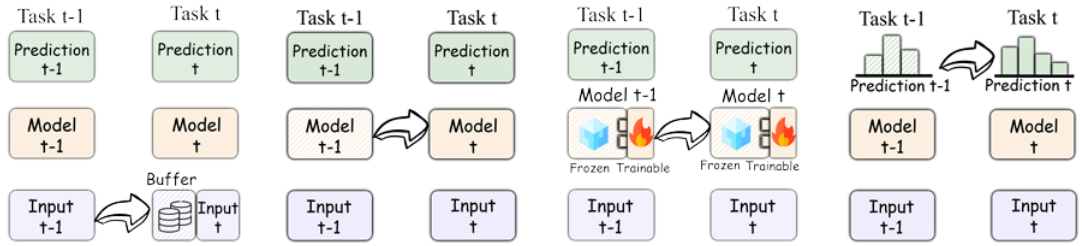
۲-۵ یادگیری پیوسته در مدل‌های بینایی-زبان

بیشتر مدل‌های زبانی بزرگ و بینایی-زبان، در شرایط ایستا آموزش می‌بینند و توانایی کمی در انطباق با داده‌های جدید، بدون بازآموزی کامل، دارند. این محدودیت، باعث توسعه و پیاده‌سازی روش‌های متنوع یادگیری پیوسته در این مدل‌ها شده است. در این زمینه، ژنگ و همکاران [۱۵] طبقه‌بندی شکل ۲-۱ را برای روش‌های متفاوت یادگیری پیوسته ارائه داده‌اند که در ادامه به آن پرداخته می‌شود.

۲-۵-۱ روش‌های مبتنی بر حافظه

در این روش، مدل بخشی از داده‌های قدیمی را ذخیره کرده و در کنار داده‌های جدید برای آموزش مجدد استفاده می‌کند. این رویکرد با هدف کاهش پدیده‌ی فراموشی طراحی شده است، که در آن مدل، دانش قبلی خود را هنگام یادگیری اطلاعات جدید از دست می‌دهد. محدودیت در ذخیره‌سازی داده‌های قدیمی چالش اصلی این روش است [۱۵] . گارگ و همکاران [۴۴]، روشی برای آموزش پیوسته‌ی مدل‌های بینایی-زبان مانند CLIP، در مواجه شدن با داده‌های وب‌مقیاس و در حال تغییر زمانی، ارائه کرده‌اند.

[32] Lightweight decoder



شکل ۲-۱: طبقه‌بندی روش‌های یادگیری پیوسته در مدل‌های زبانی بزرگ و بینایی-زبان [۱۵]. روش‌ها از سمت چپ به ترتیب، مبتنی بر حافظه، مبتنی بر تنظیم، مبتنی بر معماری و مبتنی بر تقطیر دانش می‌باشند.

این روش با بهره‌گیری از بازپخش داده‌های گذشته و استفاده از مدل پیش‌آموخته به‌عنوان نقطه شروع، امکان به‌روزرسانی کارامد مدل را، بدون نیاز به بازآموزی کامل، فراهم می‌کند.

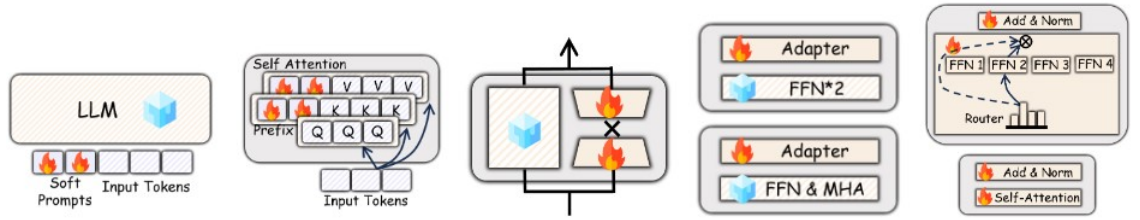
۲-۵-۲ روش‌های مبتنی بر تنظیم

در این روش، مدل از مکانیزم‌هایی مانند جریمه یا محدودسازی برای حفظ اطلاعات قبلی استفاده می‌کند. هدف این است که پارامترهایی که در یادگیری گذشته مهم بوده‌اند، هنگام آموزش جدید کمتر تغییر کنند. در این روش، در صورت حجم زیاد وظایف، اثربخشی کاهش می‌یابد [۱۵] .

۲-۵-۳ روش‌های مبتنی بر تقطیر دانش

مدل دانش خود را از مدل‌های قبلی یا معلم یاد می‌گیرد و توجه به مواردی مانند حساسیت به صحت مدل معلم و انتخاب صحیح داده‌ها برای تقطیر حائز اهمیت است [۱۵]. لو و همکاران [۱۹]، با هدف کاهش فراموشی در یادگیری افزایشی ویدیو، از روشی به نام تقطیر توجه [۳۳] استفاده کرده‌اند که در آن ویژگی‌های توجه از خروجی کدگشای ترنسفورمر CLIP، به مدل جدید منتقل می‌شود. این رویکرد به مدل کمک می‌کند تا دانش مراحل قبلی را حفظ کرده و در عین حال بتواند دسته‌های جدید را بدون نیاز به آموزش کامل مجدد یاد بگیرد.

[33] Attention distillation



(a) Prompt Tuning (b) Prefix Tuning (c) LoRA (d) Adapters (e) Mixture of Experts

شکل ۲-۲: روش‌های یادگیری پیوسته مبتنی بر معماری در مدل‌های زبانی بزرگ و بینایی-زبان [۱۵]

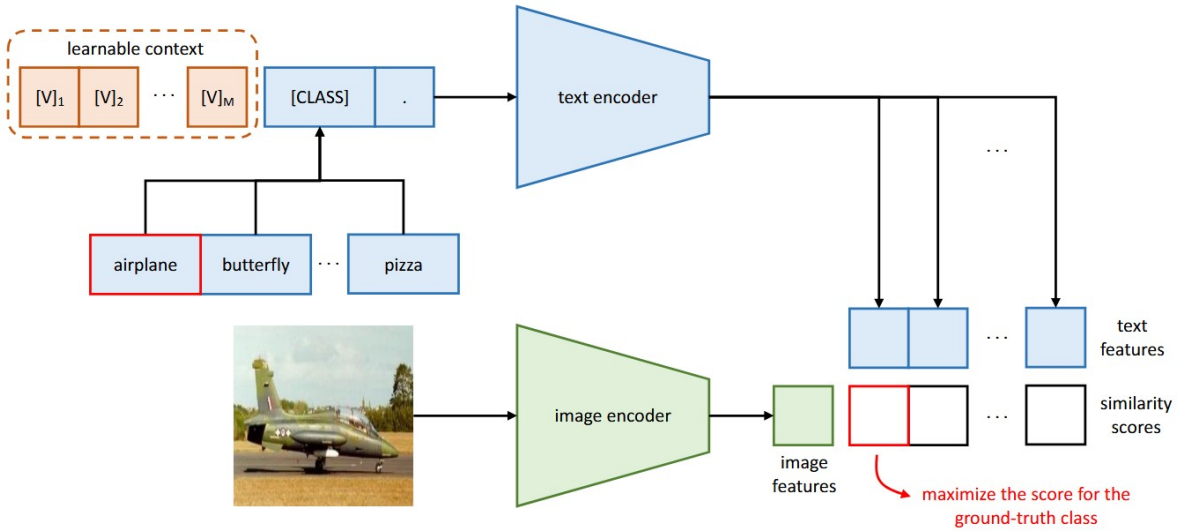
۲-۵-۴ روش‌های مبتنی بر معماری

در این رویکرد، معماری مدل برای جذب وظایف جدید، بدون تداخل با وظایف قبلی، تغییر می‌کند؛ مانند افزودن پیمانه [۳۴] استفاده از تطبیق‌دهنده‌ها و .... ژنگ و همکاران [۱۵]، با بررسی روش‌های مختلف ارائه‌شده در این رویکرد، طبقه‌بندی مطابق با شکل ۲-۲ ارائه داده اند که در ادامه به بررسی آن‌ها پرداخته می‌شود .

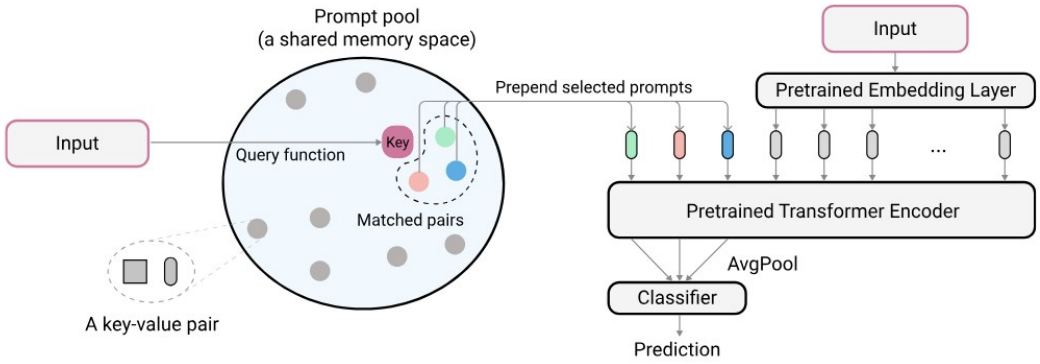
تنظیم پرامپت

همانطور که در بخش یادگیری انتقالی مدل‌های بینایی-زبان ذکر شد، در روش تنظیم پرامپت، به‌جای بازآموزی کامل یا تغییر در پارامترهای اصلی مدل، مجموعه‌ای از بردارهای قابل آموزش به‌عنوان پرامپت به ابتدای نشانه‌های ورودی [۳۵] اضافه می‌شوند. این بردارها بدون دست‌کاری در ساختار درونی مدل، نقش راهنما را ایفا کرده و جهت‌گیری مدل در تفسیر داده‌های جدید را مشخص می‌سازند. در واقع، مدل با همان دانش قبلی خود به تحلیل ورودی می‌پردازد، اما به‌واسطه‌ی پرامپت‌های جدید، قادر به تطبیق با وظایف تازه می‌شود [۱۵]. این روش به‌دلیل مصرف کم منابع محاسباتی و عدم نیاز به تغییر در پارامترهای اصلی، به‌ویژه برای سناریوهایی با دسترسی محدود به مدل یا منابع، بسیار مناسب است و الگوی مورد استفاده در این روش، می‌تواند به خوبی برای مسائل یادگیری پیوسته نیز استفاده شود [۱۵].
ژو و همکاران [۱۷]، رویکردی به نام بهینه‌سازی بافت [۳۶] (که در بخش قبل اشاره شد) ارائه کرده‌اند که به جای پرامپت‌های ثابت، از پرامپت به صورت بردارهای بافت یادگرفتنی [۳۷] که در کنار برچسب متنی داده‌ها قرار می‌گیرند، استفاده می‌کند. مطابق شکل ۲-۳، تمام وزن‌های مدل CLIP، ثابت نگه داشته شده

[34] Module
[35] Input tokens
[36] Context Optimization (CoOp)
[37] Learnable context vectors



شکل ۲-۳: طرح کلی مدل CoOp [۱۷]. در سمت چپ بالای تصویر، پرامپت‌های قابل آموزش نشان داده شده اند که به برچسب دسته‌ها چسبیده و وارد کدگذار متن می‌شوند. ویژگی کدگذار متن با ویژگی استخراج‌شده از کدگذار تصویر مقایسه شده و در روند آموزش، سعی می‌شود ویژگی تصویر و ویژگی برچسب واقعی تصویر، به یکدیگر نزدیک شوند.



شکل ۲-۴: طرح کلی مدل L2P در زمان آزمون [۷]. در بخش چپ تصویر، استخر پرامپت شامل مجموعه‌ای از جفت‌های کلید-پرامپت نمایش داده شده است. ورودی پس از عبور از تابع جستجو، با کلیدهای موجود در استخر مقایسه می‌شود. سپس تعدادی از کلیدهای مشابه‌تر انتخاب و پرامپت‌های متناظر آن‌ها به ورودی (که از لایه‌ی تعبیه عبور کرده است) اضافه می‌گردد. در بخش راست تصویر، ورودی غنی‌شده با پرامپت‌ها وارد کدگذار ترنسفورمر شده و خروجی حاصل، پس از مرحله‌ی میانگین‌گیری، به دسته‌بند منتقل می‌شود تا پیش‌بینی نهایی انجام گیرد.

و تنها بردارهای پرامپت، قابل‌آموزش هستند. به دنبال توسعه‌ی روش‌های پرامپت گذاری، روش L2P [۳۸] که توسط وانگ و همکاران [۷] ارائه شده است، یک چارچوب نوآورانه برای یادگیری پیوسته، بدون نیاز به شناسایی وظیفه در زمان آزمون، می‌باشد. همان‌طور که در شکل ۲-۴ مشاهده می‌شود، این روش به جای تغییر وزن‌های مدل پیش‌آموخته، از مجموعه‌ای از پرامپت‌های یادگرفتنی بهره می‌برد که در یک فضای حافظه اشتراکی به نام استخر پرامپت نگهداری می‌شوند. L2P از یک مکانیزم پرس‌وجوی مبتنی بر جفت‌های کلید-مقدار بهره می‌برد تا به‌صورت پویا و متناسب با ورودی، پرامپت‌های مرتبط را انتخاب کرده و به نشانه‌های ورودی مدل، اضافه کند. سپس این نشانه‌های توسعه‌یافته به مدل پیش‌آموخته تزریق شده و پیش‌بینی انجام می‌شود. در این روش، پرامپت‌ها، دانش خاص هر وظیفه یا دانش مشترک بین وظایف را به‌صورت فشرده ذخیره می‌کنند و باعث کاهش چشمگیر فراموشی مخرب در یادگیری وظایف متوالی می‌شوند. ساختار طراحی‌شده در L2P همان‌طور که در شکل ۲-۴ نشان داده شده، از یک بخش انتخاب پرامپت، لایه‌های کدگذار پیش‌آموخته [۳۹]، و دسته‌بند نهایی تشکیل شده است. جهت بهبود و مقاوم‌سازی نسبت به فراموشی در انتخاب پرامپت روش‌های پیشین، مارتین و همکاران [۴۵]، روش STAR-Prompt را معرفی کرده‌اند که از رویکردی دوسطحی برای تنظیم پرامپت، پیروی می‌کند. ابتدا از CLIP، برای تولید پرامپت‌های متنی و ساخت نمونه‌های اولیه [۴۰] پایدار دسته‌ها، استفاده می‌شود و سپس این نمونه‌های اولیه به‌عنوان کلید برای بازیابی پرامپت‌های تصویری در ترنسفورمر تصویر به‌کار می‌روند. هم‌چنین روش‌های DualPrompt [۴۶] و H-prompts [۴۷]، مانند مطالعات مذکور، در زمینه‌ی تولید پرامپت‌های مشترک و خاص وظایف ارائه شده‌اند. داهویین و همکاران [۴۸]، از پرامپت اختصاصی برای هر نمونه به‌جای هردسته، استفاده کرده‌اند. برخلاف روش‌های پرامپت‌گذاری قبلی، هنگ و همکاران [۴۹]، علاوه بر استفاده از یک پرامپت به‌جای چندپرامپت، از نمونه‌های پرت مصنوعی برای ایجاد مرز دسته‌بندی بهتر استفاده کرده‌اند. در مطالعات دیگری مانند روش یکپارچه‌سازی دانش بدون تداخل و آگاه از توزیع [۵۰] [۴۱]، به رفع چالش مداخله‌ی پرامپت در تصمیم‌گیری مکانیزم توجه پرداخته شده است. به دنبال پیشرفت‌های روش‌های پرامپت در حوزه‌ی تصویر، ویلا و همکاران [۵۱]، روشی به نام PIVOT را معرفی کرده‌اند که با بهره‌گیری از دانش پیش‌آموخته مدل تصویر-متن CLIP و استفاده از پرامپت‌های مکانی [۴۲] و زمانی، وابستگی‌های زمانی و مکانی ویدیوها را مدل کرده است. وانگ و همکاران [۱۶] نیز با

[38] Learning to Prompt for Continual Learning

[39] Pretrained encoder layers

[40] Prototypes

[41] Distribution-aware Interference-free Knowledge Integration (DIKI)

[42] Spatial prompts

هدف ارتقای عملکرد مدل CLIP در تشخیص حرکت‌های انسانی در ویدیوها، چارچوبی معرفی کرده‌اند که با استفاده از مدل‌سازی حرکتی و پرامپت‌های پویا، به شکلی مؤثر اطلاعات حرکتی را وارد فرآیند یادگیری می‌کند بدون اینکه به تغییر پارامترهای اصلی CLIP نیاز باشد. در مواردی نیز مانند روش ViLT-CLIP [۵۲]، از هر دو نوع پرامپت برای تصویر و متن برای درک ویژگی‌های ویدیویی استفاده شده‌است.

تنظیم پیشوند

در روش تنظیم پیشوند، مجموعه‌ای از پارامترهای قابل آموزش به عنوان پیشوند به ابتدای هر لایه‌ی ترنسفورمر افزوده می‌شود تا رفتار مدل را در انجام وظایف خاص تنظیم کند. این پیشوندها نقش تغییرات زمینه‌ای را ایفا کرده و برخلاف تنظیم پرامپت، چندین لایه از مدل را تحت تأثیر قرار می‌دهند [۱۵]. روی و همکاران [۵۳]، روشی معرفی کرده‌اند که با استفاده از پیشوندهای قابل یادگیری در هر لایه‌ی مدل، امکان یادگیری وظایف جدید را بدون فراموشی وظایف قبلی فراهم می‌کند. این پیشوندها با ترکیب کانولوشن و اطلاعات مشترک بین وظایف، باعث انتقال بهتر دانش و کاهش تعداد پارامترهای لازم در یادگیری پیوسته می‌شوند.

سازگارسازی رتبه‌پایین

با وارد کردن ماتریس‌های رتبه‌پایین در لایه‌های معینی از مدل پیش‌آموخته و منجمد، روش سازگارسازی رتبه‌پایین، امکان تنظیم هدفمند بخش‌هایی از مدل را بدون بازآموزی کامل فراهم می‌سازد [۱۵]. مارتین و همکاران [۵۴]، به نتیجه رسیدند که در زمینه‌ی یادگیری پیوسته، جایگزینی روش مذکور به‌جای تنظیم پرامپت، بدون افزایش چشمگیر در تعداد پارامترها، منجر به بهبود عملکرد مدل می‌گردد.

وفق دهنده

همانطور که در بخش یادگیری انتقالی مدل‌های بینایی-زبان ذکر شد، وفق دهنده‌ها شبکه‌های عصبی کوچک با ساختار فشرده‌ای هستند که بین لایه‌های مدل اصلی قرار می‌گیرند و به مدل امکان می‌دهند ویژگی‌های جدید را یاد بگیرد، بدون آن‌که نیازی به تغییر پارامترهای اصلیِ از پیش آموزش‌دیده باشد [۱۵]. به همین دلیل، می‌توان این روش را نیز در یادگیری پیوسته استفاده نمود. در برخی روش‌ها مانند DIA [۵۵] و EASE [۵۶]، از وفق دهنده‌های سبک‌وزن و اختصاصی برای هر وظیفه‌ی جدید استفاده می‌شود تا مدل بتواند بدون بازآموزی کامل یا ذخیره داده‌های قدیمی، دانش جدید را جذب کند. این

دو روش، امکان به‌روزرسانی مدل را بدون آسیب به دانش قبلی فراهم کرده و تصمیم‌گیری ترکیبی میان دسته‌های قدیم و جدید را ممکن می‌سازند. دانگ و همکاران [۵۷]، در مطالعه‌ای دیگر، روشی به نام C-ADA، ارائه داده‌اند که با استفاده از وفق دهنده‌هایی با قابلیت گسترش عاملی، یادگیری وظایف جدید را بدون نیاز به ذخیره داده‌های گذشته ممکن می‌سازد. این روش با حفظ پارامترهای قبلی و افزودن وزن‌های جدید، از تداخل دانش جلوگیری کرده و عملکرد و سرعت آموزش را به‌طور محسوسی بهبود می‌دهد. برای رفع چالش تداخل پارامترهای دسته‌های مشابه در یادگیری پیوسته، هانگ و همکاران [۵۸]، با استفاده از وفق دهنده‌های قابل تنظیم، ابتدا بازنمایی‌های متنی را متناسب با تأثیر دسته‌های جدید بر دسته‌های قدیمی اصلاح می‌کنند و سپس با یک راه‌برد تجزیه و ادغام پارامترها، فراموشی مدل در حین تنظیم وفق دهنده‌ها را کاهش می‌دهند. در حوزه‌ی ویدیو نیز، پن و همکاران [۵۹]، روشی به نام ST-Adapter پیشنهاد داده‌اند که با افزودن قابلیت زمانی-مکانی به مدل تصویری از پیش‌آموخته، آن را برای وظایف ویدیویی قابل استفاده کرده است.

مخلوط خبره‌ها

روش مخلوط خبره‌ها، با استفاده از یک مکانیزم دروازه‌ای، به‌صورت پویا تعدادی از شبکه‌های عصبی خبره را برای انجام هر وظیفه فعال می‌کند. این ساختار باعث می‌شود مدل بخش‌های مختلف خود را به وظایف متنوع اختصاص دهد و عملکرد بهتر و مقیاس‌پذیری بیشتری پیدا کند. ونگ و همکاران [۶۰]، در این زمینه رویکردی را ارائه داده‌اند که مدل به‌صورت خودکار تصمیم می‌گیرد که بسته به تغییر داده یا وظیفه، از کدام وفق دهنده‌های موجود استفاده کند یا وفق دهنده‌ی جدیدی اضافه نماید، تا تعادلی میان حفظ دانش قبلی و یادگیری دانش جدید ایجاد شود. در ادامه نیز، یو و همکاران [۱۸]، با استفاده از همین روش و با گسترش تدریجی مدل CLIP و استفاده از مسیرهای انتخابی میان وفق دهنده‌های خبره و مدل اصلی، قابلیت تشخیص یادگیری بدون نمونه حفظ شده و در عین حال بار محاسباتی به شکل چشم‌گیری کاهش می‌یابد.

۲-۶ جمع‌بندی

در این فصل، به بررسی یادگیری پیوسته با تمرکز بر حوزه‌های تصویر، ویدیو و مدل‌های بینایی-زبان پرداخته شد. ابتدا مفهوم یادگیری پیوسته معرفی شد که هدف آن توانایی سیستم‌های هوشمند در کسب و به‌روزرسانی دانش در طول زمان بدون فراموش کردن دانش پیشین است. این حوزه با چالش

اصلی «فراموشی فاجعه‌بار» روبه‌روست؛ مشکلی که هنگام ورود داده‌های جدید، باعث تضعیف یا حذف دانش قبلی مدل می‌شود.

رویکردهای اصلی مقابله با این چالش در چهار دسته‌ی مبتنی بر تنظیم، حافظه، بهینه‌سازی و معماری طبقه‌بندی شدند. هر یک از این روش‌ها مزایا و محدودیت‌های خاص خود را دارند، از جمله استفاده از حافظه خارجی برای بازپخش داده‌های قدیمی، یا تنظیم پارامترها برای حفظ دانش پیشین.

همچنین، مدل‌های بینایی-زبان معرفی و بررسی شدند. این مدل‌ها با ترکیب اطلاعات تصویری و متنی، قادر به درک عمیق‌تری از محتوای چندرسانه‌ای هستند و در وظایفی مانند توصیف تصاویر، پرسش‌وپاسخ تصویری و جستجوی مبتنی بر تصویر نقش مهمی ایفا می‌کنند. ادغام این مدل‌ها با رویکردهای یادگیری پیوسته، نیازمند راهکارهایی برای مدیریت همزمان دانش در هر دو حوزه بینایی و زبان است.

در مجموع، این فصل اهمیت یادگیری پیوسته در توسعه سامانه‌های هوشمند را برجسته کرد و نشان داد که طراحی روش‌های نوآورانه برای مدیریت داده‌های جدید و حفظ اطلاعات پیشین، به‌ویژه در مدل‌های بینایی-زبان، یکی از محورهای کلیدی پیشرفت در هوش مصنوعی آینده است.

فصل سوم

روش پیشنهادی

۳-۱ مقدمه

روش‌های موجود در زمینه یادگیری پیوسته برای داده‌های ویدیویی با وجود پیشرفت‌های اخیر، همچنان با مشکلات اساسی روبه‌رو هستند. برخی رویکردها به مدل‌های از پیش آموزش‌دیده متکی‌اند، اما برای انطباق با داده‌های ویدیویی نیاز به آموزش یا تنظیم مجدد کدگذارهای زمانی دارند، که فرآیندی زمان‌بر، پرهزینه و وابسته به منابع سخت‌افزاری سنگین است [۵۱]. هم‌چنین، روش‌های مختلف (تصویر یا ویدیو)، برای مقابله با فراموشی فاجعه‌بار به استفاده از بافرهای بازپخش یا ذخیره‌سازی داده‌های قبلی متکی هستند که نیازمند حافظه بالا و ناسازگار با محدودیت‌های حریم خصوصی است [۵۱، ۲۹، ۶۱] . علاوه بر این، برخی از روش‌ها از ساختارهای وابسته به وظیفه [۱] استفاده می‌کنند که مدیریت و نگهداری آن‌ها در سناریوهای واقعی و وظایف متوالی دشوار بوده و باعث کاهش تعمیم‌پذیری می‌شود [۶۲، ۶۳].

با توجه به این محدودیت‌ها، نیاز به رویکردهایی احساس می‌شود که بتوانند بدون وابستگی به ذخیره‌سازی وسیع داده‌های گذشته یا آموزش سنگین کدگذارها، عملکرد بهتری در داده‌های ویدیویی ارائه دهند و در عین حال ماهیت مستقل از وظیفه [۲] داشته باشند. هدف چنین رویکردهایی این است که از ظرفیت مدل‌های بزرگ و از پیش آموزش‌دیده استفاده کرده و با اضافه کردن لایه‌های سبک یا پرامپت‌های یادگیرنده، بدون تغییر مستقیم پارامترهای اصلی مدل، دانش قبلی را حفظ کنند و با داده‌های جدید تطبیق یابند.

در راستای رفع محدودیت‌های مذکور، در این فصل به ارائه روشی با عنوان ProActionCLIP[۳]، پرداخته می‌شود که با ترکیب قابلیت‌های روش Open-VCLIP [۲۰] و ایده‌های روش L2P [۷]، از پرامپت‌های سبک و پویا، برای انطباق با وظایف جدید، استفاده می‌کند. بنابراین در این روش، نیازی به طی کردن فرآیند آموزش کدگذارهای زمانی، که از نظر محاسباتی هزینه‌بر است، وجود ندارد. روش ProActionCLIP، با بهره‌گیری بهینه از دانش مدل‌های پیش‌آموزش‌دیده، کاهش فراموشی فاجعه‌بار و حفظ کارآیی در سناریوهای واقعی را هدف قرار داده است. این روش می‌تواند با بهره‌گیری بهینه از منابع محاسباتی و نیاز کمتر به سخت‌افزار، توانایی یادگیری پیوسته وظایف جدید را داشته باشد و بدون وابستگی به شناسه وظایف، عملکردی کارآمد و مقیاس‌پذیر ارائه دهد.

[1] Task specific

[2] Task-agnostic

[۳] این نام مخفف Prompt Action recognition CLIP می‌باشد که به استفاده از روش پرامپت‌گذاری برای تشخیص حرکت توسط مدل کلیپ، اشاره می‌کند.

۳-۲ روش ProActionCLIP

روش ProActionCLIP، که برای یادگیری پیوسته‌ی تشخیص حرکت انسان معرفی شده، بر پایه‌ی ترکیب دو رویکرد Open-VCLIP و L2P، بنا شده است. ایده‌ی اصلی این روش آن است که از قابلیت‌های Open-VCLIP برای استخراج ویژگی‌های چندماهیتی (تصویر-متن) بهره گرفته و در عین حال از سازوکار پرامپت‌های یادگیرنده در L2P استفاده می‌کند. به‌این‌ترتیب مدل می‌تواند بدون نیاز به تغییر مستقیم پارامترهای کدگذار اصلی، خود را با وظایف متوالی تطبیق دهد. این ترکیب باعث می‌شود که مشکل فراموشی فاجعه‌بار کاهش یافته، حافظه‌ی مورد نیاز برای ذخیره‌سازی نمونه‌ها به حداقل برسد و مدل به صورت مستقل از وظیفه، وظایف جدید را پردازش کند. به عبارت دیگر، با این رویکرد سعی شده است مزیت‌های هر دو روش با هم ترکیب شود که عبارت‌اند از: قدرت تعمیم‌دهی و دانش وسیع Open-VCLIP و انعطاف‌پذیری L2P در مدیریت وظایف پیوسته. مدل ProActionCLIP، شامل دو مرحله‌ی آموزش و آزمون می‌باشد که پس از معرفی مدل‌های پایه‌ی بکار برده شده در ProActionCLIP، به توضیح آن‌ها پرداخته می‌شود.

۳-۳ مدل Open-VCLIP

مدل‌های بینایی-زبان مانند CLIP، به دلیل توانایی یادگیری بازنمایی‌های مشترک تصویر و متن، عملکرد قابل توجهی در وظایف بینایی و زبانی داشته‌اند. با این حال، این مدل‌ها در حالت پایه برای داده‌های ایستا (تصاویر) طراحی شده‌اند. Open-VCLIP با گسترش معماری CLIP و افزودن قابلیت درک اطلاعات زمانی، این محدودیت را برطرف می‌کند و روشی کارآمد برای تحلیل ویدیو ارائه می‌دهد. مدل Open-VCLIP هم‌چنین به‌منظور حفظ قابلیت یادگیری بدون نمونه‌ی مدل پایه‌ی CLIP و جلوگیری از فراموشی آن در فرآیند یادگیری داده‌های ویدیویی، از دو تکنیک منظم‌سازی وزن‌های درون‌یابی [۴] و میانگین تصادفی وزن‌ها [۵] استفاده می‌کند. تکنیک اول با بکارگیری روش ترکیب مدل قدیمی و جدید و تغییراتی در آن، از فراموشی دانش قبلی جلوگیری کرده و هم‌زمان باعث یادگیری دانش جدید می‌شود. در تکنیک دوم، با میانگین‌گیری از وزن‌های مدل در نقاط مختلف، باعث بهبود وزن‌های نهایی از لحاظ تعمیم‌پذیری می‌شوند. در ادامه، ابتدا نحوه‌ی تعمیم مدل CLIP برای داده‌های ویدیویی بررسی شده و سپس تکنیک‌های منظم‌سازی وزن‌های درون‌یابی و میانگین تصادفی وزن‌ها شرح داده خواهد شد.

[4] Interpolation Weight Regularization (IWR)

[5] Stochastic Weight Averaging (SWA)

۳-۳-۱ تعمیم مدل CLIP برای داده های ویدیویی

برای تعمیم مدل CLIP برای داده‌های ویدیویی، ورودی ویدیویی به دنباله‌ای از قاب‌ها تبدیل می‌شود و هر قاب توسط کدگذار تصویری CLIP به یک بردار ویژگی تبدیل می‌گردد. سپس این ویژگی‌ها در قالب دنباله‌ای زمانی قرار گرفته و با سازوکار توجه ترکیب می‌شوند. مفهومی به نام وصله‌ی[۶] تصویر که در ادامه ذکر شده است، ناحیه‌ای مستطیلی و پیوسته از تصویر اصلی با ابعاد مشخص است که به‌منظور بررسی جزئیات محلی و ویژگی‌های بخش‌های مختلف تصویر استفاده می‌شود (مانند نمونه‌ی تیره شده در شکل ۳-۱). در سازوکار توجه فرمول محاسبه‌ی خروجی مطابق رابطه‌ی (۳-۱) می‌باشد:

$$y'_{s,t} = \text{Softmax}\left(\frac{q_{s,t}E_t^\top}{\sqrt{d}}\right)V_t. \tag{۳-۱}$$

در این رابطه، $q_{s,t}$ بردار پرسمان[۷] برای یک وصله از تصویر، E_t و V_t به ترتیب نشان‌دهنده‌ی ماتریس کلید و ماتریس مقدار متناظر با قاب یا تصویر t هستند که از طریق سازوکار توجه ترکیب می‌شوند. از d به‌عنوان عامل نرمال‌سازی استفاده می‌شود. به این ترتیب، با به کارگیری رابطه‌ی (۳-۱)، ارتباط یک وصله از تصویر با خودش و بقیه‌ی وصله‌های تصویر در $y'_{s,t}$ قرار داده می‌شود. به‌منظور تعمیم مدل برای داده‌های ویدیویی، مدل Open-VCLIP، بردار پرسمان وصله‌ی قاب فعلی را در ماتریس کلید قاب فعلی، بعدی و قبلی ($E_{(t-1)\sim(t+1)}$) ضرب می‌کند و پس از اعمال تابع softmax، آن را در ماتریس مقدار قاب فعلی، بعدی و قبلی ($V_{(t-1)\sim(t+1)}$) ضرب می‌کند و سازوکار توجه را براساس رابطه‌ی (۳-۲)، محاسبه می‌کند. به این ترتیب ارتباط وصله قاب فعلی با سایر وصله‌های این قاب و قاب‌های قبل و بعد از قاب فعلی درنظر گرفته می‌شود (مطابق با شکل ۳-۱). این راهکار به ظاهر ساده، توانست تحول خوبی در زمینه‌ی سازگاری مدل CLIP با داده‌ی ویدیویی ایجاد کند [۲۰].

$$y'_{s,t} = \text{Softmax}\left(\frac{q_{s,t}\left[E_{(t-1)\sim(t+1)}\right]^\top}{\sqrt{d}}\right)\left[V_{(t-1)\sim(t+1)}\right]. \tag{۳-۲}$$

[6] Patch

[7] Query

شکل ۳-۱: در سازوکار تغییریافته‌ی توجه، وصله‌های در نظر گرفته شده برای هر وصله از قاب (مانند وصله‌ی تیره در شکل)، شامل وصله‌های قاب فعلی و بعدی و قبلی می‌باشند.

۳-۳-۲ منظم‌سازی مبتنی بر درون‌یابی وزن‌ها

همانطور که ذکر شد، برای جلوگیری از فراموشی دانش پیش‌آموزش و در عین حال سازگار کردن مدل با داده‌های جدید، منظم‌سازی مبتنی بر درون‌یابی وزن‌ها معرفی شده است. در این روش، وزن‌های مدل به‌صورت ترکیبی از وزن‌های اولیه (پیش‌آموزش) و وزن‌های به‌روزرسانی‌شده در وظیفه جدید، تنظیم می‌شوند. این درون‌یابی به مدل کمک می‌کند تا در حین یادگیری، تعادلی میان دانش قدیمی و اطلاعات تازه برقرار کرده و از بیش‌برازش[8] جلوگیری کند. روش مذکور، تعمیم ایده‌ی گابریل و همکاران [۶۴]، که مطابق با رابطه‌ی (۳-۳) است، می‌باشد:

$$\theta = \lambda\theta_A + (1-\lambda)\theta_B. \tag{۳-۳}$$

در این رابطه، θ از ترکیب خطی وزن‌های مدل پایه θ_A و مدل به‌روزرسانی‌شده θ_B، با ضریب λ، تشکیل می‌شود تا صحت مدل در وظایف جدید افزایش یابد، بدون آنکه عملکرد آن در سایر وظایف که از پیش بهینه بوده‌اند، کاهش پیدا کند. با توجه به این که λ یک ابرپارامتر[9] است و در رابطه‌ی (۳-۳) هیچ فرآیند بهینه‌سازی مستقیمی روی مدل نهایی انجام نمی‌شود، مدل ترکیبی ممکن است روی داده‌های جدید (θ_B) عملکرد ضعیفی داشته باشد و دچار زیربرازش[10] شده و همچنین، کیفیت مدل به شدت به مقدار پارامتر تعادلی λ وابسته شود [۲۰]. پس در مدل Open-VCLIP راه حلی ارائه شد که عملکرد مدل ترکیبی را در برابر بازه‌ای از مقادیر λ بهینه کند. مطابق با رابطه‌ی (۳-۴)، وزن‌های جدید در آموزش

[8] Overfitting
[9] Hyper parameter
[10] Underfitting

به سمتی می‌رود که هم زیان مدل جدید و هم زیان مدل ترکیبی با ضریب α روی داده‌های جدید حداقل شود. در این حالت، ترکیب‌های مختلف مدل قبلی و جدید در مراحل مختلف آموزش در نظر گرفته می‌شود که در واقع، عملکرد بهتری در تحلیل داده‌های نادیده خواهد داشت. در انتها با استفاده از رابطه‌ی (۳-۳)، مدل جدید و قدیم ترکیب خواهند شد. با این تفاوت که وزن‌های مدل جدید، در طول آموزش با در نظر گرفتن عدم فراموشی مدل قبلی، یاد گرفته شده‌اند:

$$\arg\min_{\theta_B} \mathcal{L} = L(\theta_B; D_B) + \beta L\big(\alpha\theta_A + (1-\alpha)\theta_B; D_B\big). \tag{۴-۳}$$

پارامتر α در رابطه‌ی (۴-۳)، از یک توزیع یکنواخت در بازه $(0, \lambda)$ نمونه‌برداری می‌شود و ضریب β به‌عنوان یک پارامتر تنظیم‌کننده برای کنترل میزان تاثیر عبارت درون‌یابی تعریف شده است. همچنین مقدار β به‌صورت $\beta = C\frac{1}{1-\alpha}$ محاسبه می‌شود که در آن C یک مقدار ثابت برای کنترل بزرگی β است.

۳-۳-۳ میانگین‌گیری تصادفی وزن‌ها

در فرآیند آموزش شبکه‌های عصبی، مدل‌ها معمولاً در نواحی کمینه تیز[11] در فضای وزن‌ها قرار می‌گیرند؛ این نقاط اگرچه روی داده‌های آموزش خطای پایینی دارند، اما به دلیل حساسیت زیاد به تغییرات وزن، معمولاً تعمیم خوبی روی داده‌های جدید ندارند. به‌منظور بهبود قابلیت تعمیم مدل، پاول و همکاران [۶۵]، روش میانگین‌گیری تصادفی وزن‌ها[12] را معرفی کردند که با میانگین‌گیری از وزن‌های مدل در نقاط مختلف، طی فرآیند آموزش، مدل را به سمت کمینه مسطح[13] هدایت می‌کند. این نواحی از فضای وزن‌ها تغییرات کمتری در تابع زیان دارند و باعث کاهش خطای آزمون و بهبود توانایی تعمیم مدل می‌شوند. به منظور بهبود تعمیم‌پذیری مدل Open-VCLIP نیز، از روش مذکور استفاده شده است. فرمول نهایی مدل Open-VCLIP با احتساب روش میانگین‌گیری تصادفی وزن‌ها در J گام آموزش، در رابطه‌ی (۵-۳) آورده شده است:

$$\sum_{i}^{J} \frac{\lambda\theta_A + (1-\lambda)\theta_i}{J} = \lambda\theta_A + (1-\lambda)\left(\frac{1}{J}\sum_{i}^{J}\theta_i\right). \tag{۵-۳}$$

[11] Sharp minimum

[12] Stochastic Weight Averaging (SWA)

[13] Flat minimum

در نهایت همان‌طور که ذکر شد، مدل Open-VCLIP نهایی، نسخه‌ی تنظیم دقیق‌شده از مدل CLIP با سازوکار توجه تغییریافته، است. این مدل به‌منظور جلوگیری از فراموشی دانش مدل اولیه و در عین حال حفظ قابلیت یادگیری بدون نمونه، دو تکنیک مکمل منظم‌سازی مبتنی بر درون‌یابی وزن‌ها و میانگین‌گیری تصادفی وزن‌ها را به کار گرفته است.

۳-۴ مدل L2P

روش ارائه‌شده در این مقاله، با هدف بهبود یادگیری پیوسته، از سازوکاری مبتنی بر «استخر پرامپت» استفاده می‌کند. در این رویکرد، به‌جای تغییر پارامترهای اصلی مدل، مجموعه‌ای از پرامپت‌های قابل‌آموزش طراحی می‌شود که مدل با استفاده از آن‌ها قادر به استخراج اطلاعات مهم از داده‌های ورودی است. در هر مرحله یادگیری، پرامپت‌های مناسب بر اساس شباهت با داده‌های جدید انتخاب می‌شوند و این امر باعث می‌شود مدل بتواند دانش جدید را یاد بگیرد، بدون آنکه دانش قبلی را فراموش کند. این روش با بهره‌گیری از معماری ترنسفورمر، توانسته است تعادل موثری میان حفظ دانش گذشته و یادگیری وظایف جدید برقرار کند. همانطور که در شکل ۲-۴ مشاهده شد، این مدل دارای دو بخش انتخاب پرامپت‌ها و یادگیری و به‌روزرسانی پرامپت‌ها است که هر یک در ادامه توضیح داده می‌شود.

۱-۴-۳ انتخاب پرامپت

در بخش انتخاب پرامپت[۱۴]، از استخر پرامپت تعدادی پرامپت متناسب با ورودی تصویری انتخاب می‌شود. استخر پرامپت شامل تعدادی پرامپت به تعداد M بوده که به صورت $\mathbf{P} = \{P_1, P_2, \cdots, P_M\}$ نمایش داده می‌شوند. هر $P_j \in \mathbb{R}^{L_p \times d_{in}}$ یک پرامپت منفرد با تعداد نشانه‌ی L_p و اندازه تعبیه‌ی[۱۵] d_{in} مشابه اندازه‌ی تعبیه‌ی ورودی است. همچنین هر پرامپت، دارای یک کلید است و فرآیند انتخاب بر اساس شباهت این کلیدها با بردار ویژگی مرتبط با کلاس ورودی انجام می‌گیرد. بردار ویژگی موردنظر از طریق استخراج‌کننده‌ی ویژگی، به‌دست می‌آید و سپس با کلیدهای موجود در استخر مقایسه می‌شود. در نهایت، تعدادی از پرامپت‌های کلیدهایی که بیشترین شباهت را با بردار ویژگی دارند انتخاب می‌شوند (مطابق با شکل ۲-۴). مجموعه‌ای از کلیدها به‌صورت $\mathbf{K} = \{k_i\}_{i=1}^{M}$ نمایش داده می‌شود که هر $k_i \in \mathbb{R}^{d_k}$ است. استخراج‌کننده‌ی ویژگی برای مقایسه با کلیدها، به صورت $z : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^{d_k}$

[14] Prompt selection

[15] Embedding size

معرفی می‌شود که بردار ویژگی x با ابعاد ارتفاع (H) در عرض (W) در تعداد کانال رنگی (C)، را به ابعاد کلید نگاشت می‌کند. در واقع از مدل ازپیش‌آموزش‌دیده‌ی منجمد[۱۶] به‌عنوان استخراج‌کننده ویژگی استفاده می‌شود: $z(x) = f(x)[0, :]$ (که در آن از بردار ویژگی متناظر با کلاس استفاده می‌شود). تابع $\gamma : \mathbb{R}^{d_k} \times \mathbb{R}^{d_k} \to \mathbb{R}$ به‌عنوان معیاری برای سنجش میزان تطابق بین بردار ویژگی ورودی و کلید پرامپت تعریف می‌شود (مانند فاصله کسینوسی). در نهایت، انتخاب N شبیه‌ترین کلید طبق رابطه‌ی (۳-۶) به صورت زیر نشان داده شده است:

$$\mathbf{K}_x = \underset{\{s_i\}_{i=1}^{N} \subseteq [1,M]}{\arg\min} \sum_{i=1}^{N} \gamma\big(z(x), \mathbf{k}_{s_i}\big), \tag{۳-۶}$$

$\{s_i\}_{i=1}^{N}$ به‌عنوان یک زیرمجموعه از N کلید از بازه $[1, M]$ در نظر گرفته می‌شود و $\mathbf{K}_x$ نشان‌دهنده یک زیرمجموعه از بهترین N کلید انتخاب‌شده از $\mathbf{K}$ به‌طور خاص برای نمونه x است.

علاوه بر این، به روش انتخاب پرامپت، قابلیت اضافه‌ای نیز اضافه شده است به این صورت که برای پرامپت‌های قبلا به‌روزرسانی شده، جریمه در نظر گرفته است تا متناسب با تعداد تکرارشان در وظایف قبلی، جریمه‌ی بیشتری برای انتخاب بگیرند. در این حالت، پرامپت‌های با تکرار کم نیز شانس انتخاب شدن پیدا می‌کنند و پرامپت‌ها با تکرار بیشتر، کمتر تغییر می‌کنند و تداخل کمتر می‌شود.

۳-۴-۲ یادگیری پرامپت

در بخش یادگیری پرامپت، پرامپت‌های انتخاب‌شده به‌همراه داده‌ی ورودی به مدل تغذیه شده و پس از عبور از لایه‌های ترنسفورمر، بخش خروجی مربوط به پرامپت‌ها، استخراج شده و با میانگین‌گیری، به دسته‌بند[۱۷] منتقل می‌شود. سپس با انجام عملیات پس‌انتشار[۱۸]، وزن‌های پرامپت‌ها و کلیدهای متناظر آن‌ها به‌روزرسانی می‌شوند. این فرآیند باعث می‌شود مدل ضمن یادگیری وظایف جدید، قابلیت تعمیم خود را افزایش داده و دانش قبلی را حفظ کند (مطابق با شکل ۲-۴). تابع زیان در این مدل مطابق با (۳-۷)، به صورت ترکیبی معرفی می‌شود به این صورت که بخش اول شامل زیان بین برچسب تصویر و پیش‌بینی دسته‌بند از ورودی دارای پرامپت و بخش دوم شامل تفاوت بین کلیدهای انتخاب شده و ویژگی استخراجی از ورودی می‌باشد. به عبارتی تلاش بر این است که علاوه بر تقویت پرامپت‌ها، کلیدهای

[16] Frozen
[17] Classifier
[18] Backpropagation

مرتبط نیز، به ویژگی ورودی متناظرشان نزدیک‌تر شوند.

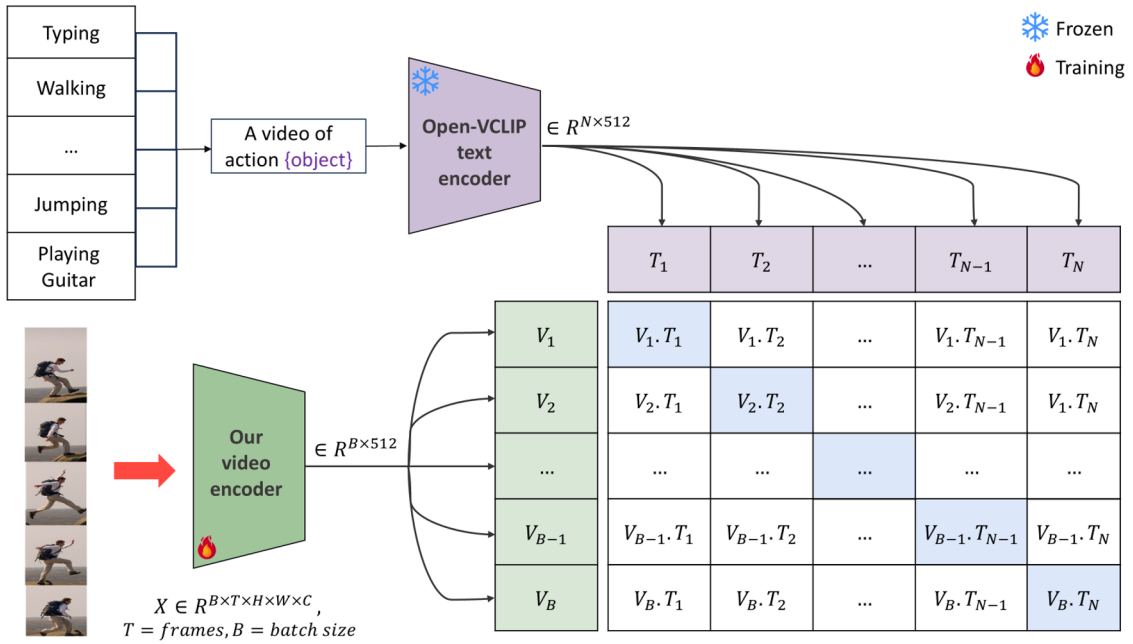
$$\min_{\mathbf{P},\mathbf{K},\phi} \mathcal{L}\big(g_\phi(f_r^{\text{avg}}(x_p)), y\big) + w \sum_{\mathbf{K}_x} \gamma\big(z(x), \mathbf{k}_{s_i}\big) \tag{۷-۳}$$

همانطور که دیده می‌شود، بخش اول شامل تابع زیان (کراس انتروپی) بین خروجی دسته‌بند g_ϕ و برچسب واقعی y محاسبه می‌شود. تابع $f_r^{\text{avg}} = \text{AvgPool}(f_r(x_p)[0 : NL_p, :])$، به این معناست که بردارهای پنهان خروجی متناظر با مکان‌های $N \cdot L_p$ پرامپت‌ها، پیش از ورود به لایه‌ی دسته‌بند، به‌صورت میانگین‌گیری شده ترکیب می‌شوند. در عبارت $f_r(x_p)$، تابع f_r نشان‌دهنده بخش باقی‌مانده از مدل ازپیش‌آموزش‌دیده است که پس از الحاق پرامپت‌ها به ورودی، عملیات پردازش ویژگی را انجام می‌دهد. بردار x_p نیز از اتصال پرامپت‌ها به بردار تعبیه‌شده ورودی حاصل می‌شود. در بخش دوم، پارامتر w وزن تنظیم‌کننده است که میزان اهمیت معیار نزدیکی بین کلیدهای انتخاب‌شده و ویژگی ورودی را کنترل می‌کند. فاصله بین بردار $z(x)$ و کلیدهای انتخاب‌شده $\mathbf{k}_{s_i}$ را کاهش می‌دهد. مجموعه $\mathbf{K}_x$ نیز نشان‌دهنده کلیدهای انتخاب‌شده برای ورودی x است که طبق معادله (۶-۳) به دست می‌آیند.

در ادامه به توضیح مدل پیشنهادی و اجزای آن پرداخته می‌شود.

۵-۳ مرحله‌ی آموزش ProActionCLIP

همانطور که پیش‌تر اشاره شد، در مرحله‌ی آموزش، تمرکز بر یادگیری پرامپت‌های مناسب برای دسته‌های مختلفی است که به صورت پیوسته به مدل اضافه می‌شوند. طرح کلی مدل در مرحله‌ی آموزش در شکل ۲-۳، نشان داده شده است که الگو گرفته از مدل CLIP می‌باشد. این مدل شامل یک کدگذار ویدیو (قابل به‌روزرسانی) و کدگذار متن منجمد (غیر قابل به‌روزرسانی) از مدل Open-VCLIP است به این صورت که بردار ویژگی ویدیوی استخراج شده از کدگذار ویدیو و بردارهای ویژگی برچسب‌های موجود استخراج شده از کدگذار متن، طبق روش تقابلی مقایسه شده و طبق نزدیک شدن موارد متناظر و دور شدن موارد نامتناظر، وزن‌های پرامپت‌ها تغییر داده می‌شوند. دو بخش اصلی مدل به نام کدگذار ویدیو و یادگیری پرامپت در ادامه توضیح داده خواهند شد.



شکل ۳-۲: طرح کلی مرحله‌ی آموزش مدل ProActionCLIP. در بخش پایین سمت چپ تصویر، ویدیو وارد کدگذار ویدیویی مدل شده و خروجی آن به‌صورت برداری با ابعاد ۵۱۲ استخراج می‌شود. این بردار با ویژگی‌های متنی حاصل از کدگذار متن (مربوط به برچسب‌ها) مقایسه می‌گردد. در فرآیند آموزش، پرامپت‌های یادگیرنده در کدگذار ویدیو، به‌گونه‌ای به‌روزرسانی می‌شوند که ویژگی‌های ویدیو به ویژگی‌های متناظر با برچسب واقعی خود نزدیک‌تر شوند.

۳-۵-۱ کدگذار ویدیو

این بخش، از ترکیب L2P و Open-VCLIP تشکیل شده است. به طور کلی مطابق شکل ۳-۳، ویدیو به عنوان ورودی، وارد کدگذار Open-VCLIP و لایه‌ی کانولوشنی دوبعدی می‌شود. ویژگی کلاس[19] از خروجی کدگذار Open-VCLIP با ۵۱۲ بعد، با کلیدهای داخل استخر پرامپت مقایسه می‌شود. به تعداد N پرامپت از مشابه‌ترین کلیدها انتخاب می‌شوند. از طرف دیگر ویدیو از لایه‌ی کانولوشنی عبور کرده و پرامپت‌ها به هر قاب، به صورت جداگانه متصل می‌شوند. سپس به کدگذار ترنسفورمر معرفی شده در Open-VCLIP، وارد شده و به ازای هر قاب، یک ویژگی کلاس بدست می‌آید که میانگین آن‌ها محاسبه و به عنوان خروجی نهایی این بخش، ارائه می‌گردد. در مدل پیشنهادی، صرفا پرامپت‌ها و کلیدهای متناظر آن‌ها، قابل یادگیری هستند و بقیه‌ی اجزا به صورت منجمد استفاده می‌شوند. در این بخش از تحقیق، آزمایش‌های مختلفی اجرا شد که بر اساس نوع انتخاب پرامپت و شرایط استخر پرامپت می‌توان به دسته‌های زیر تقسیم نمود:

- **مقداردهی اولیه‌ی کلید پرامپت:** از آن جایی که ابتدای آموزش، مقادیر اولیه به صورت تصادفی

[19] Class feature (CLS)

هستند، در این آزمایش، مقادیر کلیدها، معادل ویژگی‌های برچسب دسته‌های استخراج شده از کدگذار متن قرار داده شد. در این صورت، مقایسه‌ی ویژگی ویدیو و کلیدها به صورت بهینه‌تر و دقیق‌تری صورت می‌گیرد.

- **وزن‌دهی به کلید پرامپت‌های از قبل انتخاب شده:** مطابق با روش اضافه‌ای که برای انتخاب پرامپت در L2P مطرح شد، تعداد تکرار پرامپت‌ها در هر وظیفه محاسبه می‌شود. در وظیفه‌ی بعدی، هرچه تکرار پرامپت بیشتر بوده باشد، تاثیرش در انتخاب کمتر می‌شود.

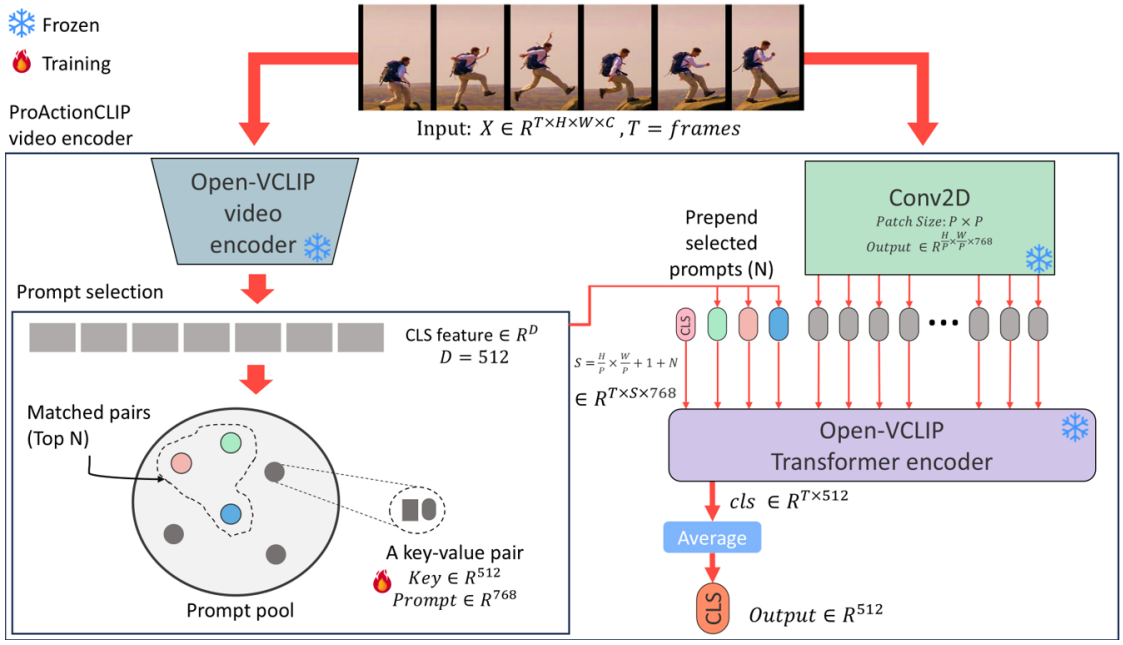
- **منجمد کردن پرامپت‌های قبلی:** یکی از راه‌های استفاده از وزن‌های قبلی، این است که به صورت منجمد استفاده شوند و به‌روزرسانی نشوند. این روش کمک می‌کند اطلاعات اختصاصی هر وظیفه از بین نرود و اگر داده‌ی جدید اشتراکی با قبلی‌ها داشته باشد، پرامپت آن‌ها را انتخاب خواهد کرد.

- **پویا بودن تعداد پرامپت‌های استخر پرامپت:** در فرض اولیه، در استخر پرامپت تعدادی ثابت پرامپت وجود داشت اما برای بهینه‌بودن مدل برای دسته‌های بیشتر و موجود بودن پرامپت کافی در هر وظیفه، در این قسمت پرامپت‌ها در ابتدای هر وظیفه افزایش میابد.

ترکیب برخی از این روش‌ها نیز آزمایش شده است مانند استفاده از استخر پویا و مقداردهی اولیه کلیدها.

۳-۵-۲ یادگیری پرامپت

وقتی ویدیو از کدگذار ویدیوی پیشنهادی عبور کرد، وارد مرحله‌ی نهایی برای به‌روزرسانی وزن‌های پرامپت‌ها و کلیدهای متناظرشان می‌شود. مطابق با شکل ۳-۲، ویژگی برچسب‌ها از طریق کدگذار متن مدل Open-VCLIP بدست می‌آید. تابع زیان همانند روش L2P، شامل دو بخش است. اولین بخش شامل زیان بین ویژگی استخراجی از ویدیو و ویژگی برچسب متناظر آن و بخش بعدی مختص زیان بین ویدیو و کلیدهای انتخاب شده برای آن می‌باشد. در این صورت پرامپت‌ها در بخش اول و کلیدها در بخش دوم تابع زیان مورد تمرکز قرار می‌گیرند. تابع زیان مدل پیشنهادی با L2P تفاوت‌هایی دارد که در ادامه بررسی می‌کنیم. روش مدل پیشنهادی برگرفته از مدل CLIP، برخلاف L2P که با دسته‌بند است، به صورت تقابلی می‌باشد. مطابق (۳-۸)، پرامپت‌های انتخابی، به ورودی عبور کرده از لایه‌ی کانولوشنی، ملحق شده و $x_{p,t}$ را تشکیل می‌دهند. t نشان‌دهنده‌ی هر قاب است. سپس از لایه‌ی ترنسفورمر معرفی



شکل ۳-۳: سازوکار بخش کدگذار ویدیو مدل ProActionCLIP. ویدیو وارد کدگذار ویدیو می‌شود. در مسیر سمت چپ تصویر، خروجی کدگذار Open-VCLIP تولید شده و با کلیدهای موجود در استخر پرامپت مقایسه می‌گردد تا پرامپت‌های متناظر با شبیه‌ترین کلیدها انتخاب شوند. در مسیر سمت راست، هر قاب ویدیو پس از عبور از لایه‌ی کانولوشنی، با پرامپت‌های انتخاب‌شده ترکیب می‌شود. این ترکیب‌ها از کدگذار ترنسفورمر Open-VCLIP عبور می‌کند. سپس نتایج به‌صورت میانگین‌گیری ادغام شده و در نهایت ویژگی کلاس برای تولید خروجی نهایی انتخاب می‌شود.

شده، عبور کرده و به صورت $a(x_{p,t})$ نمایش داده می‌شود. خروجی ترنسفورمر، در بردارهای ویژگی برچسب‌ها که از کدگذار متن بدست آمده (W)، ضرب می‌شود. از τ نیز به عنوان ضریب مقیاس‌دهی در این ضرب استفاده می‌شود. در این حالت شباهت ویژگی بدست آمده از قاب t از ویدیو، با برچسب‌ها سنجیده می‌شود. سپس میانگین شباهت‌ها برای قاب‌های ویدیو (T) محاسبه می‌شود. تابع زیان کراس انتروپی بین خروجی و برچسب متناظر اجرا می‌شود. بخش دوم، مانند روش L2P، بر نزدیک کردن کلیدهای انتخاب شده، به ویژگی کلاس ویدیوی نمونه، سعی دارد که در بخش قبل به تفصیل شرح داده شد.

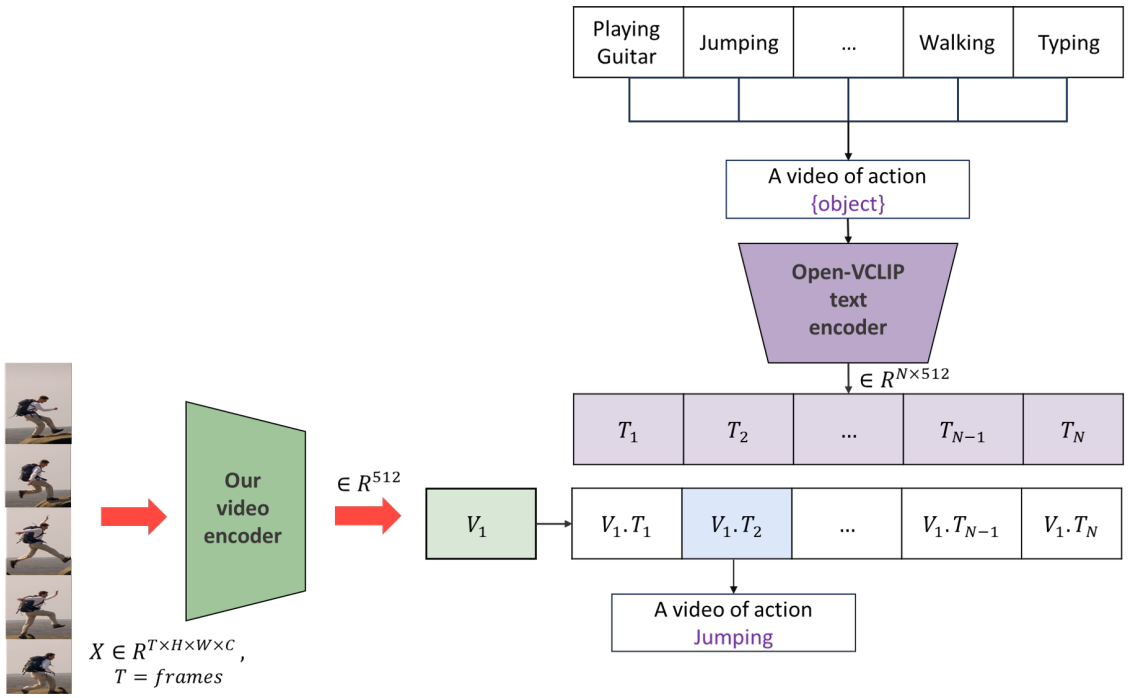
$$\mathcal{L} = \mathrm{CE}\left(\frac{1}{T}\sum_{t=1}^{T}\left[\tau \cdot \mathbf{a}(\mathbf{x_{p,t}})\mathbf{W}^{\top}\right], y\right) + w\sum_{\mathbf{K}_x}\gamma\big(z(x), \mathbf{k}_{s_i}\big) \qquad (۸-۳)$$

۳-۶ مرحله‌ی آزمون ProActionCLIP

برای آزمون مدل، مطابق شکل ۳-۴، ویدیو وارد کدگذار ویدیو می‌شود و به تعداد N نزدیکترین کلید به ویژگی ویدیو، را انتخاب کرده و به هر یک از قاب‌های ویدیو ملحق کرده و از ترنسفورمر عبور می‌دهد. خروجی نهایی این بخش را با ویژگی استخراج شده‌ی برچسب‌ها از کدگذار متن، مقایسه کرده و برچسب با بیشترین شباهت انتخاب می‌شود.

۳-۷ جمع‌بندی

در این فصل مدل نهایی پیشنهادی این پژوهش با نام ProActionCLIP معرفی شد که ترکیبی از دو مدل L2P و Open-VCLIP بوده که با تلفیق قابلیت یادگیری پرامپت در L2P و توانایی استخراج ویژگی‌های ویدیویی در Open-VCLIP طراحی شده است. در این فصل، ساختار کلی مدل پیشنهادی و اجزای اصلی آن شامل کدگذار ویدیو، کدگذار متن منجمد و سازوکار یادگیری پرامپت تشریح شد. تکنیک‌های مختلف استفاده شده در طراحی استخر پرامپت‌ها و روش انتخاب پرامپت‌های مناسب برای هر ورودی، نیز شرح داده شد. همچنین نحوه تعامل این اجزا با یکدیگر برای یادگیری پیوسته و به‌روزرسانی وزن‌های پرامپت‌ها از طریق الگوریتم پس‌انتشار توضیح داده شد. در نهایت، فرآیند آموزش مدل و محاسبه تابع زیان که شامل عبارت منظم‌سازی (برای نزدیک کردن کلیدها و ویژگی ویدیو) و معیار تقابلی است، مورد بررسی قرار گرفت.



شکل ۳-۴: طرح کلی مرحله‌ی آزمون مدل ProActionCLIP. در بخش پایین سمت چپ تصویر، ویدیو وارد کدگذار ویدیویی مدل شده و خروجی آن به‌صورت برداری با ابعاد ۵۱۲ استخراج می‌شود. این بردار با ویژگی‌های متنی حاصل از کدگذار متن (مرتبط با برچسب‌های مختلف) مقایسه می‌گردد. در نهایت، برچسبی که بیشترین شباهت را با ویژگی استخراج‌شده از ویدیو دارد به‌عنوان برچسب پیش‌بینی‌شده انتخاب می‌شود.

فصل چهارم

نتایج آزمایش‌های تجربی

۴-۱ مقدمه

در این فصل، عملکرد مدل ProActionCLIP در زمینه‌ی یادگیری پیوسته‌ی تشخیص حرکت انسان مورد ارزیابی قرار می‌گیرد. هدف از این ارزیابی، بررسی میزان صحت مدل در یادگیری وظایف جدید، ارزیابی میزان فراموشی دانش پیشین و تحلیل بهره‌وری محاسباتی آن از نظر مصرف سخت‌افزار و تعداد پارامترهای قابل‌آموزش است. به منظور ارزیابی جامع، مدل پیشنهادی با روش‌های مطرح در این حوزه مانند PIVOT [۵۱]و سایر رویکردهای مرجع مقایسه می‌شود. برای این منظور، از مجموعه‌داده‌های نظیر UCF101 [۲۱] و HMDB51 [۲۲] استفاده شده است.

در ادامه‌ی این فصل، ابتدا معیارهای ارزیابی شامل صحت، میزان فراموشی معرفی می‌شوند. سپس، جزئیات مربوط به مجموعه‌های داده، تنظیمات آزمایشی، نتایج و مقایسه ارائه خواهد شد تا ارزیابی مدل پیشنهادی به‌طور کامل و شفاف صورت گیرد. در اخر نیز پیچیدگی محاسباتی بررسی خواهد شد.

۴-۲ معیارهای ارزیابی

در یادگیری پیوسته، دو معیار ارزیابی اهمیت دارند. یکی از آن‌ها میانگین صحت وظایف با وجود یادگیری سایر وظایف بوده و دیگری میزان فراموشی مدل پس از یادگیری هر وظیفه است که در ادامه هر یک شرح داده خواهد شد.

۴-۲-۱ میانگین صحت

در یادگیری پیوسته، عملکرد مدل روی تمام وظایف یادگرفته شده تاکنون توسط میانگین صحت ارزیابی می‌گردد. دو نوع صحت $top1$ و $top5$ در این تحقیق بررسی می‌شود. صحت $top1$ بیان‌گر این است که پیش‌بینی‌کننده، از بین احتمالات بدست آمده، برچسب با بیشترین احتمال را انتخاب می‌کند و سپس برابر بودن آن با برچسب واقعی داده‌ی بررسی‌شده، سنجیده می‌شود. اما در صحت $top5$، پنج تا برچسب با بیشترین احتمال، انتخاب شده و وجود یا عدم وجود برچسب واقعی در این پنج برچسب بررسی می‌شود. مطابق با (۴-۱) و (۴-۲)، میانگین صحت در وظیفه‌ی فعلی، برای همین وظیفه و وظایف قبلی، در حالت $top1$ و $top5$ بدست می‌آید. $\mathrm{ACC}_{i,t}^{top1}$ نشانگر صحت $top1$ برای وظیفه‌ی i بعد از یادگیری وظیفه‌ی t

و $\mathrm{ACC}_{i,t}^{top5}$ نشانگر صحت $top5$ برای وظیفه‌ی i بعد از یادگیری وظیفه‌ی t می‌باشد.

$$A_{top1}(t) = \frac{1}{t}\sum_{i=1}^{t}\mathrm{ACC}_{i,t}^{top1} \quad (۱-۴)$$

$$A_{top5}(t) = \frac{1}{t}\sum_{i=1}^{t}\mathrm{ACC}_{i,t}^{top5} \quad (۲-۴)$$

۴-۲-۲ میزان فراموشی

در یادگیری پیوسته، مدل باید بتواند وظایف جدید را یاد بگیرد بدون اینکه دانش وظایف قبلی را فراموش کند. اما معمولاً پدیده‌ی فراموشی فاجعه‌آمیز رخ می‌دهد؛ یعنی مدل پس از یادگیری وظایف جدید، صحت آن روی وظایف قدیمی کاهش پیدا می‌کند. معیار فراموشی برای اندازه‌گیری میزان این افت عملکرد تعریف می‌شود و به صورت میانگین کاهش صحت در وظایف قبلی است. بر اساس (۴-۳)، فراموشی مدل روی وظیفه‌ی i پس از یادگیری وظیفه‌ی t بدست می‌آید. ماتریس ACC، شامل صحت مدل روی هر وظیفه پس از یادگیری همان وظیفه و وظیفه‌های دیگر است. به دنبال آن، $\mathrm{ACC}_{i,k}$ نشان‌دهنده‌ی صحت مدل روی وظیفه‌ی i پس از یادگیری وظیفه‌ی k است. به این ترتیب، اختلاف بین بیشترین صحتی که وظیفه‌ی i پس از یادگیری وظایف مختلف بدست آورده و صحتی که پس از وظیفه‌ی t (آخرین وظیفه‌ی یادگرفته شده) بدست آمده، در $f_i(t)$ قرار می‌گیرد. در نهایت فراموشی برای هر وظیفه محاسبه شده و میانگین آن‌ها به عنوان فراموشی مدل پس از یادگیری وظیفه‌ی t، در نظر گرفته می‌شود ((۴-۴)).

$$f_i(t) = \max_{k\leq t}\mathrm{ACC}_{i,k} - \mathrm{ACC}_{i,t} \quad (۳-۴)$$

$$F(t) = \frac{1}{t}\sum_{i=1}^{t} f_i(t) \quad (۴-۴)$$

شکل ۴-۱: نمونه‌ای از مجموعه‌داده‌ی UCF101[۲۱].

۴-۳ مجموعه‌داده

در این تحقیق از دو مجموعه‌داده‌ی مرسوم تشخیص حرکت با نام‌های UCF101[۲۱] و HMDB51 [۲۲] استفاده شده است که در ادامه هریک شرح داده خواهد شد.

۴-۳-۱ مجموعه‌داده‌ی UCF101

مجموعه‌داده‌ی UCF101 [۲۱] یکی از مجموعه‌داده‌های استاندارد و پرکاربرد در زمینه تشخیص حرکات انسانی در ویدیو است. این مجموعه شامل 13,320 ویدیو از 101 دسته‌ی مختلف فعالیت انسانی است که طیف وسیعی از حرکات روزمره، ورزشی و تعاملی را پوشش می‌دهد. نمونه‌های این مجموعه‌داده از ویدیوهای واقعی و متنوع جمع‌آوری شده‌اند و شامل تغییرات قابل توجه در شرایط نور، پس‌زمینه، زاویه دید و ظاهر اشخاص هستند. دسته‌های موجود در UCF101 به پنج گروه کلی تقسیم می‌شوند:

۱. **فعالیت‌های ورزشی**

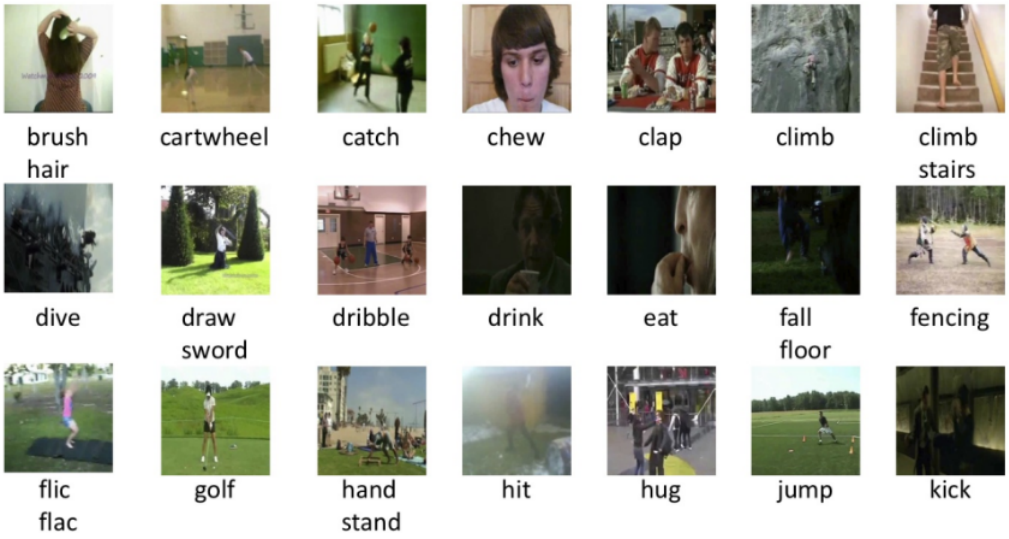
۲. **فعالیت‌های تعاملی انسان با اشیا**

۳. **فعالیت‌های تعاملی انسان با انسان**

۴. **فعالیت‌های بدنی عمومی**

۵. **فعالیت‌های نواختن موسیقی**

ویژگی مهم این مجموعه‌داده، تنوع بالای آن در شرایط تصویربرداری و پیچیدگی حرکات است که آن را به یک معیار معتبر برای ارزیابی مدل‌های تشخیص حرکات‌های انسانی و یادگیری پیوسته تبدیل می‌کند. نمونه‌هایی از داده‌های این مجموعه‌داده در شکل ۴-۱ قابل مشاهده است.



شکل ۴-۲: نمونه‌ای از مجموعه‌داده‌ی HMDB51[۲۲].

۴-۳-۲ مجموعه‌داده‌ی HMDB51

مجموعه‌داده HMDB51 [۲۲] یکی از مجموعه‌داده‌های پرکاربرد در زمینه‌ی تشخیص حرکات انسانی در ویدیو است که برای ارزیابی عملکرد الگوریتم‌های تشخیص حرکت طراحی شده است. این مجموعه شامل 6,766 ویدیو در 51 دسته‌ی فعالیت انسانی است که هر دسته تقریباً ۱۰۰ نمونه دارد. نمونه‌های موجود در HMDB51 از منابع متنوعی همچون فیلم‌های سینمایی، ویدیوهای اینترنتی و فیلم‌های خانگی جمع‌آوری شده‌اند و طیف وسیعی از حرکات انسانی شامل فعالیت‌های بدنی، تعامل انسان با اشیاء و تعامل انسان با انسان را پوشش می‌دهند. یکی از ویژگی‌های مهم این مجموعه‌داده، تنوع بالا در صحنه، پس‌زمینه، زاویه دوربین و کیفیت ویدیوها است که تشخیص حرکات را به یک چالش واقعی تبدیل می‌کند. به دلیل اندازه متوسط و تنوع مناسب، HMDB51 به‌طور گسترده برای آموزش و ارزیابی مدل‌های تشخیص حرکات و یادگیری پیوسته مورد استفاده قرار می‌گیرد. فعالیت‌های موجود در HMDB51 را می‌توان در پنج گروه کلی دسته‌بندی کرد:

۱. **فعالیت‌های عمومی صورت**

۲. **فعالیت‌های صورت همراه با تعامل با اشیا**

۳. **حرکات عمومی بدن**

۴. **حرکات بدن همراه با تعامل با اشیا**

۵. **حرکات بدن در تعامل انسان با انسان**

نمونه‌ای از این داده‌ها در شکل ۴-۲ قابل مشاهده است.

۴-۴ تنظیمات آزمایش

آزمایش‌ها بر روی دو مجموعه‌داده‌ی HMDB51[۲۲] و UCF101[۲۱] به صورت جداگانه اجرا شده و در هر دو حالت، مدل پایه‌ی مورد استفاده، Open-VCLIP [۲۰] می‌باشد که در فرآیند آموزش از وزن‌های پیش‌آموخته‌ی CLIP [۱۳] (نسخه‌ی ViT-B/16) بهره برده است. آموزش با نرخ اولیه‌ی یادگیری 3.33×10^{-6} (مقدار معین شده در تنظیمات Open-VCLIP) تست گردید اما به علت یادگیری آهسته و دریافت نتیجه‌ی نامطلوب، نرخ اولیه‌ی یادگیری، بعد از آزمایش‌های مختلف، با مقدار 1.5×10^{-3} برای مجموعه‌داده‌ی UCF101 و 1.5×10^{-2} برای مجموعه‌داده‌ی HMDB51 مورد استفاده قرار گرفت. مجموعه‌داده‌ها به نسبت‌های 60% برای آموزش، 30% برای آزمون و 10% برای اعتبارسنجی تقسیم شدند. تعداد وظیفه برای هر اجرا، 10 در نظر گرفته شد و تعداد دسته در هر وظیفه، برابر با تعداد کل دسته‌های مجموعه‌داده تقسیم بر تعداد کل وظایف خواهد بود. تنظیمات مربوط به پرامپت، شامل طول هر پرامپت، مقدار بیشینه تعداد پرامپت انتخابی در هر اجرا، به ترتیب ۵ (طبق مقداردهی L2P [۷])و ۲ در نظر گرفته شده است. طول استخر پرامپت متناسب با سناریوی استفاده شده، متفاوت خواهد بود. سایر تنظیمات، مانند تنظیمات معرفی شده در Open-VCLIP، در نظر گرفته شده‌اند. در ادامه تنظیمات اجرا برای هر دو مجموعه‌داده‌ی مذکور بررسی خواهند شد. سخت افزار استفاده شده در این تحقیق، GPU L40S با 48 گیگابایت حافظه‌ی رم می‌باشد.

۴-۴-۱ آزمایش با مجموعه‌داده‌ی UCF101

در این آزمایش ابتدا تعداد 5 ایپاک برای هر وظیفه در نظر گرفته شد و پس از بررسی نمودار صحت بر حسب ایپاک طی شده، مشاهده شد که صحت وظایف پس از سه ایپاک چه در اعتبارسنجی و چه در آموزش ثابت می‌شوند. بنابراین ایپاک نهایی برای این مجموعه‌داده، 3 در نظر گرفته شد. آزمایش این مجموعه‌داده در سه سناریو بررسی شد که در ادامه توضیح داده خواهد شد:

۱. **استخر ثابت با جریمه‌ی وزن‌های پیشین:** دراین حالت طول استخر پرامپت ثابت خواهد بود که در این آزمایش، به اندازه‌ی 202 مقداردهی شده است. تعداد تکرار انتخاب هر پرامپت در

یادگیری پیوسته، ذخیره می‌شود و هنگام انتخاب پرامپت در وظیفه‌ی جدید، متناسب با تعداد تکرار هر پرامپت، جریمه‌ای در نظر گرفته می‌شود. بنابراین احتمال انتخاب پرامپت‌های پیشین، تا حدودی کاهش می‌یابد(الگو گرفته از روش اضافه‌ی ذکر شده در L2P [۷]).

۲. **استخر پویا با مقداردهی تصادفی:** در این حالت طول استخر پرامپت اولیه به اندازه‌ی 20 در نظر گرفته شده است. هر وظیفه که اضافه می‌شود، به طول استخر 20 واحد اضافه می‌شود. نکته‌ی حائز اهمیت در سناریوی فعلی و بعدی، این است که پرامپت‌های استفاده شده در وظایف قبلی، هنگام یادگیری وظیفه‌ی جدید، منجمد شده و صرفا قابلیت انتخاب شدن، دارند و تغییری نخواهند کرد.

۳. **استخر پویا با مقداردهی مبتنی بر کدگذار متن CLIP:** در این حالت، تمام تنظیمات مدل مشابه سناریوی قبلی است و تنها تفاوت، در نحوه‌ی مقداردهی اولیه‌ی کلیدهای استخر پرامپت است. در ابتدا، به ازای هر دسته در هر وظیفه، به تعداد از پیش‌تعیین‌شده کلید، مقداردهی اولیه می‌شود؛ به این صورت که برچسب هر دسته به همراه عبارت‌های معنادار ثابت، از کدگذار متن مدل CLIP عبور کرده و ویژگی‌های استخراج‌شده در کلیدها قرار می‌گیرند. به عنوان مثال اگر برچسب یک ویدیو، jumping person باشد، عبارتی که وارد کدگذار می‌شود برابر با `"a video of jumping person."` خواهد بود. در این سناریو، مقادیر اولیه‌ی کلیدها معنادار بوده و نتایج نشان داده اند که تاثیر مثبتی در انتخاب درست پرامپت‌ها داشته‌اند. در آزمایش اولیه، تعداد کلیدهای مقداردهی اولیه برای هر دسته از هر وظیفه، برابر با 2 در نظر گرفته شد. در حالت نهایی تست‌شده، برای بهبود انتخاب پرامپت‌ها، این تعداد به 5 کلید برای هر دسته افزایش یافت که این تغییر باعث شد که فضای انتخاب پرامپت‌ها غنی‌تر شود و مدل در انتخاب پرامپت‌های مناسب عملکرد بهتری داشته باشد.

۴-۴-۲ آزمایش با مجموعه‌داده‌ی HMDB51

یش ابتدا تعداد 5 ایپاک برای هر وظیفه در نظر گرفته شد و پس از بررسی نمودار صحت بر حسب ایپاک طی شده، مشاهده شد که یادگیری به خوبی صورت نگرفته است. بنابراین پس از افزایش ایپاک‌ها در طی آزمایش، ایپاک نهایی برای این مجموعه‌داده، 7 در نظر گرفته شد. مانند مجموعه‌داده‌ی پیشین، آزمایش این مجموعه‌داده نیز در سه سناریو بررسی شد که در ادامه توضیح داده خواهد شد:

۱. **استخر ثابت با جریمه‌ی وزن‌های پیشین:** این سناریو نیز مانند توضیحات ذکر شده در قسمت قبل، اجرا شده است با این تفاوت که طول استخر پرامپت در این جا 102 در نظر گرفته شده است.

۲. **استخر پویا با مقداردهی تصادفی:** در این حالت طول استخر پرامپت اولیه ابتدا به اندازه‌ی 10 در نظر گرفته شد و سپس به علت کسب نتیجه‌ی بهتر، به 25 تغییر داده شد. هر وظیفه که اضافه می‌شود، به طول استخر 25 واحد اضافه می‌شود. نکته‌ی حائز اهمیت در سناریوی فعلی و بعدی، این است که پرامپت‌های استفاده شده در وظایف قبلی، هنگام یادگیری وظیفه‌ی جدید، منجمد شده و صرفا قابلیت انتخاب شدن، دارند و تغییری نخواهند کرد.

۳. **استخر پویا با مقداردهی مبتنی بر کدگذار متن CLIP:** همانند آن‌چه در قسمت مجموعه‌داده‌ی UCF101 گفته شد، تنظیمات مانند سناریوی پیشین باقی می‌مانند و صرفا مقداردهی اولیه‌ی کلیدها با کدگذار CLIP صورت می‌گیرد. در این جا نیز ابتدا تعداد کلیدهای مقداردهی اولیه برای هر دسته از هر وظیفه، برابر با 2 در نظر گرفته شد و در نهایت، این تعداد به 5 کلید برای هر دسته افزایش یافت. به عبارتی، طول اولیه‌ی استخر پرامپت از 10 به 25 تغییر داده شد.

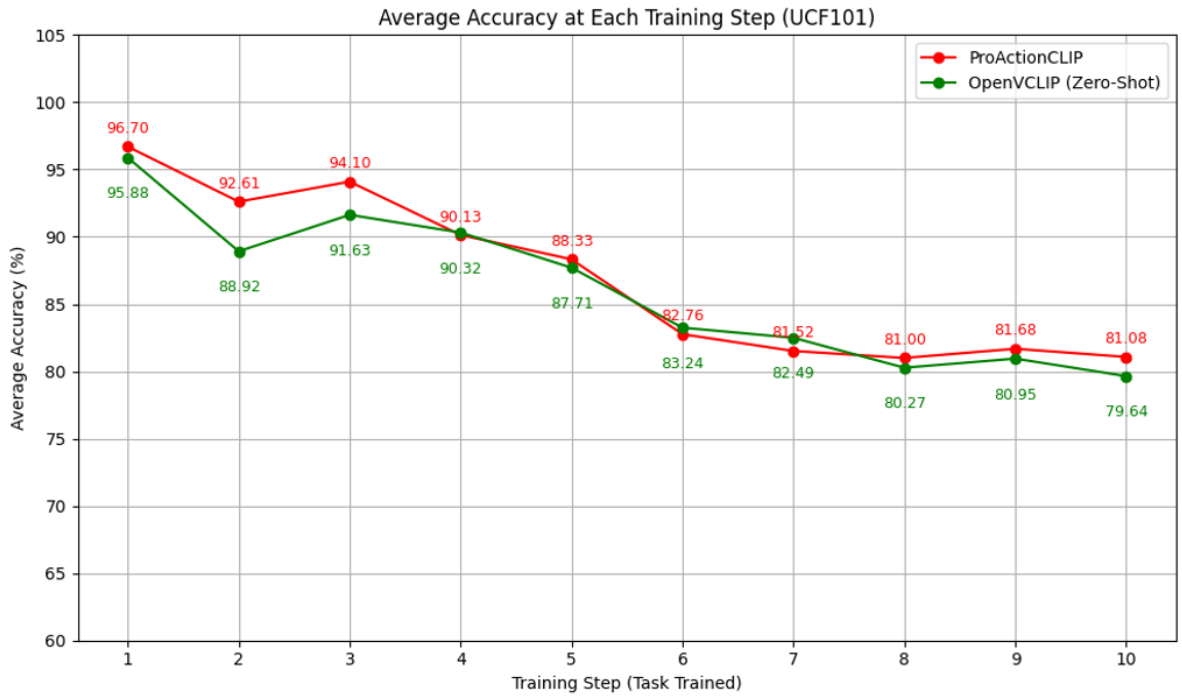
۴-۵ ارزیابی نتایج

به‌منظور ارزیابی عملکرد روش ProActionCLIP با سایر روش‌های مشابه، ابتدا مدل تنها با داده‌های مربوط به وظیفه‌ی اول آموزش داده می‌شود و سپس وظایف بعدی به‌ترتیب به مدل معرفی و آموزش داده می‌شوند. در این فرآیند، مدل باید ضمن یادگیری اطلاعات مربوط به هر وظیفه‌ی جدید، دانش حاصل از وظایف قبلی را نیز حفظ کند. پس از اتمام یادگیری تمامی ده وظیفه، دو معیار میزان فراموشی و میانگین صحت، برای ارزیابی عملکرد مدل، مورد استفاده قرار می‌گیرد. میانگین صحت، نشان‌دهنده‌ی توانایی مدل در حفظ عملکرد بر روی تمام وظایف است و میزان فراموشی، نشان‌دهنده‌ی کاهش صحت مدل، بر روی وظایف اولیه، پس از یادگیری وظایف جدید می‌باشد. این دو معیار برای مقایسه‌ی دقیق‌تر بین روش پیشنهادی و سایر روش‌های موجود به‌کار گرفته شده‌اند. نتایج حاصل از ارزیابی عملکرد روش ProActionCLIP و سایر روش‌های منتخب، در جدول ۴-۱ آورده شده است. در این جدول علاوه بر میانگین صحت و میزان فراموشی، تعداد پارامترهای آموزش‌پذیر مدل و میزان حافظه برای نگه‌داری نمونه‌های ویدیویی وظایف پیشین نیز مقایسه شده است. روش‌های مقایسه‌شده شامل iCaRL[۳۱]، EWC[۲۷]، L2P[۷]، TCD[۶۶]، Pivot[۵۱] و Open-VCLIP[۲۰] می‌باشند. این روش‌ها نماینده‌ی

جدول ۴-۱: مقایسه روش ProActionCLIP با سایر روش‌های یادگیری پیوسته روی مجموعه‌داده UCF101 بر اساس صحت، میزان فراموشی، تعداد پارامترها و حافظه مورد استفاده.

Model	Memory usage. Number of Video Instances	Params	UCF101	
			Accuracy [%]	Forgetting [%]
PIVOT	1010	9M	93.36	4.47
iCaRL	2020	0.5M	80.97	18.11
EWC	None	0.4	9.51	98.94
L2P	None	124K	78.35	5.46
TCD	250	-	74.89	-
Open-VCLIP	None	None	79.64	-
ProActionCLIP (Ours)	None	2M	81.07	9.44

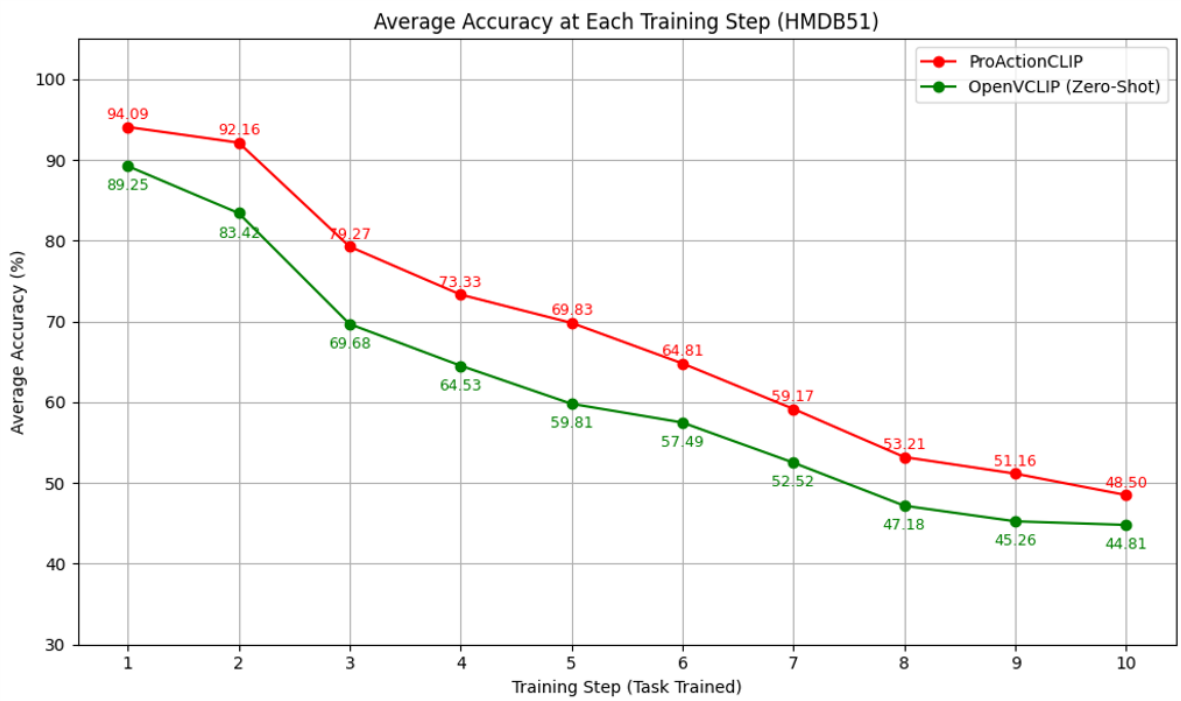
رویکردهای متنوع در حوزه‌ی یادگیری پیوسته هستند. روش‌های iCaRL، EWC و L2P در اصل برای داده‌های تصویری طراحی شده‌اند، اما با استناد به دو مطالعه‌ی پیشین [۶۷، ۶۸] که این روش‌ها را بر روی داده‌های ویدیویی مجموعه‌داده‌ی UCF101 مورد ارزیابی قرار داده‌اند، نتایج مربوط به آن‌ها نیز در جدول ۴-۱ گزارش شده‌اند. در جدول فوق همچنین از مدل Open-VCLIP در حالت منجمد استفاده شده است. در این حالت، وظایف به ترتیب به مدل ارائه شدند تا صحت آن در مواجهه با وظایف جدید اندازه‌گیری شود. برای این مدل در هر مرحله از آزمایش تنها تعداد دسته‌ها نسبت به مرحله قبل افزایش یافته است. در جدول ۴-۱، همان‌طور که مشاهده می‌شود مدل پیشنهادی، از لحاظ میانگین صحت و میزان فراموشی، نتایج بهتری نسبت به سایر روش‌های مشابه داشته است. البته روش PIVOT در مقایسه با روش پیشنهادی، صحت بیشتر و میزان فراموشی کمتری کسب کرده است. علت این امر، استفاده این روش از حافظه برای ذخیره‌ی داده‌های ویدیویی وظایف پیشین بوده است. علاوه بر حافظه مصرفی، تعداد پارامترهای قابل یادگیری این روش به میزان قابل توجهی نسبت به روش پیشنهادی، بیشتر است. انتظار می‌رود استفاده از حافظه و تعداد پارامترهای قابل یادگیری بیشتر، در افزایش صحت و کاهش میزان فراموشی مدل‌ها موثر باشند. این امر در جدول ۴-۱ قابل مشاهده است. چنان‌که مدل Open-VCLIP به صورت منجمد استفاده شده است و از حافظه و پارامتر قابل‌آموزش بهره نبرده است، صحت کمتری نسبت به PIVOT ارائه داده است. با این حال، این مدل علیرغم آنکه برای یادگیری پیوسته طراحی نشده است، نتایج نسبتا خوبی در حالت یادگیری بدون نمونه ارائه داده است. یکی از دلایلی که این روش در معماری ProActionCLIP به‌کار گرفته شده است، همین موضوع می‌باشد.همان‌طور که در جدول ۴-۱ مشاهده می‌شود، مدل ProActionCLIP نسبت به مدل TCD میانگین صحت بالاتری کسب کرده است. لازم به ذکر است که مدل TCD از حافظه استفاده کرده و در گام نخست ۵۰ دسته را به‌صورت یکجا آموزش داده و سپس بقیه را به‌صورت پیوسته در ۱۰ وظیفه به مدل افزوده است. آموزش یکجای ۵۰ دسته می‌تواند احتمال بروز فراموشی فاجعه‌بار را کاهش دهد، با این حال مدل ProActionCLIP عملکرد



شکل ۴-۳: میانگین صحت وظایف در هر گام آموزشی برای دو روش ProActionCLIP و Open-VCLIP برای مجموعه‌داده‌ی UCF101. محور عمودی نشان‌دهنده‌ی میانگین صحت مدل بر وظایف و محور افقی نشان‌دهنده‌ی گام‌های آموزش است. در هر گام آموزشی، یک وظیفه اضافه می‌شود. با افزایش تعداد وظایف و در نتیجه افزایش تعداد دسته‌ها، میانگین صحت کاهش می‌یابد. نمودار قرمزرنگ مربوط به مدل پیشنهادی بوده و نمودار سبزرنگ مربوط به مدل Open-VCLIP می‌باشد.

بهتری داشته است. در مدل EWC، از حافظه استفاده نشده است. این مدل برای تصویر آموزش داده شده است و نتایج نامطلوب ارائه شده برای ویدیو، به این علت است که از ویدیو، صرفا تعداد محدودی قاب برای دسته‌بندی انتخاب شده است و از مدل‌های از پیش آموزش دیده مانند CLIP نیز استفاده نکرده است. در مدل iCaRL، استفاده از حافظه، باعث افزایش صحت شده است اما میزان فراموشی هم‌چنان نسبت به مدل پیشنهادی، بالاتر است. در روش L2P با وجود استفاده نکردن از حافظه و آموزش پارامترهای قابل یادگیری کم، توانسته است نتایج خوبی ارائه دهد. این ویژگی‌های مثبت در روش Open-VCLIP و L2P موجب ارائه‌ی روش ProActionCLIP گردید که توانسته است نسبت به این دو روش، نتایج بهتری را ارائه دهد.

به منظور ارائه‌ی مقایسه‌ای دقیق‌تر بین روش پیشنهادی و Open-VCLIP، میانگین صحت وظایف در هر گام آموزشی محاسبه گردید که نتایج آن برای مجموعه‌داده‌های UCF101 و HMDB51 به ترتیب در نمودارهای شکل ۴-۳ و شکل ۴-۴ آورده شده است. همان‌طور که مشاهده می‌شود، در این نمودارها با افزایش تعداد وظایف، میانگین صحت روندی نزولی داشته است. یکی از دلایل این امر آن است که



شکل ۴-۴: میانگین صحت وظایف در هر گام آموزشی برای دو روش ProActionCLIP و Open-VCLIP برای مجموعه‌داده‌ی HMDB51 . محور عمودی نشان‌دهنده‌ی میانگین صحت مدل بر وظایف و محور افقی نشان‌دهنده‌ی گام‌های آموزش است. در هر گام آموزشی، یک وظیفه اضافه می‌شود. با افزایش تعداد وظایف و در نتیجه افزایش تعداد دسته‌ها، میانگین صحت کاهش می‌یابد. نمودار قرمزرنگ مربوط به مدل پیشنهادی بوده و نمودار سبزرنگ مربوط به مدل Open-VCLIP می‌باشد.

با اضافه شدن هر وظیفه‌ی جدید، تعداد دسته‌ها افزایش یافته و ویژگی استخراج‌شده از ویدیو باید با تعداد بیشتری دسته مقایسه شود. این موضوع احتمال بروز خطا در تصمیم‌گیری و کاهش صحت را بالا می‌برد. علاوه بر این، شباهت معنایی برخی برچسب‌ها نیز می‌تواند عامل دیگری برای افت صحت باشد. برای مثال، در مجموعه‌داده‌ی UCF101 دسته‌هایی مانند Playing Guitar، Playing Piano، Playing Tabla، Playing Violin، Playing Cello، Playing Daf، Playing Dhol، Playing Flute و Playing Sitar همگی عبارت مشترک "Playing" را در برچسب خود دارند. این شباهت باعث می‌شود که در فرایند کدگذاری متن توسط مدل CLIP، نمایش‌های برداری این برچسب‌ها به یکدیگر نزدیک شوند و در نتیجه، کلیدهای پرامپت این دسته‌ها نیز، شباهت بالایی به یکدیگر پیدا کنند. در چنین شرایطی، ممکن است مدل برای یک ویدیوی مرتبط با یکی از این دسته‌ها، به اشتباه پرامپت دسته‌ی مشابه دیگری را انتخاب کند. با این حال، نتایج نشان می‌دهد که مدل پیشنهادی در هر دو مجموعه‌داده عملکرد بهتری در کاهش این افت صحت داشته است.

۴-۵-۱ مطالعات فرسایشی

به‌منظور بررسی تأثیر مؤلفه‌های مختلف مدل پیشنهادی ProActionCLIP، یک مطالعه‌ی فرسایشی[1] انجام گرفته است. مدل ProActionCLIP در این پژوهش به‌عنوان ترکیبی از دو روش Open-VCLIP و L2P طراحی شده و در دو مجموعه‌داده‌ی معتبر HMDB51 و UCF101 مورد ارزیابی قرار گرفته است. برای تحلیل دقیق‌تر، مطابق با آن‌چه در قسمت تنظیمات آزمایش ذکر شد، عملکرد مدل در سه پیکربندی متفاوت از نظر ساختار و رفتار استخر پرامپت‌ها مورد آزمایش قرار گرفت. این پیکربندی‌ها عبارتند از:

۱. **استخر ثابت با جریمه:** طول استخر ثابت نگه داشته شده و با استفاده از سازوکار جریمه، احتمال استفاده‌ی مجدد از پرامپت‌های پرتکرار کاهش می‌یابد.

۲. **استخر پویا با مقداردهی تصادفی:** با ورود هر وظیفه، پرامپت‌های جدید به‌صورت تصادفی اضافه شده و پرامپت‌های قبلی منجمد می‌شوند.

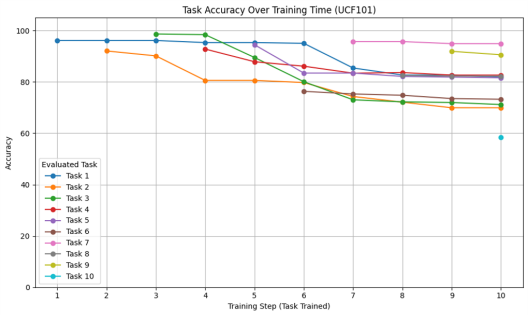
۳. **استخر پویا با مقداردهی معنایی:** مشابه سناریوی قبل، اما مقداردهی اولیه‌ی پرامپت‌ها با استفاده از خروجی کدگذار متن CLIP انجام می‌شود.

[1] Ablation study

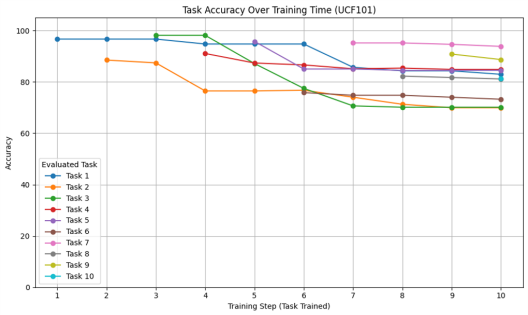
نتایج حاصل از این آزمایش‌ها برای مجموعه‌داده‌های UCF101 و HMDB51 به ترتیب در شکل ۴-۵ و شکل ۴-۶ آورده شده است که در ادامه هر یک بررسی خواهد شد. مطابق آنچه در شکل ۴-۵ مشاهده می‌شود، نتایج روی مجموعه‌داده‌ی UCF101 نشان می‌دهد که استفاده از مقداردهی اولیه با کدگذار CLIP موجب حفظ بهتر صحت در طول گام‌های آموزشی می‌شود و این اثر با افزایش تعداد پرامپت‌های هر دسته از دو به پنج، تقویت می‌گردد. مقایسه‌ی شکل ۴-۵آ و شکل ۴-۵ب نشان می‌دهد که افزایش تعداد پرامپت‌ها باعث بهبود عملکرد، به‌ویژه در گام‌های پایانی آموزش و کاهش نرخ افت صحت یا همان فراموشی شده است. در مقابل، در دو پیکربندی با مقداردهی اولیه تصادفی، عملکرد کلی پایین‌تر بوده و افت صحت در طول مراحل آموزش محسوس‌تر است. همچنین مقایسه‌ی شکل ۴-۵ج و شکل ۴-۵د نشان می‌دهد که استفاده از استخر پرامپت پویا در مقداردهی تصادفی نسبت به استخر ثابت، موجب بهبود نسبی حفظ صحت می‌شود. این نتایج بیانگر آن است که هم نوع مقداردهی اولیه و هم طراحی استخر پرامپت نقش کلیدی در کاهش فراموشی و بهبود پایداری مدل در یادگیری وظایف پیوسته دارند. با توجه به اینکه نتایج سناریوی آخر در مجموعه‌داده‌ی UCF101 عملکرد بهتری داشت، همان سناریو برای مجموعه‌داده‌ی HMDB51 نیز مورد ارزیابی قرار گرفت و نتایج آن در شکل ۴-۶ ارائه شده است. همان‌طور که مشاهده می‌شود، هر دو پیکربندی از مقداردهی اولیه با کدگذار CLIP استفاده می‌کنند و تفاوت اصلی آن‌ها در تعداد پرامپت‌های هر دسته است. مقایسه‌ی شکل ۴-۶آ و شکل ۴-۶ب نشان می‌دهد که افزایش تعداد پرامپت‌ها از دو به پنج موجب بهبود عملکرد مدل، به‌ویژه در مراحل پایانی آموزش، و کاهش نرخ افت صحت یا همان فراموشی شده است. این نتایج نشان می‌دهد که حتی در شرایطی که نوع مقداردهی اولیه ثابت باشد، افزایش تعداد پرامپت‌های هر دسته می‌تواند نقش مؤثری در حفظ دانش و پایداری مدل در طول یادگیری وظایف پیوسته ایفا کند.

۴-۶ جمع‌بندی

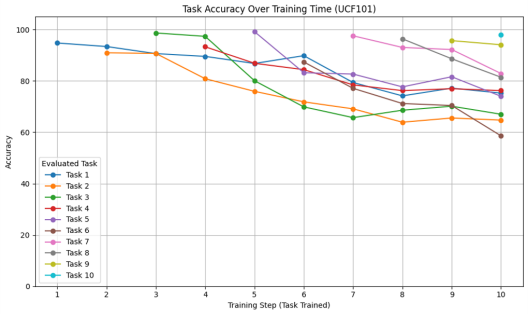
نتایج این فصل نشان می‌دهد که مدل ProActionCLIP در یادگیری پیوسته‌ی تشخیص حرکت انسان، با دستیابی به میانگین صحت بالا، کاهش محسوس میزان فراموشی وظایف پیشین و بهره‌وری بالای محاسباتی، عملکردی برتری نسبت به روش‌های مرجع مانند PIVOT دارد. همچنین آزمایش‌های تکمیلی نشان دادند که استفاده از مقداردهی اولیه با CLIP و استخر پرامپت پویا، به‌ویژه با تعداد پرامپت‌های بیشتر، نقش مهمی در بهبود عملکرد مدل پیشنهادی ایفا می‌کند.



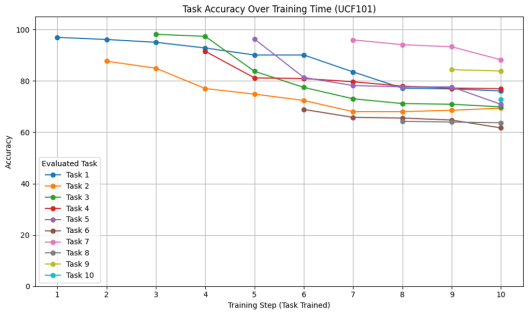
(ب) استخر پویا – 2/class – CLIP init



(آ) استخر پویا – 5/class – CLIP init

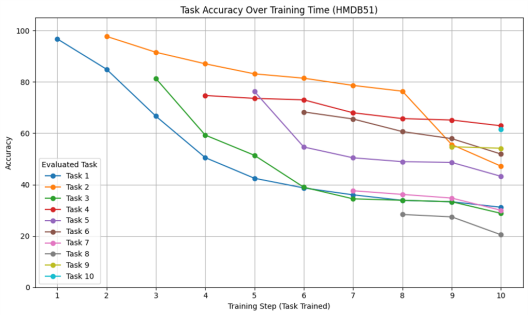


(د) استخر ثابت – Random init

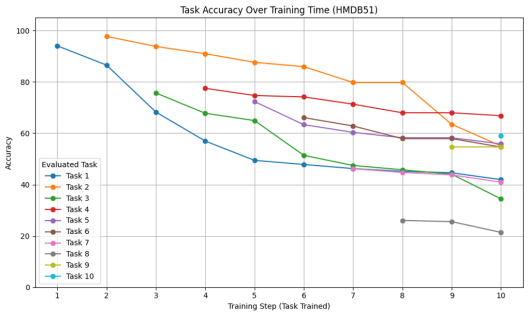


(ج) استخر پویا – 2/class – Random init

شکل ۴-۵: تغییرات صحت هر وظیفه در مراحل مختلف آموزش برای چهار پیکربندی متفاوت. صحت هر وظیفه به صورت جداگانه در هر نمودار، از زمان اضافه شدن آن به مدل، نشان داده شده است. در نمودار (الف) از مقداردهی اولیه با کدگذار CLIP استفاده شده و هر دسته شامل ۵ پرامپت است و استخر پرامپت به صورت پویا به‌روزرسانی می‌شود. نمودار (ب) نیز از مقداردهی اولیه با کدگذار CLIP بهره می‌برد اما هر دسته ۲ پرامپت دارد و استخر پرامپت همچنان پویاست. در نمودار (ج) مقداردهی اولیه به صورت تصادفی انجام شده و به ازای هر دسته ۲ پرامپت به استخر پرامپت اضافه می‌شود. نمودار (د) نشان‌دهنده‌ی حالتی است که مقداردهی اولیه تصادفی بوده و از یک استخر پرامپت ثابت با اندازه ۱۰۲ استفاده شده است.



(ب) استخر پویا – 2/class – CLIP init



(آ) استخر پویا – 5/class – CLIP init

شکل ۴-۶: تغییرات صحت هر وظیفه در مراحل مختلف آموزش برای دو پیکربندی متفاوت. صحت هر وظیفه به صورت جداگانه در هر نمودار، از زمان اضافه شدن آن به مدل، نشان داده شده است. در نمودار (الف) از مقداردهی اولیه با کدگذار CLIP استفاده شده و هر دسته شامل ۵ پرامپت است و استخر پرامپت به صورت پویا به‌روزرسانی می‌شود. نمودار (ب) نیز از مقداردهی اولیه با کدگذار CLIP بهره می‌برد اما هر دسته ۲ پرامپت دارد و استخر پرامپت همچنان پویاست.

فصل پنجم

جمع‌بندی و نتیجه‌گیری و پیشنهادات

در این پایان نامه به معرفی مدل پیشنهادی ProActionCLIP برای یادگیری پیوسته در داده‌های ویدیویی پرداخته شد. این روش با ترکیب توانایی‌های مدل Open-VCLIP در استخراج ویژگی‌های ویدیو و سازوکار پرامپت‌های یادگیرنده در L2P، توانسته است بدون تغییر مستقیم پارامترهای مدل پایه‌ی Open-VCLIP، به یادگیری پیوسته‌ی وظایف در حوزه‌ی ویدیو بپردازد. به عبارت دیگر، ProActionCLIP با بهره‌گیری از مزیت‌های هر دو مدل مرجع، راهکاری مؤثر در یادگیری پیوسته، برای تشخیص حرکت انسان در حوزه‌ی ویدیو ارائه می‌دهد. نتایج حاصل نشان داد که این مدل، ضمن حفظ دانش پیشین و کاهش قابل‌توجه فراموشی فاجعه‌بار، از لحاظ مصرف حافظه و منابع محاسباتی عملکرد بهتری نسبت به سایر روش‌های مشابه داشته است.

۵-۱ پیشنهادات

با توجه به نتایج و تحلیل‌های ارائه‌شده، چند مسیر پژوهشی برای بهبود روش پیشنهادی قابل بررسی است:

۱. **کاهش شباهت کلیدها بین کلاس‌های مشابه:** یکی از چالش‌های اصلی در افزایش تعداد وظایف، شباهت میان برچسب‌ها است که منجر به شباهت کلیدهای متناظر آن‌ها می‌شود. این امر می‌تواند باعث انتخاب نادرست پرامپت‌ها گردد. پیشنهادی که می‌توان مطرح کرد، در نظر گرفتن یک مؤلفه در تابع زیان است که فاصله‌ی کلیدهای مربوط به برچسب‌های متفاوت را افزایش دهد. به این ترتیب، احتمال شباهت کلیدها بین کلاس‌های نزدیک کاهش یافته و دقت انتخاب پرامپت‌ها بهبود می‌یابد.

۲. **افزایش تعداد پرامپت‌های اختصاصی برای هر کلاس:** اختصاص تعداد بیشتری پرامپت به هر کلاس می‌تواند انعطاف‌پذیری مدل را در یادگیری ویژگی‌های متنوع آن کلاس افزایش دهد و عملکرد کلی سیستم را بهبود بخشد.

منابع و مراجع

[1] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Ł ukasz, and Polosukhin, Illia. Attention is all you need. in Guyon, I., Luxburg, U. Von, Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds. , Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., 2017.

[2] He, Kaiming, Gkioxari, Georgia, Dollár, Piotr, and Girshick, Ross. Mask r-cnn. in 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988, 2017.

[3] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[4] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. in Burstein, Jill, Doran, Christy, and Solorio, Thamar, eds. , Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold,

Georg, Gelly, Sylvain, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[6] Argyriou, Andreas, Evgeniou, Theodoros, and Pontil, Massimiliano. Multi-task feature learning. NIPS'06, p. 41–48, Cambridge, MA, USA, 2006. MIT Press.

[7] Wang, Zifeng, Zhang, Zizhao, Lee, Chen-Yu, Zhang, Han, Sun, Ruoxi, Ren, Xiaoqi, Su, Guolong, Perot, Vincent, Dy, Jennifer, and Pfister, Tomas. Learning to prompt for continual learning. in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 139–149, 2022.

[8] Wang, Liyuan, Zhang, Xingxing, Su, Hang, and Zhu, Jun. A comprehensive survey of continual learning: Theory, method and application. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[9] Mai, Zheda, Li, Ruiwen, Jeong, Jihwan, Quispe, David, Kim, Hyunwoo, and Sanner, Scott. Online continual learning in image classification: An empirical survey. Neurocomputing, 469:28–51, 2022.

[10] French, Robert M. Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 3(4):128–135, 1999.

[11] Chen, Wuyang, Zhou, Yanqi, Du, Nan, Huang, Yanping, Laudon, James, Chen, Zhifeng, and Cui, Claire. Lifelong language pretraining with distribution-specialized experts. in International Conference on Machine Learning, pp. 5383–5395. PMLR, 2023.

[12] Achiam, Josh, Adler, Steven, Agarwal, Sandhini, Ahmad, Lama, Akkaya, Ilge, Aleman, Florencia Leoni, Almeida, Diogo, Altenschmidt, Janko, Altman, Sam, Anadkat, Shyamal, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[13] Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, Krueger,

Gretchen, and Sutskever, Ilya. Learning transferable visual models from natural language supervision. in International Conference on Machine Learning, 2021.

[14] Zhang, Jingyi, Huang, Jiaxing, Jin, Sheng, and Lu, Shijian. Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8):5625–5644, 2024.

[15] Zheng, Junhao, Qiu, Shengjie, Shi, Chengming, and Ma, Qianli. Towards lifelong learning of large language models: A survey. ACM Comput. Surv., 57(8), March 2025.

[16] Wang, Qiang, Du, Junlong, Yan, Ke, and Ding, Shouhong. Seeing in flowing: Adapting clip for action recognition with motion prompts learning. in Proceedings of the 31st ACM International Conference on Multimedia, MM '23, p. 5339–5347, New York, NY, USA, 2023. Association for Computing Machinery.

[17] Zhou, Kaiyang, Yang, Jingkang, Loy, Chen Change, and Liu, Ziwei. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9):2337–2348, July 2022.

[18] Yu, Jiazuo, Zhuge, Yunzhi, Zhang, Lu, Hu, Ping, Wang, Dong, Lu, Huchuan, and He, You. Boosting continual learning of vision-language models via mixture-of-experts adapters. pp. 23219–23230, 06 2024.

[19] Lu, Shuyun, Jiao, Jian, Wang, Lanxiao, Qiu, Heqian, Lin, Xingtao, Mei, Hefei, and Li, Hongliang. Video class-incremental learning with clip based transformer. in 2024 IEEE International Conference on Image Processing (ICIP), pp. 500–506, 2024.

[20] Weng, Zejia, Yang, Xitong, Li, Ang, Wu, Zuxuan, and Jiang, Yu-Gang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. in ICML, 2023.

[21] Soomro, Khurram, Zamir, Amir Roshan, and Shah, Mubarak. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[22] Kuehne, Hildegard, Jhuang, Hueihan, Garrote, Estíbaliz, Poggio, Tomaso, and Serre, Thomas. Hmdb: a large video database for human motion recognition. in 2011 International conference on computer vision, pp. 2556–2563. IEEE, 2011.

[23] van de Ven, Gido M. and Tolias, Andreas S. Three scenarios for continual learning, 2019.

[24] Churamani, Nikhil, Kara, Ozgur, and Gunes, Hatice. Domain-Incremental Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition . IEEE Transactions on Affective Computing, 14(04):3191–3206, October 2023.

[25] Ma, Jiawei, Tao, Xiaoyu, Ma, Jianxing, Hong, Xiaopeng, and Gong, Yihong. Class incremental learning for video action classification. in 2021 IEEE International Conference on Image Processing (ICIP), pp. 504–508, 2021.

[26] Park, Jaeyoo, Kang, Minsoo, and Han, Bohyung. Class-incremental learning for action recognition in videos. in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13678–13687, 2021.

[27] Kirkpatrick, James, Pascanu, Razvan, Rabinowitz, Neil, Veness, Joel, Desjardins, Guillaume, Rusu, Andrei A., Milan, Kieran, Quan, John, Ramalho, Tiago, Grabska-Barwinska, Agnieszka, Hassabis, Demis, Clopath, Claudia, Kumaran, Dharshan, and Hadsell, Raia. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13):3521–3526, 2017.

[28] Li, Zhizhong and Hoiem, Derek. Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40:2935–2947, 2016.

[29] Shin, Hanul, Lee, Jung Kwon, Kim, Jaehong, and Kim, Jiwon. Continual learning with deep generative replay. in Guyon, I., Luxburg, U. Von, Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds. , Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., 2017.

[30] Chaudhry, Arslan, Ranzato, Marc'Aurelio, Rohrbach, Marcus, and Elhoseiny, Mohamed. Efficient lifelong learning with a-gem. ArXiv, abs/1812.00420, 2018.

[31] Rebuffi, Sylvestre-Alvise, Kolesnikov, Alexander, Sperl, Georg, and Lampert, Christoph H. icarl: Incremental classifier and representation learning. in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5533–5542, 2017.

[32] Aljundi, Rahaf, Caccia, Lucas, Belilovsky, Eugene, Caccia, Massimo, Lin, Min, Charlin, Laurent, and Tuytelaars, Tinne. Online continual learning with maximally interfered retrieval. ArXiv, abs/1908.04742, 2019.

[33] Minhas, Rashid, Mohammed, Abdul Adeel, and Wu, Q. M. Jonathan. Incremental learning in human action recognition based on snippets. IEEE Transactions on Circuits and Systems for Video Technology, 22(11):1529–1541, 2012.

[34] Li, Tianjiao, Ke, Qiuhong, Rahmani, Hossein, Ho, Rui En, Ding, Henghui, and Liu, Jun. Else-net: Elastic semantic network for continual action recognition from skeleton data. in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13414–13423, 2021.

[35] Cheng, Jian, Liu, Haijun, Wang, Feng, Li, Hongsheng, and Zhu, Ce. Silhouette analysis for human action recognition based on supervised temporal t-sne and incremental learning. IEEE Transactions on Image Processing, 24(10):3203–3217, 2015.

[36] Parisi, German I., Tani, Jun, Weber, Cornelius, and Wermter, Stefan. Lifelong learning of human actions with deep neural network self-organization. Neural Networks, 96:137–149, 2017.

[37] Singh, Amanpreet, Hu, Ronghang, Goswami, Vedanuj, Couairon, Guillaume, Galuba, Wojciech, Rohrbach, Marcus, and Kiela, Douwe. Flava: A foundational language and vision alignment model. in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15617–15629, 2022.

[38] Li, Liunian Harold, Zhang, Pengchuan, Zhang, Haotian, Yang, Jianwei, Li, Chunyuan, Zhong, Yiwu, Wang, Lijuan, Yuan, Lu, Zhang, Lei, Hwang, Jenq-Neng, Chang, Kai-Wei, and Gao, Jianfeng. Grounded language-image pre-training. in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10955–10965, 2022.

[39] Wang, Mengmeng, Xing, Jiazheng, and Liu, Yong. Actionclip: A new paradigm for video action recognition. ArXiv, abs/2109.08472, 2021.

[40] Gao, Peng, Geng, Shijie, Zhang, Renrui, Ma, Teli, Fang, Rongyao, Zhang, Yongfeng, Li, Hongsheng, and Qiao, Yu. Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision, 132(2):581–595, Feb 2024.

[41] Wortsman, Mitchell, Ilharco, Gabriel, Kim, Jong Wook, Li, Mike, Kornblith, Simon, Roelofs, Rebecca, Lopes, Raphael Gontijo, Hajishirzi, Hannaneh, Farhadi, Ali, Namkoong, Hongseok, and Schmidt, Ludwig. Robust fine-tuning of zero-shot models. in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7949–7961, 2022.

[42] Gu, Xiuye, Lin, Tsung-Yi, Kuo, Weicheng, and Cui, Yin. Open-vocabulary detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921, 2021.

[43] Lüddecke, Timo and Ecker, Alexander. Image segmentation using text and image prompts. in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7076–7086, 2022.

[44] Garg, Saurabh, Farajtabar, Mehrdad, Pouransari, Hadi, Vemulapalli, Raviteja, Mehta, Sachin, Tuzel, Oncel, Shankar, Vaishaal, and Faghri, Fartash. Tic-clip: Continual training of clip models. in The Twelfth International Conference on Learning Representations (ICLR), 2024.

[45] Menabue, Martin, Frascaroli, Emanuele, Boschini, Matteo, Sangineto, Enver, Bonicelli, Lorenzo, Porrello, Angelo, and Calderara, Simone. Semantic residual prompts for continual learning. in Leonardis, Aleš, Ricci, Elisa, Roth, Stefan, Russakovsky, Olga, Sattler, Torsten, and Varol, Gül, eds. , Computer Vision – ECCV 2024, pp. 1–18, Cham, 2025. Springer Nature Switzerland.

[46] Wang, Zifeng, Zhang, Zizhao, Ebrahimi, Sayna, Sun, Ruoxi, Zhang, Han, Lee, Chen-Yu, Ren, Xiaoqi, Su, Guolong, Perot, Vincent, Dy, Jennifer, and Pfister, Tomas. Dualprompt: Complementary prompting for rehearsal-free continual learning. in Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI, p. 631–648, Berlin, Heidelberg, 2022. Springer-Verlag.

[47] Zuo, Yukun, Yao, Hantao, Yu, Lu, Zhuang, Liansheng, and Xu, Changsheng. Hierarchical prompts for rehearsal-free continual learning. ArXiv, abs/2401.11544, 2024.

[48] Jung, Dahuin, Han, Dongyoon, Bang, Jihwan, and Song, Hwanjun. Generating instance-level prompts for rehearsal-free continual learning. in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11813–11823, 2023.

[49] Huang, Wei-Cheng, Chen, Chun-Fu, and Hsu, Hsiang. OVOR: Oneprompt with virtual outlier regularization for rehearsal-free class-incremental learning. in The Twelfth International Conference on Learning Representations, 2024.

[50] Tang, Longxiang, Tian, Zhuotao, Li, Kai, He, Chunming, Zhou, Hantao, Zhao, Hengshuang, Li, Xiu, and Jia, Jiaya. Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models. in Leonardis, Aleš, Ricci, Elisa, Roth, Stefan, Russakovsky, Olga, Sattler, Torsten, and Varol, Gül, eds. , Computer Vision – ECCV 2024, pp. 346–365, Cham, 2024. Springer Nature Switzerland.

[51] Villa, Andrés, Alcázar, Juan León, Alfarra, Motasem, Alhamoud, Kumail, Hurtado, Julio, Heilbron, Fabian Caba, Soto, Alvaro, and Ghanem, Bernard. Pivot: Prompting for video continual learning. in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24214–24223, 2023.

[52] Wang, Hao, Liu, Fang, Jiao, Licheng, Wang, Jiahao, Hao, Zehua, Li, Shuo, Li, Lingling, Chen, Puhua, and Liu, Xu. Vilt-clip: Video and language tuning clip with multimodal prompt learning and scenario-guided optimization. Proceedings of the AAAI Conference on Artificial Intelligence, 38(6):5390–5400, Mar. 2024.

[53] Roy, Anurag, Moulick, Riddhiman, Verma, Vinay, Ghosh, Saptarshi, and Das, Abir. Convolutional prompting meets language models for continual learning. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2024.

[54] Wistuba, Martin, Teja Sivaprasad, Prabhu, Balles, Lukas, and Zappella, Giovanni. Choice of PEFT Technique in Continual Learning: Prompt Tuning is Not All You Need. arXiv e-prints, p. arXiv:2406.03216, June 2024.

[55] Li, Jiashuo, Wang, Shaokun, Qian, Bo, He, Yuhang, Wei, Xing, Wang, Qiang, and Gong, Yihong. Dynamic integration of task-specific adapters for class incremental learning, 2025.

[56] Zhou, Da-Wei, Sun, Hai-Long, Ye, Han-Jia, and Zhan, De-Chuan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. in 2024

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 23554–23564, 2024.

[57] Gao, Xinyuan, Dong, Songlin, He, Yuhang, Wang, Qiang, and Gong, Yihong. Beyond prompt learning: Continual adapter for efficient rehearsal-free continual learning. in Leonardis, Aleš, Ricci, Elisa, Roth, Stefan, Russakovsky, Olga, Sattler, Torsten, and Varol, Gül, eds. , Computer Vision – ECCV 2024, pp. 89–106, Cham, 2025. Springer Nature Switzerland.

[58] Huang, Linlan, Cao, Xusheng, Lu, Haori, and Liu, Xialei. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. in Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LIV, p. 214–231, Berlin, Heidelberg, 2024. Springer-Verlag.

[59] Pan, Junting, Lin, Ziyi, Zhu, Xiatian, Shao, Jing, and Li, Hongsheng. ST-adapter: Parameter-efficient image-to-video transfer learning. in Oh, Alice H., Agarwal, Alekh, Belgrave, Danielle, and Cho, Kyunghyun, eds. , Advances in Neural Information Processing Systems, 2022.

[60] Wang, Huiyi, Lu, Haodong, Yao, Lina, and Gong, Dong. Self-expansion of pre-trained models with mixture of adapters for continual learning. in NeurIPS 2024 Workshop on Scalable Continual Learning for Lifelong Foundation Models, 2024.

[61] Lu, Shuyun, Jiao, Jian, Wang, Lanxiao, Qiu, Heqian, Lin, Xingtao, Mei, Hefei, and Li, Hongliang. Video class-incremental learning with clip based transformer. in 2024 IEEE International Conference on Image Processing (ICIP), pp. 500–506, 2024.

[62] Li, Jiashuo, Wang, Shaokun, Qian, Bo, He, Yuhang, Wei, Xing, Wang, Qiang, and Gong, Yihong. Dynamic integration of task-specific adapters for class incremental learning, 2025.

[63] Wang, Huiyi, Lu, Haodong, Yao, Lina, and Gong, Dong. Self-expansion of pre-trained models with mixture of adapters for continual learning. in Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 10087–10098, 2025.

[64] Ilharco, Gabriel, Wortsman, Mitchell, Gadre, Samir Yitzhak, Song, Shuran, Hajishirzi, Hannaneh, Kornblith, Simon, Farhadi, Ali, and Schmidt, Ludwig. Patching open-vocabulary models by interpolating weights. in Oh, Alice H., Agarwal, Alekh, Belgrave, Danielle, and Cho, Kyunghyun, eds. , Advances in Neural Information Processing Systems, 2022.

[65] Izmailov, Pavel, Podoprikhin, Dmitrii, Garipov, Timur, Vetrov, Dmitry, and Wilson, Andrew. Averaging weights leads to wider optima and better generalization. 03 2018.

[66] Park, Jaeyoo, Kang, Minsoo, and Han, Bohyung. Class-incremental learning for action recognition in videos. in Proceedings of the IEEE/CVF international conference on computer vision, pp. 13698–13707, 2021.

[67] Villa, Andrés, Alhamoud, Kumail, Escorcia, Victor, Caba, Fabian, Alcázar, Juan León, and Ghanem, Bernard. vclimb: A novel video class incremental learning benchmark. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19035–19044, 2022.

[68] Hong, Kiseong, Kim, Gyeong-hyeon, and Kim, Eunwoo. Rainbowprompt: Diversity-enhanced prompt-evolving for continual learning. arXiv preprint arXiv:2507.22553, 2025.

Abstract

In recent years, machine learning—and in particular, deep neural networks—have made remarkable advancements, achieving near-human performance in fields such as computer vision, natural language processing, and speech recognition. Achieving such performance typically requires pre-training these networks on large-scale datasets, a process that enables the reuse of trained models across a variety of tasks. In many real-world applications, however, data becomes available incrementally and in the form of sequential tasks, highlighting the need for continual learning approaches. A major challenge in this domain is catastrophic forgetting, wherein the model suffers a significant drop in performance on previously learned tasks after acquiring new ones. Despite recent developments in large language models and vision-language models, limitations such as high memory consumption remain. This research introduces ProActionCLIP, a method for continual learning on video data that combines the feature extraction capabilities of the vision-language model Open-VCLIP with the learnable prompt mechanism from the L2P model. This integration enables adaptation to sequential tasks without altering the main parameters of the base Open-VCLIP model, while optimizing prompt selection to mitigate catastrophic forgetting. Experimental results demonstrate that the proposed method not only preserves prior knowledge and effectively learns new knowledge, but also offers high efficiency in terms of memory and computational resource usage. Consequently, it provides an effective solution for continual learning in human action recognition within video datasets.

Keywords:

Continual learning, vision-language model, prompt learning, catastrophic forgetting

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Computer engineering

M. Sc. Thesis

Action recognition with continual Learning

By

Parisa Mollahoseini

Supervisor

Dr. Mohammad Rahmati

August & 2025