به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس مدل های مولد عمیق

مدرس: دکتر مصطفی توسلی‌پور

تمرین شماره ۴

دی ماه ۱۴۰۳

فهرست

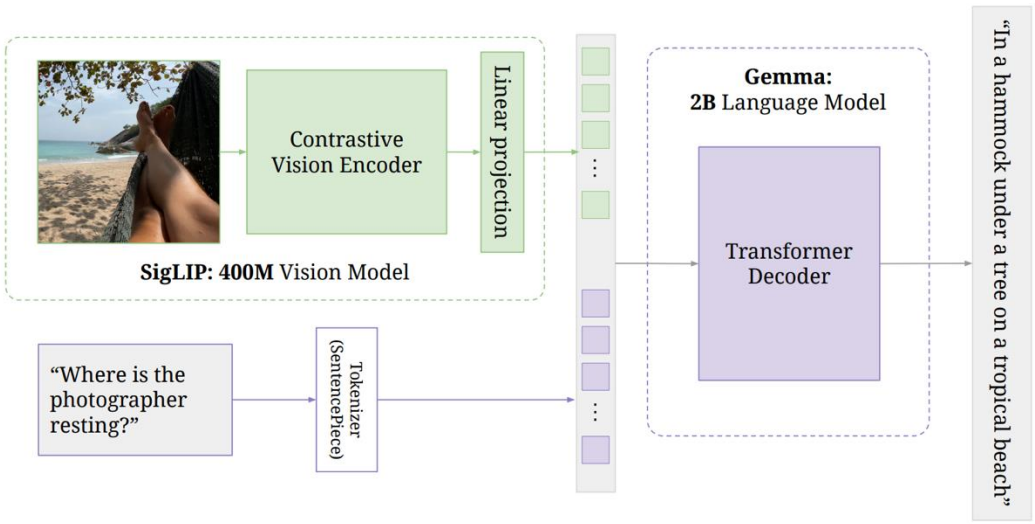
سوال اول : *VISION-LANGUAGE MODEL*

در این سوال قصد داریم از یک مدل پیش آموزش‌دیده‌ی زبان بصری [۱] استفاده کرده و پروژه‌ای در زمینه پاسخگویی به سوالات مرتبط با تصویر [۲] پیاده‌سازی کنیم. در این پروژه، شما با نحوه فاین‌تیون کردن یک مدل چندمدی [۳] از پیش آموزش‌دیده بر روی داده‌های جدید و چالش‌های آن آشنا خواهید شد. این چالش‌ها شامل درک عملکرد مدل، پیش‌پردازش داده‌ها و نحوه استفاده از روش‌های PEFT برای فاین‌تیون کردن مدل‌های بزرگ هستند.

بخش اول - مدل PALIGEMMA

مدل Paligemma توسط شرکت گوگل معرفی شده و جزو مدل‌های سبک [۴] در حوزه زبان بصری محسوب می‌شود. برخلاف سایر مدل‌های زبان بصری که به طور مداوم در حال افزایش اندازه خود هستند تا به هدف هوش مصنوعی عمومی [۵] نزدیک‌تر شوند، تمرکز اصلی این مدل بر روی فشرده‌سازی بهینه است تا کاربران بتوانند به راحتی آن را بر روی داده‌های خود فاین‌تیون و شخصی‌سازی کنند.

همانطور که در شکل ۱ دیده می‌شود، این مدل از یک مدل Visual Encoder به نام SigLIP تشکیل شده است که برای پردازش ورودی‌های تصویری اختصاص یافته است. علاوه بر این، در ساختار این مدل، ساختار Decoder-only مدل Gemma برای پردازش توکن‌های ورودی و تولید متن خروجی در نظر گرفته شده است.



شکل ۱. ساختار مدل PaliGemma

---

۱ Vision-language model (VLM)
۲ Visual Question Answering
۳ Multimodal
۴ Light-weight
۵ Artificial General Intelligence

زیربخش اول – VLM

سوال ۱: درباره مدل‌های زبان بصری و تفاوت‌های آن‌ها با مدل‌های سنتی تصویری یا متنی توضیح دهید. همچنین، کاربردها و نحوه عملکرد آن‌ها را به طور مختصر شرح دهید. **(۲/۵ نمره)**

سوال ۲: برای آموزش یک مدل زبان بصری ، رویکردهای مختلفی وجود دارد. یکی از این رویکردها که در مدل Paligemma استفاده شده، آموزش مدل‌های ماژولار و به‌کارگیری آن‌ها در ساختار مدل است. رویکرد دیگر که در مدل‌هایی مانند DALL·E و Imagen به کار رفته، آموزش end-to-end مدل است. در این بخش، به توضیح این دو رویکرد و مزایا و معایب هرکدام پرداخته و معماری مدل Paligemma را با دو مدل DALL·E و Imagen به طور کلی مقایسه کنید. **(۲/۵ نمره)**

زیربخش دوم – SIGLIP IMAGE ENCODER

یکی از بخش‌های مهم معماری Paligemma، مدل image encoder آن است که داده‌های تصویری را پردازش کرده و اطلاعات آن‌ها را برای ورود به decoder مدل زبانی آماده می‌کند. Image encoder استفاده شده در این مدل از مدل SigLIP گرفته شده است.

سوال ۱: درباره معماری، نحوه آموزش فضای joint embedding و تابع loss بکار رفته در آموزش مدل SigLIP مختصرا توضیح دهید. **(۵ نمره)**

سوال ۲: همچنین در آموزش این مدل، ایده‌ای به‌کاررفته است که باعث مزیت سرعت آموزش این مدل نسبت به مدل CLIP[۱] می‌شود. این ایده و مزیت آن را نیز توضیح دهید. **(۵ نمره)**

زیربخش سوم –PRE-TRAINING

سوال ۱ : همان‌طور که در مقاله Paligemma توضیح داده شده است، پس از آموزش image encoder و text decoder به‌طور جداگانه و بر روی داده‌های uni-modal، این مدل‌ها با یکدیگر ادغام شده و در معماری کلی مدل قرار می‌گیرند. پس از طی فاز اولیه پیش‌آموزش مدل، در مراحل دوم و سوم، مدل مجدداً آموزش داده می‌شود. در این بخش، نحوه پیش‌آموزش مدل در طی این دو مرحله و لزوم هر یک از این مراحل توضیح دهید. **(۶ نمره)**

[۱] Learning Transferable Visual Models From Natural Language Supervision (2021)

سوال ۲ : ورودی مدل gemma یک token به فرم زیر است:

Tokens = [image tokens..., BOS, prefix tokens..., SEP, suffix tokens..., EOS, PAD...]

توضیح دهید prefix tokens و suffix tokens به چه منظور در این مدل استفاده می‌شود. **(۴ نمره)**

سوال ۳ : یکی از چالش‌های موجود در ساختار چنین مدلی این است که مدل زبانی استفاده‌شده بر روی داده‌های متنی آموزش دیده است و بنابراین نحوه پردازش و مدیریت توکن‌های متنی را به خوبی فرا گرفته است. در مواجهه با داده‌های تصویری، ممکن است عملکرد آن کاهش یابد. به طور کلی، چه روشی برای رفع این مشکل می‌توان استفاده کرد و مقاله Paligemma چگونه این مسئله را حل کرده است؟ **(۵ نمره)**

زیربخش چهارم –TRANSFER LEARNING

همان‌طور که قبلاً اشاره شد، مدل Paligemma به‌منظور تسهیل فرآیند فاین‌تیونینگ بر روی داده‌های خاص، با ساختاری فشرده‌تر از سایر مدل‌های زبان بصری طراحی شده است. با این وجود، کوچک‌ترین نسخه این مدل حدود ۳ میلیارد پارامتر دارد که با توجه به محدودیت‌های سخت‌افزاری موجود برای کاربران عادی، فاین‌تیون کردن این مدل‌ها برای آن‌ها عملاً غیرممکن است. به همین دلیل، برای فاین‌تیون کردن چنین مدل‌های بزرگی از روش‌های PEFT[۱] استفاده می‌شود که یکی از معروف‌ترین آن‌ها، که ما نیز در این تمرین قصد استفاده از آن را داریم، روش QLoRA است.

سوال ۱: تفاوت‌های اصلی بین QLora و LoRA را بیان کنید و مزایا و معایب هرکدام را توضیح دهید.
همچنین، یکی از هایپرپارامترهای مهم در استفاده از این تکنیک‌ها Rank است. در این راستا، تأثیر این پارامتر بر عملکرد مدل نهایی را توضیح دهید.
علاوه بر این، یکی از datatype هایی که برای کوانتیزه کردن مدل در این روش‌ها استفاده می‌شود، NF4 است. درباره این نوع datatype نیز به‌طور مختصر توضیح دهید. **(۵ نمره)**

سوال ۲: در این بخش می‌خواهیم حافظه مورد نیاز برای فاین‌تایون کردن یک مدل را در دو حالت مقایسه کنیم:

حالت اول: زمانی که تمام پارامترهای مدل به طور کامل تغییر می‌کنند.
حالت دوم: زمانی که از روش LoRA برای فاین‌تایون کردن استفاده می‌شود.

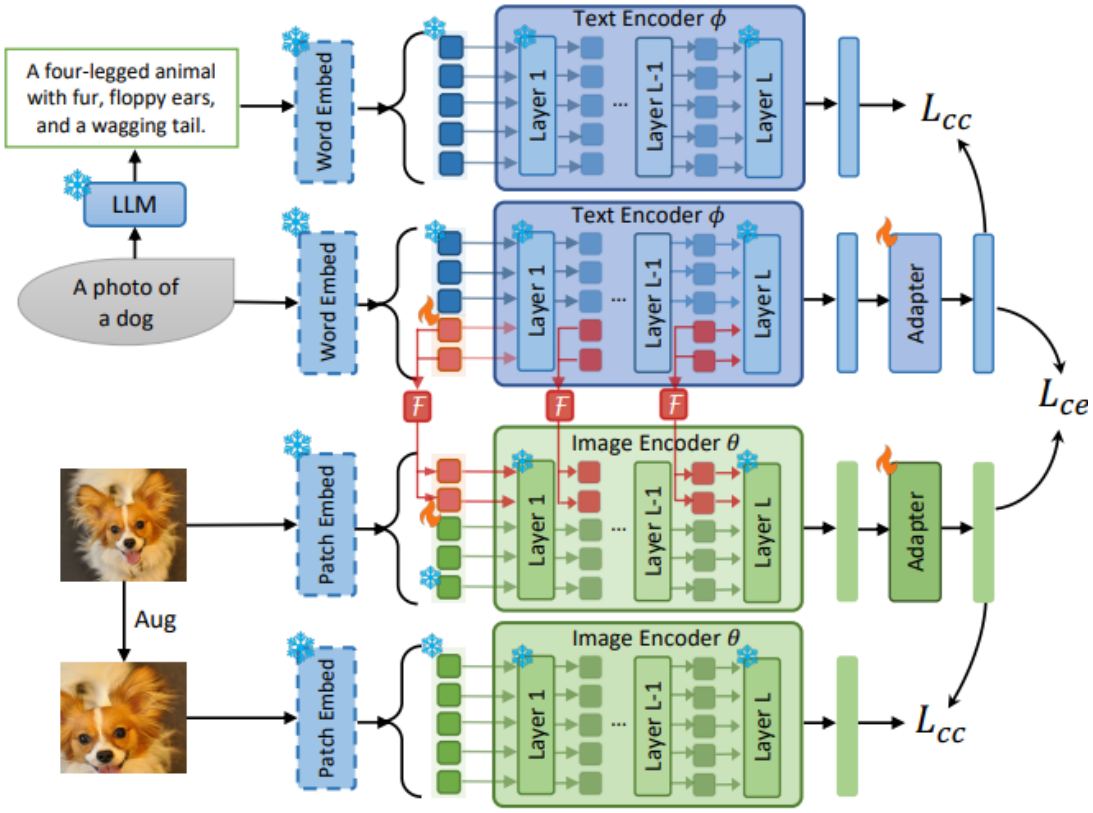
دقت کنید که برای این سوال تمامی پارامترهایی که در حین فاین‌تیونینگ در حافظه قرار می‌گیرند را در

[۱] Parameter-Efficient Fine-Tuning

محاسبات خود در نظر بگیرید. برای اینکه نتایج به‌دست‌آمده قابل صحت‌سنجی باشند، می‌توانید از اطلاعات زیر برای انجام محاسبات خود استفاده کنید. **(۵ نمره)**

Model Data type	bfloat16
Optimizer	AdamW
LoRA Rank	16
LoRA Target Modules	q_proj, o_proj, k_proj, v_proj,gate_proj,up_proj,down_pro

سوال ۳: یکی از مشکلاتی که پس از فاین‌تیونینگ مدل‌ها به وجود می‌آید، overfitting است که در آن مدل قدرت generalization خود را از دست می‌دهد و در مواردی حتی عملکرد مدل بدتر از چیزی می‌شود که قبل از فاین‌تیونینگ بود. این پدیده را به‌طور خاص catastrophic forgetting می‌نامند. این مقاله با ایجاد تغییراتی در ساختار مدل اولیه، مدل را به‌گونه‌ای فاین‌تیون می‌کند که مشکلات مطرح‌شده تا حد زیادی برطرف می‌شوند. ساختار پیشنهادی در مقاله در شکل ۲ نشان داده شده است. پس از مطالعه این مقاله، ایده نوآورانه آن و نحوه تأثیر آن بر قدرت generalization مدل را مطرح کنید. **(۱۰ نمره)**



شکل ۲. ساختار مدل در مقاله CoPrompt

زیربخش پنجم – پیاده سازی

معرفی دیتاست

همان‌طور که اشاره کردیم، در این تمرین قصد داریم مدل Paligemma را برای تسک پاسخگویی به سوالات مرتبط با تصویر فاین‌تیون کنیم. از آنجا که با محدودیت‌های سخت‌افزاری مواجه هستیم، یکی از چالش‌های اصلی ما پیدا کردن یک دیتاست غنی و با حجم کم است که بتوان آن را بر روی سخت‌افزارهای در دسترس استفاده کرد. علاوه بر این، بسیاری از benchmark های موجود برای این تسک دارای سوگیری هستند، چرا که در متن‌های مربوط به تصاویر، اطلاعاتی وجود دارد که مدل می‌تواند از آن‌ها برای پاسخ به سوالات استفاده کند. این امر باعث می‌شود که مدل به جای استفاده از استدلال[۱] برای پاسخ‌دهی، از اطلاعات مستقیم موجود در متن استفاده کند. در این تمرین، ما از دیتاست CLEVR استفاده می‌کنیم که یک مجموعه داده تشخیصی است و طیفی از توانایی‌های استدلال بصری را آزمایش می‌کند. این دیتاست دارای حداقل سوگیری است و حاشیه‌نویسی‌های دقیقی

[۱] Reasoning

دارد که نوع استدلال مورد نیاز برای هر سوال را توصیف می‌کند. از این مجموعه داده برای تجزیه و تحلیل انواع سیستم‌های استدلال بصری مدرن استفاده می‌شود و بینش جدیدی در مورد توانایی‌ها و محدودیت‌های آن‌ها ارائه می‌دهد. این دیتاست به‌طور آزاد از طریق Hugging Face قابل دسترسی است، ولی با توجه به حجم زیاد آن و محدودیت‌های سخت‌افزاری، شما می‌توانید از ۱٪ یا بیشتر از کل داده‌ها استفاده کرده و آن‌ها را به دو بخش داده‌های آموزش و تست تقسیم کنید (یا از تقسیم‌بندی پیش‌فرض دیتاست استفاده کنید).

سوال ۱: آموزش مدل

برخی از مدل‌های از پیش‌آموزش‌دیده در Huggingface موجود هستند، بنابراین برای فاین‌تیون کردن Paligemma، نیاز است از کتابخانه Transformers و ابزارها و مدل‌های ارائه شده توسط آن استفاده شود. این ابزارها برای آماده‌سازی داده‌ها و آموزش مدل نیز قابل‌استفاده هستند.(برای کوانتیزه کردن مدل نیز می‌توانید از BitsAndBytes استفاده کنید)

همچنین در فرآیند آموزش، نمایش میزان loss در هر مرحله ضروری است تا پیشرفت مدل به‌درستی ارزیابی شود. **(۲۰ نمره)**

سوال ۲ معیار سنجش عملکرد(ROUGE SCORE)

با توجه به اینکه خروجی نهایی مدل متن است، برای ارزیابی عملکرد می‌توان از معیارهای استاندارد سنجش مدل‌های زبانی، مانندROUGE score ، استفاده کرد. ابتدا باید توضیحی کامل درباره این معیار ارائه شود تا اهمیت و نحوه کارکرد آن روشن شود.

(۱۰ نمره)

سوال ۳ ارزیابی مدل بر اساسROUGE SCORE

برای مقایسه عملکرد مدل، معیار ROUGE score باید به دو صورت اجرا و گزارش شود:

بر روی داده‌های تست و مدل اولیه.

بر روی داده‌های تست و مدل فاین‌تیون‌شده.

این مقایسه میزان بهبود عملکرد مدل را نشان می‌دهد. **(۷ نمره)**

سوال ۴ نمایش خروجی های نمونه

به صورت تصادفی ۱۰ نمونه از داده‌های تست انتخاب کنید و در گزارش خود موارد زیر را ارائه دهید. **(۳ نمره)**

با توجه به محدودیت‌های سخت‌افزاری، ما سعی کردیم مدل و مجموعه داده را به‌گونه‌ای انتخاب کنیم که در صورت

پیاده‌سازی صحیح و تنظیم مناسب هایپرپارامترها، بتوانید مدل را در محیط Colab یا Kaggle آموزش دهید و پیش از اتمام محدودیت زمانی استفاده از GPU، نتایج قابل قبولی را مشاهده کنید.

سوال دوم: FLOW MATCHING

مدل‌های FlowMatching با تعریف یک فرآیند دینامیکی پیوسته، داده‌ها را از یک توزیع اولیه (معمولاً نویز گوسی) به توزیع هدف هدایت می‌کنند. این مدل‌ها به جای تخمین مستقیم توزیع یا گرادیان آن، بر یادگیری میدان‌های برداری متمرکز هستند که مسیرهای انتقال داده را مشخص می‌کنند. با استفاده از معادلات دیفرانسیل عادی[1] و اطلاعات زمانی صریح، این مدل‌ها رفتار پیچیده‌تری را در طول مسیر تحول داده‌ها یاد می‌گیرند. رویکرد FlowMatchingبه دلیل کاهش نیاز به نمونه‌گیری‌های مکرر، از نظر محاسباتی کارآمدتر است و تفسیرپذیری بالاتری نسبت به مدل‌های Diffusion ارائه می‌دهد. این ویژگی‌ها آن را برای تولید داده‌های جدید، بهینه‌سازی، و مدل‌سازی سیستم‌های دینامیکی پیچیده مناسب می‌سازد.

زیربخش اول

مطابق آن‌چه در مقاله‌ی flow matching بیان شده است، نشان دهید که رابطه زیر تحت چه شرایطی و چرا برقرار است.

$$\nabla_\theta L_{FM}(\theta) = \nabla_\theta L_{CFM}(\theta)$$

(۵ نمره)

زیربخش دوم

با مطالعه بخش های ابتدایی مقاله‌ی flow matching با توجه به رابطه ۱۱ و Theorem ۳ نشان دهید:

$$u_t(x \mid x_1) = \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}\big(x - \mu_t(x_1)\big) + \mu_t'(x_1).$$

(۱۰ نمره)

---

[1] ODE

زیربخش سوم

شرح مختصری از مسئله Optimal transport را بیان کرده و توضیح دهید که از این مسئله چگونه در مدل‌های flow matching می‌توان استفاده کرده؟ مزایا و معایت استفاده از optimal transport در این مدل‌ها چیست؟**(۵ نمره)**

مراجع

1. PaliGemma: A versatile 3B VLM for transfer
2. Sigmoid Loss for Language Image Pre-Training
3. LoRA: Low-Rank Adaptation of Large Language Models
4. Consistency-guided Prompt Learning for Vision-Language Models
5. Flow Matching for Generative Modeling

نکات تحویل

- مهلت ارسال این تمرین تا پایان روز " ۱۶ دی ماه" خواهد بود.
- این زمان قابل تمدید نیست و درصورت نیاز می‌توانید از grace time استفاده کنید.
- در نظر داشته باشید که حداکثر مهلت آپلود تمرین در سامانه تا ۷ روز پس مهلت تحویل است و پس از آن سامانه بسته خواهد شد.
- پیاده سازی با زبان برنامه نویسی پایتون باید باشد و کدهای شما باید قابل اجرا بوده و به همراه گزارش آپلود شوند.
- انجام این تمرین به صورت یک نفره می‌باشد.
- در صورت مشاهده هر گونه تشابه در گزارش کار یا کدهای پیاده‌سازی، این امر به منزله تقلب برای طرفین در نظر گرفته خواهد شد.
- استفاده از کدهای آماده بدون ذکر منبع و بدون تغییر به منزله تقلب خواهد بود و نمره تمرین شما **صفر** در نظر گرفته می‌شود
- در صورت رعایت نکردن فرمت گزارش کار نمره گزارش به شما تعلق نخواهد گرفت.
- تحویل تمرین به صورت **دستنویس** قابل پذیرش نیست.
- تمامی تصاویر و جداول مورد استفاده در گزارش کار باید دارای توضیح (caption) و شماره باشند.
- بخش زیادی از نمره شما مربوط به گزارش کار و روند حل مسئله است.
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه بارگذاری نمائید.
- HW1 _[Lastname]_[StudentNumber].zip
- در صورت وجود سوال و یا ابهام میتوانید از طریق رایانامه زیر با موضوع TAI_HW1 با دستیاران آموزشی در ارتباط باشید:
    - سوال اول

      m.dadkhah99@gmail.com
    - سوال دوم

      fatemehnadir@gmail.com

با آرزوی سلامتی و موفقیت روزافزون.