دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین

پاییز ۱۴۰۳

استاد: علی شریفی زارچی
مسئول تمرین: نسرین امجدی

تمرین ششم

مهلت ارسال نهایی: ۱ بهمن

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- این تمرین ، امتیازی می باشد و امکان آپلود تمرین پس از روز ارسال نهایی وجود ندارد .
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.

گردآورندگان تمرین: ریحانه حلوایی ، امیر رضا توکلی ، مرتضی شهرابی ، احسان شبیری ، مهدی رحیم سیرت

سوالات نظری (۱۰۰ نمره)

۱. (۲۰ نمره) به سوالات زیر پاسخ دهید :

مفهوم وظیفه‌ی پیش‌متن (pretext task) را در یادگیری خودنظارتی (self-supervised learning) توضیح دهید. همچنین، سه وظیفه‌ی پیش‌متن زیر را به طور خلاصه توضیح دهید و نوع ویژگی‌هایی را که به مدل‌ها آموزش می‌دهند، بیان کنید:

الف) پیش بینی چرخش(Rotation Prediction)

ب) رنگ آمیزی (Colorization)

ج) حل پازل(Jigsaw puzzle Solving)

۲. (۲۰ نمره) فرض کنید مجموعه‌ای از تصاویر ماهواره‌ای مناطق شهری بدون برچسب در اختیار دارید. می خواهید از یک مدل یادگیری خود نظارتی برای استخراج ویژگی های مفید (برای استفاده در وظایف بعدی مانند تشخیص ساختمان ها یا طبقه بندی کاربری زمین) استفاده کنید . از میان سه وظیفه ی پیش متن معرفی شده در سوال اول ، کدام را انتخاب می کنید ؟ انتخاب خود را با دلایل زیر توضیح دهید :

الف) چرا این وظیفه با ساختار و ویژگی‌های تصاویر ماهواره‌ای همخوانی دارد؟

ب) چگونه می‌توان این وظیفه پیش‌متن را روی این داده‌ها اعمال کرد؟

ج) دو وظیفه‌ی دیگر چه محدودیت‌هایی برای این نوع داده‌ها دارند؟

۳. (۲۰ نمره) به یک مدل ViT (Vision Transformer) تصویری با ابعاد 224×224 پیکسل داده شده است. این تصویر به پچ‌هایی (patch) با ابعاد 16×16 پیکسل تقسیم می‌شود. هر پچ به یک بردار مسطح تبدیل شده و سپس با یک لایه خطی (Linear) به یک جاسازی (embedding) با ابعاد 128 نگاشت می‌شود.

با توجه به این توضیحات، به سوالات زیر پاسخ دهید:

الف) ابتدا تعداد کل بچ‌ها (N) را محاسبه کنید. سپس فرض کنید هر پچ شامل 256 مقدار پیکسل (16×16) باشد، فرآیند تبدیل این بچ به جاسازی 128 بعدی با استفاده از ماتریس تبدیل خطی را توضیح دهید.

ب) توضیح دهید چگونه جاسازی موقعیتی (Positional Embedding) به جاسازی های پچ اضافه می شود و دلیل اهمیت این کار را توضیح دهید .

ج) دنباله ورودی برای رمزگذار (Encoder) مدل ترنسفورمر را شامل توکن ویژه [CLS] بسازید. ابعاد این بردار را هم بیان کنید و در انتها، نقش توکن ویژه (CLS) و کاربرد و نحوه‌ی استفاده از آن را بیان کنید.

۴. (۲۰ نمره) به سوالات زیر پاسخ دهید :

فرض کنید مدل CLIP تصویری از یک «سیب قرمز» و سه متن توصیفی زیر دریافت می‌کند:

- «یک سیب قرمز آبدار روی میز»
- «یک سیب سبز آویزان از درخت»
- «یک توپ قرمز درخشان»

الف) توضیح دهید CLIP چگونه امتیاز شباهت را برای هر جفت تصویر و متن محاسبه می‌کند. به نظر شما کدام جفت احتمالاً بالاترین امتیاز را می‌گیرد؟ دلیل خود را بیان کنید.

ب)اگر مدل، توصیف «یک توپ قرمز درخشان» را بالاتر از «یک سیب سبز» رتبه‌بندی کند، این رفتار چه چیزی درباره‌ی فضای نمایشی (embedding space) یادگرفته‌شده توسط CLIP نشان می‌دهد؟

۵. (۲۰ نمره)

با توجه به مقاله به سوالات زیر پاسخ دهید.

الف) مکانیزم attention pooling را با global average pooling از نظر نحوه عملکرد و تولید خروجی با هم مقایسه کنید .

ب) ماتریس لیبل، به اندازه $N \times N$ شامل درایه‌های صفر و یک برای آموزش مدل به‌صورت contrastive learning می‌باشد. تعداد درایه‌های صفر در این ماتریس را به‌صورت رابطه‌ای از N بنویسید.

ج) مطابق مقاله، قدرت zero-shot مدل کلیپ در چه نوع تسک‌هایی ضعیف‌تر است؟ بنظر شما علت این امر در چیست؟