به نام خدا

درس یادگیری عمیق

تمرین سری پنجم

استاد درس : دکتر محمدرضا محمدی

دستیاران : فاطمه ستوده، نفیسه احمدی،

محمدمصطفی رستم خانی، بهداد نادری فرد

دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر

نیمسال دوم تحصیلی ۱۴۰۳ - ۱۴۰۴

مهلت تحویل : ۱۴۰۴/۰۳/۰۶

لطفا به نکات موجود در سند قوانین انجام و تحویل تمرین ها دقت فرمایید.

سوالات تئوری

۱.  مقاله Composing Parameter-Efficient Modules with Arithmetic Operations را مطالعه کنید و به سوالات زیر پاسخ دهید(۱۵ نمره):

(آ) با توجه به مقاله، هدف اصلی استفاده از روش‌های PEFT در مدل‌های زبانی چیست؟ همچنین دو روش PEFT بررسی‌شده در این مقاله کدام‌اند و ویژگی‌های متمایز آن‌ها نسبت به روش‌های سنتی finetuning چیست؟

هدف اصلی استفاده از روش‌های PEFT، تطبیق LLM های از پیش آموزش‌دیده با وظایف جدید، بدون نیاز به بازآموزی یا تغییر تمام پارامترهای مدل اصلی است. در روش سنتی تنظیم دقیق کامل، کل پارامترهای مدل برای وظیفه جدید به‌روزرسانی می‌شوند که این امر منجر به مشکلاتی مانند هزینه محاسباتی بالا، نیاز به حافظه زیاد، فراموشی فاجعه‌بار می شود.

روش‌های PEFT با ثابت نگه داشتن پارامترهای مدل اصلی و افزودن تعداد بسیار کمی پارامتر قابل آموزش (در قالب یک ماژول کوچک)، این مشکلات را حل می‌کنند. این رویکرد، فرآیند تطبیق را بسیار کارآمدتر و سبک‌تر می‌سازد.

این مقاله به طور خاص دو روش محبوب PEFT را بررسی و با هم مقایسه می‌کند:

- LoRA:

- این روش به جای تغییر مستقیم وزن‌های سنگین مدل، دو ماتریس کوچک و کم‌رتبه [1] را به لایه‌های مدل تزریق می‌کند. در حین آموزش، فقط این ماتریس‌های کوچک به‌روزرسانی می‌شوند. حاصل‌ضرب این دو ماتریس، تغییری را که در حالت تنظیم دقیق کامل باید در وزن اصلی ایجاد می‌شد، «شبیه‌سازی» یا «تقریب» می‌زند.
- ویژگی متمایز: تعداد پارامترهای قابل آموزش در LoRA بسیار کمتر از تنظیم دقیق کامل است و به اندازه مدل اصلی بستگی ندارد، بلکه به رتبه [2] انتخابی برای ماتریس‌ها وابسته است.

- $(IA)^3$:

- این روش حتی از LoRA هم پارامتر-کارآمدتر است $(IA)^3$ به جای افزودن ماتریس‌های جدید، بردارهای کوچکی را برای مقیاس‌دهی مجدد [3] به بردارهای فعال‌سازی [4] در معماری ترنسفورمر اضافه می‌کند. این بردارهای آموخته‌شده، برخی از ویژگی‌های داخلی مدل را تقویت [5] و برخی دیگر را تضعیف [6] می‌کنند تا رفتار مدل را برای وظیفه جدید تنظیم کنند.
- ویژگی متمایز: این روش تعداد پارامترهای بسیار ناچیزی را آموزش می‌دهد، زیرا فقط بردارهای مقیاس‌دهی را یاد می‌گیرد نه ماتریس‌های کامل را.

ویژگی کلیدی هر دو روش که این مقاله بر آن تمرکز دارد، ماژولار بودن آن‌هاست. از آنجایی که تغییرات در یک ماژول کوچک و جداگانه ذخیره می‌شود، می‌توان این ماژول‌ها را با استفاده از عملیات ساده حسابی (مانند جمع و تفریق) با یکدیگر ترکیب کرد تا بدون نیاز به آموزش، قابلیت‌های جدیدی در مدل ایجاد نمود، کاری که با مدل‌های تنظیم دقیق کامل غیرممکن است.

(ب) مقاله نشان می‌دهد که ترکیب ماژول‌های PEFT نسبت به استفاده‌ی منفرد از آن‌ها عملکرد بهتری دارد. دلایل این بهبود عملکرد چیست؟ همچنین توضیح دهید چرا در روش پیشنهادی مقاله نیازی به آموزش مجدد ماژول‌ها وجود ندارد.

ترکیب ماژول‌های PEFT عملکرد بهتری نسبت به استفاده‌ی منفرد از آن‌ها دارد زیرا این کار

[1] Low-Rank
[2] rank
[3] rescaling
[4] activations
[5] amplify
[6] inhibit

باعث تجمیع و یکپارچه‌سازی دانش‌ها و مهارت‌های مختلف می‌شود. به جای داشتن یک مدل متخصص در یک حوزه محدود، با ترکیب ماژول‌ها می‌توان یک مدل جامع‌تر یا با قابلیت جدید و هدفمند ساخت.

دلایل این بهبود عملکرد در مقاله به دو شکل اصلی نشان داده شده است:

1. ترکیب برای تعمیم‌پذیری[7]:

وقتی دو ماژول که هر کدام بر روی یک توزیع داده متفاوت (مثلاً یکی روی نظرات مثبت و دیگری روی نظرات منفی) آموزش دیده‌اند با هم ترکیب می‌شوند، ماژول حاصل دانش هر دو را در خود دارد. این ماژول جدید می‌تواند عملکرد بهتری روی یک مجموعه داده کلی که شامل هر دو توزیع است نشان دهد، زیرا توانایی درک هر دو جنبه را پیدا کرده است.

این موضوع به وضوح در بخش ۲.۴ (Composition for Distribution Generalization) نشان داده شده است. مقاله می‌گوید: یافته‌های ما نشان می‌دهد که یادگیری ماژولار، امکان یکپارچه‌سازی توانایی‌ها از طریق عمل جمع را فراهم می‌کند.

2. ایجاد مهارت جدید از طریق قیاس[8]:

در سناریوهای پیچیده‌تر، ترکیب ماژول‌ها فراتر از یک میانگین‌گیری ساده است. برای مثال، مقاله نشان می‌دهد که می‌توان مهارت طبقه‌بندی نظرات از یک دامنه (مثل محصولات آمازون) را به دامنه دیگر (مثل رستوران‌های Yelp) منتقل کرد. این کار با افزودن مهارت طبقه‌بندی آمازون به "تفاوت زبانی" بین دو دامنه انجام می‌شود. این عمل یک مهارت کاملاً جدید و بهینه‌شده برای دامنه هدف ایجاد می‌کند که از ابتدا وجود نداشت.

این رویکرد در بخش ۵.۴ (Composition for Domain Transfer) و با استفاده از "معادله قیاسی"[9] معروف "ملکه = پادشاه + زن - مرد" توضیح داده شده است.

در روش پیشنهادی مقاله نیازی به آموزش مجدد وجود ندارد زیرا این روش بر پایه انجام مستقیم عملیات حسابی (جمع و تفریق) بر روی پارامترهای ماژول‌های از قبل آموزش‌دیده استوار است.

به عبارت دیگر، این فرآیند یک محاسبه است، نه یک فرآیند یادگیری. وقتی دو ماژول با هم ترکیب می‌شوند، پارامترهای آن‌ها (مثلاً ماتریس‌های A و B در LoRA) به صورت

[7] Aggregation of Knowledge

[8] Targeted Skill Creation

[9] analogy equation

عددی با هم جمع یا از هم کم می‌شوند تا پارامترهای ماژول جدید را بسازند. در این میان هیچ‌گونه فرآیند بهینه‌سازی، تابع هزینه[10] یا پس‌انتشار[11] وجود ندارد. تمام دانش لازم از قبل در ماژول‌های اولیه موجود است و فقط آن‌ها به شیوه‌ای هوشمندانه در کنار هم قرار داده می شوند.

(ج) طبق یافته‌های مقاله، چرا ترکیب PEMهایی که با مقداردهی اولیه متفاوت آموزش دیده‌اند، ممکن است منجر به کاهش کارایی مدل ترکیبی شود؟

طبق یافته‌های مقاله، ترکیب PEM هایی که با مقداردهی اولیه متفاوت آموزش دیده‌اند ممکن است به کاهش کارایی منجر شود، زیرا مقداردهی اولیه متفاوت باعث می‌شود که هر ماژول در طی آموزش، یک مسیر منحصربه‌فرد را طی کند و در نهایت در یک حوضچه خطاloss basin متفاوت در فضای پارامترها قرار بگیرد.

(د) مقاله چگونه نشان می‌دهد که ترکیب PEMها در فضای وزن می‌تواند منجر به عملکردی فراتر از حالت تک‌وظیفه‌ای شود؟ با استناد به جدول‌ها یا نمودارهای مقاله، توضیح دهید که این ترکیب چگونه موجب تعمیم بهتر به وظایف یا داده‌های جدید می‌شود.

مقاله با ارائه سناریوهایی که در آن‌ها ترکیب PEM ها منجر به خلق قابلیت‌های جدید یا بهبود چشمگیر در تعمیم‌پذیری می‌شود، نشان می‌دهد که این روش فراتر از یک مدل تک وظیفه‌ای ساده عمل می‌کند. روش پیشنهادی مقاله می‌تواند مهارتی کاملاً جدید را بدون آموزش مستقیم بسازد.

این موضوع از دو طریق کلیدی نشان داده می‌شود:

- ساخت یک مهارت جدید از طریق قیاس: پیشرفته‌ترین نمونه، سناریوی انتقال دامنه است. در این حالت، مقاله یک طبقه‌بند برای نظرات رستوران‌های Yelp می‌سازد، در حالی که هیچ داده برچسب‌دار طبقه‌بندی از Yelp در اختیار ندارد. این کار با معادله قیاسی زیر انجام می‌شود:

$$\text{طبقه‌بند}Yelp = \text{طبقه‌بند}Amazon + (\text{زبان}Amazon - \text{زبان}Yelp)$$

- این عملیات یک مهارت کاملاً جدید (طبقه‌بندی در دامنه Yelp) را از ترکیب مهارت‌های موجود «تولید» می‌کند که این توانایی بسیار فراتر از انجام چند وظیفه از پیش تعریف‌شده

[10] loss function

[11] backpropagation

است.

- تجمیع دانش برای تعمیم‌پذیری بهتر :در سناریوی تعمیم توزیع، مقاله دو ماژول را که هر کدام روی بخش خاص و متفاوتی از داده‌ها آموزش دیده‌اند، با هم ترکیب می‌کند. ماژول ترکیبی نهایی، نه تنها هر دو مهارت را دارد، بلکه در مواجهه با داده‌های جدید و جامع، عملکردی بهتر از میانگین دو ماژول اولیه از خود نشان می‌دهد، زیرا به درک کامل‌تری از توزیع کلی داده‌ها دست یافته است.

این بهبود در تعمیم‌پذیری به وضوح در جدول‌های مقاله قابل مشاهده است:

- جدول ۲ (Table 2 - Composition for Distribution Generalization): این جدول نتایج ترکیب دو PEM را که روی توزیع‌های داده‌ای متفاوت آموزش دیده‌اند، نشان می‌دهد. ستون PEM(Merged PEM ادغام‌شده(به طور مداوم عملکردی بهتر از میانگین عملکرد دو ماژول تکی دارد. برای مثال، در مجموعه داده RTE، ماژول ترکیبی LoRA توانسته ۲.۵ واحد بهبود مطلق نسبت به میانگین دو ماژول اولیه کسب کند. این نشان می‌دهد که ماژول ترکیبی به شکل بهتری به توزیع کلی داده‌ها تعمیم یافته است.
- جدول ۵ (Table 5 - Composition for Domain Transfer): این جدول عملکرد مدل ترکیبی برای طبقه‌بندی در دامنه جدید را با یک "خط پایه انتقال ساده" (Vanilla Transfer) مقایسه می‌کند. نتایج نشان می‌دهد که ترکیب حسابی ماژول‌ها توانسته است مهارتی ایجاد کند که به شکل بسیار بهتری به یک وظیفه و دامنه کاملاً جدید تعمیم پیدا می‌کند.

(ه) با وجود مزایای ترکیب ماژول‌های PEFT، مقاله به چه محدودیت‌هایی در به‌کارگیری این روش در کاربردهای واقعی مدل‌های زبانی بزرگ اشاره می‌کند؟

مقاله با وجود تاکید بر مزایای ترکیب ماژول‌های PEFT، به چند محدودیت مهم در به‌کارگیری این روش در کاربردهای واقعی اشاره می‌کند:

- به ارث بردن سوگیری‌ها و مشکلات ایمنی: اگر ماژول‌های اولیه‌ای که با هم ترکیب می‌شوند دارای سوگیریbias یا مشکلات ایمنی باشند، ماژول ترکیبی نهایی نیز این مشکلات را به ارث خواهد برد. این موضوع در کاربردهای واقعی که ایمنی و انصاف اهمیت بالایی دارد، یک نگرانی جدی است.
- ماهیت جعبه-سیاه و ریسک‌های پنهان : مقاله هشدار می‌دهد که حتی پس از «سم‌زدایی» یک مدل با کم کردن یک ماژول سمی، به دلیل ماهیت جعبه-سیاه شبکه‌های عصبی،

هیچ تضمینی وجود ندارد که سمی‌بودن به شکل پنهان در بخش‌های دیگر مدل باقی نمانده باشد و در شرایط پیش‌بینی‌نشده بروز نکند.

- محدودیت در معماری و مقداردهی اولیه : نویسندگان اذعان می‌کنند که اکثر آزمایش‌هایشان تحت شرایط کنترل‌شده انجام شده است؛ یعنی ماژول‌هایی که با هم ترکیب شده‌اند دارای معماری یکسان و در اغلب موارد مقداردهی اولیه یکسان بوده‌اند. در دنیای واقعی، ممکن است کاربران بخواهند ماژول‌هایی با ساختارهای متفاوت مثلاً ترکیب یک ماژول LoRA با یک ماژول $(IA)^3$ را با هم ادغام کنند که عملکرد روش در این حالت هنوز بررسی نشده است.
- نیاز به تنظیم دستی ابرپارامتر[۱۲]: رویکرد پیشنهادی برای یافتن بهترین ترکیب، به تنظیم دستی ابرپارامتر وزن λ نیاز دارد. این موضوع فرآیند را از یک راه‌حل ساده و سریع دور کرده و نیازمند آزمایش و تنظیمات اضافی برای رسیدن به نتیجه مطلوب است که در عمل می‌تواند هزینه‌بر باشد.

۲.  درمورد روش MOCO به سوالات زیر پاسخ دهید(۱۵ نمره).

برای پاسخ این سوال، از پاسخ خانم میاهی نیا و آقای حسین زاده استفاده شده است.

(آ) انتخاب ضریب تکانه مناسب برای به روز رسانی کدگذارکلید [۱۳] اهمیت بالایی دارد. دلیل آن چیست؟ اگر این ضریب بیش از حد کوچک یا بزرگ انتخاب شود چه مشکلاتی ایجاد میکند؟

در روش MoCo، ضریب تکانه نقش مهمی در به‌روزرسانی شبکه key encoder دارد، زیرا این شبکه مستقیما آموزش نمیبیند و از طریق میانگین گیری نمایی از پارامترهای شبکه query به‌روزرسانی میشود. اگر این ضریب بیش از حد کوچک باشد، شبکه key encoder ناپایدار شده و باعث بی اعتباری key های قبلی و کاهش دقت مدل میشود. از طرف دیگر، اگر ضریب خیلی بزرگ باشد، به‌روزرسانی شبکه encoder key بسیار کند شده و از تغییرات مدل عقب میماند، که باعث کاهش سرعت یادگیری و کارایی مدل میشود. بنابراین، انتخاب مناسب این ضریب برای پایداری و کارایی روش بسیار حیاتی است.

(ب) نقش صف نمونه های منفی را در این الگوریتم شرح دهید. در الگوریتم MoCo صف نمونه‌های منفی نقش کلیدی در آموزش کنتراستیو دارد. این صف شامل ویژگی‌های استخراج شده از

[12] Manual Hyperparameter Tuning
[13] Key Encoder

تصاویر key در مراحل قبلی آموزش است و به عنوان منابع منفی ثابت و متنوع برای مقایسه با نمونه مثبت استفاده میشود. در فرایند آموزش، هدف مدل این است که embedding های query را به نمونه مثبتش نزدیک کرده و از نمونه‌های منفی متمایز کند. برای انجام این کار به تعداد زیادی نمونه منفی نیاز است، اما بارگذاری و پردازش همزمان آنها از نظر حافظه دشوار است. MoCo این مشکل را با استفاده از یک صف حافظه چرخشی حل میکند: پس از هر مرحله، ویژگی کلید جدید به صف افزوده میشود و قدیمی ترین نمونه حذف میگردد. بنابراین، این صف باعث میشود که بدون نیاز به پردازش همزمان حجم زیادی از داده‌ها، مدل همیشه به مجموعه‌ای بزرگ و به‌روز از نمونه‌های منفی دسترسی داشته باشد و بتواند آموزش کنتراستیو مؤثر و پایدار انجام دهد.

(ج) فرض کنید میخواهیم برای مسئله دسته بندی تصاویر MRI، از روش MOCO برای یادگیری بازنمایی استفاده کنیم. با توجه به اینکه تعداد تصاویر بدون برچسب فقط ۲۰۰۰ نمونه است، یادگیری بازنمایی به خوبی انجام نمی شود. اگر با استفاده از تکنیک های داده افزایی، از هر نمونه تعداد ۱۰ نمونه جدید ساخته شود تا در نهایت روش MOCO با ۲۲۰۰۰ نمونه آموزش ببیند،چه تاثیری خواهد داشت؟ در نهایت دقت مسئله دسته بندی چه تغییری خواهد کرد؟
این روش میتواند به آشفتگی در انتخاب نمونه‌های مثبت یا منفی منجر شود. MoCo بر پایه این فرض کار میکند که «دو نما ازیک تصویر»مثبت هستند و «هر چیزدیگر»منفی. اگر همه augment هایی یک تصویررا جداگانه به مدل بدهیم، MoCo نمیداند که آنها از یک پدرند و برخی از آنها را به عنوان منفی نسبت به هم در صف قرار میدهد؛ پس مدل به جای نزدیک کردن نماهای یکسان، سعی میکند آنها را از هم دور کند.

سوالات عملی

۳.  در این سوال قصد داریم به قابلیت دسته بندی بدون نمونه در مدل Clip بپردازیم. بدین منظور از کتابخانه open-clip استفاده میکنیم.(۳۰ نمره)

(آ) ابتدا نسخه ConvNext-Base را فراخوانی کنید. سپس از دیتاست Stanford-dogs (مجموعه آزمون)برای ارزیابی توانایی یادگیری بدون نمونه این مدل بهره بجویید. برای انتخاب قالب پرامپت متن مجاز هستید از یک قالب دلخواه استفاده کنید.

(ب) چند قالب متن دیگر را به دلخواه امتحان کنید و نتایج آنها را مقایسه کنید. توجه داشته باشید،

به علت تعداد بالای کلاس های مجموعه داده، انتخاب قالب های متفاوت برای کلاس های مختلف عملی به نظر نمیرسد.

(ج) accuracy، precision، recall و F1 را به تفکیک کلاس گزارش کنید. کلاسی که بدترین F1 را دارد را اعلام کنید. سعی کنید برای این کلاس قالب های متفاوتی پیدا کنید تا عملکرد مدل درمورد آن کلاس بهبود یابد. نتایج را مقایسه کنید.

(د) برای هر کلاس یک جاگذاری قابل آموزش تعریف کنید و به کمک گرادیان کاهشی آنرا آموزش دهید(مجموعه آموزش). سپس دقت را روی مجموعه آزمون محاسبه و گزارش کنید.

(ه) رمزگذار متن را کنار بگذارید. رمزگذار تصویر را منجمد کرده و یک لایه تمام متصل اضافه کنید. لایه اضافه شده را آموزش دهید(مجموعه آموزش). سپس دقت را روی مجموعه آزمون محاسبه و گزارش کنید.

(و) اینبار مدل ConvNext-XXLarge را فراخوانی کرده و آزمایشهای قسمت الف و ب را با آن تکرار کرده، نتایج را مقایسه کنید.

توجه: برای گزارش این سوال نتایج را مقایسه، تحلیل و سپس توجیه کنید.

به نوتبوک Q3.ipynb مراجعه شود.

۴.  در این سوال قصد داریم با استفاده از یک مدل زبانی فارسی با soft prompting آشنا شده و سپس به روش‌هایی از Reasoning بپردازیم. قسمت های مشخص شده در نوت بوک LLM.ipynb را تکمیل کنید(۲۰ نمره).

به نوتبوک Q4.ipynb مراجعه شود.

۵.  در این تمرین با استفاده از کتابخانه‌های HuggingFace و PEFT، یک مدل Vision Transformer (ViT) را روی دیتاست Food101 آموزش خواهید داد. هدف اصلی استفاده از روش Low-Rank Adaptation (LoRA) برای آموزش مؤثرتر و کم‌هزینه‌تر مدل است. لطفا نوتبوک peft_lora.ipynb را کامل کنید(۲۰ نمره).

- داده‌ها را بارگذاری و آماده‌سازی کنید (از دیتاست food101 استفاده شده است).
- مدل ViT را از HuggingFace بارگیری کنید.
- با استفاده از PEFT و LoRA، بخش‌هایی از مدل را قابل آموزش قرار دهید.
- مدل را با استفاده از Trainer آموزش دهید.

- در پایان، از مدل آموزش‌دیده برای پیش‌بینی تصویر جدید استفاده کنید.

به نوتبوک Q5.ipynb مراجعه شود.