**دانشگاه صنعتی امیرکبیر**

**دانشکده مهندسی کامپیوتر**

# پروژه هفتم درس یادگیری عمیق

**زمستان ۱۴۰۲**

**هدف:** آشنایی با شبکه ترنسفورمر و مدل‌های بنیادی در یادگیری عمیق

**پیاده‌سازی:** استفاده از تمامی زبان‌های برنامه‌نویسی مجاز است. در این تمرین استفاده از کتابخانه‌های یادگیری عمیق مجاز است.

**گزارش:** ملاک اصلی ارزیابی پروژه‌ها، گزارش آن‌ها است. گزارش پروژه باید در قالب pdf باشد و در آن تمامی تصاویر ورودی و خروجی و توضیحات مربوطه ذکر گردد. توجه کنید گزارش شما باید بدون نیاز به مراجعه به فایل‌های پیاده‌سازی قابل درک باشد.

**تذکر:** مطابق قوانین دانشگاه هر گونه کپی‌برداری و اشتراک کار دانشجویان غیرمجاز بوده و شدیدا برخورد خواهد شد. استفاده از کدها و توضیحات اینترنت به منظور یادگیری بلامانع است، اما کپی کردن غیرمجاز است.

**راهنمایی:** در صورت نیاز، سوالات خود را در گروه تلگرام درس یا با ایمیل زیر مطرح کنید:

E-mail: ann.ceit.aut@gmail.com

**ارسال پاسخ‌ها:** فایل‌های کد و گزارش را در قالب یک فایل فشرده با فرمت DL#_StudentID.zip که # شماره پروژه است در سامانه کورسز بارگذاری کنید. تاریخ مجاز پروژه در سامانه کورسز قابل مشاهده است.

**قوانین تاخیر:** در طول ترم در مجموع مجاز به حداکثر ۱۰ روز تاخیر در ارسال پاسخ‌ها هستید. این مدت برای تمام پروژه‌ها بوده و تصمیم‌گیری در مورد میزان استفاده از آن در هر پروژه به عهده شما است. پس از اتمام این ۱۰ روز، هر روز تاخیر اضافه منجر به کسر ۱۰ درصد از نمره پروژه مربوطه خواهد شد.

مدل‌های بنیادی[1] که در حوزه پردازش زبان‌های طبیعی با نام مدل‌های زبانی بزرگ[2] نیز شناخته می‌شوند، در حال ایجاد تحول در همه حوزه‌های یادگیری عمیق هستند. ایده اصلی همه این مدل‌ها، آموزش شبکه‌های عمیق چندمنظوره (معمولا مبتنی بر ترنسفورمر) بر روی حجم بسیار زیادی از داده‌های موجود در اینترنت (معمولا به صورت بدون نظارت) است. مثال‌های معروفی از این شبکه‌ها، BERT، GPT و DALL-E هستند. به لطف داده‌های فراوان، آموزش منعطف و البته قدرت پردازشی صدها عدد پردازنده گرافیکی، این مدلها با کمترین تنظیم دقیق و یا حتی بدون آن، در بسیاری از وظایف سطح بالای حوزه یادگیری ماشین به بهترین نتایج دست یافته‌اند.

[1] Foundation Models
[2] Large Language Models

یکی از اولین مدل‌های بنیادی حوزه بینایی ماشین، مدل CLIP است[3]. در مقاله CLIP محققان از ۴۰۰ میلیون جفت تصویر-توضیح موجود در اینترنت (شکل ۱)، برای آموزش توامان یک مدل استخراج ویژگی از تصویر (ResNet50) و یک مدل استخراج ویژگی از متن (Transformer) استفاده کرده‌است . نکته اصلی استفاده از این داده‌ها این است که جمع‌آوری آن‌ها نیاز به نظارت و هزینه بسیار اندکی دارد. ایده CLIP نیز بسیار ساده است: هر دو مدل تصویر و متن باید به ازای یک جفت تصویر-توضیح، بردار یکسانی استخراج کند.

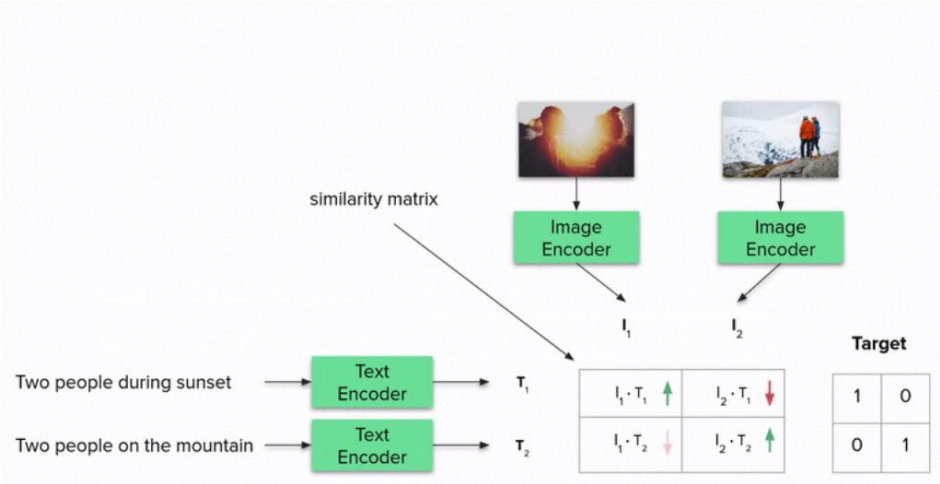

*شکل ۱- نمونه جفت عکس-توضیح موجود در اینترنت برای آموزش CLIP*

به طور دقیق‌تر فرض کنید یک دسته از نمونه‌های آموزشی شامل دو جفت تصویر-توضیح باشد(شکل ۳). در مرحله اول از دو تصویر و دو توضیح متناظر با آن‌ها به ترتیب با شبکه ResNet50 و Transformer ویژگی استخراج می‌شود. این ویژگی‌ها به وسیله یک لایه تماما متصل به یک فضا با ابعاد یکسان نگاشت می‌شود. از ضرب داخلی دو به دو ویژگی‌های استخراج شده از تصویرها و توضیح‌ها، یک ماتریس دو در دو به دست می‌آید که شباهت هر جفت ویژگی را نشان می‌دهد. از آنجایی که هر تصویر باید ویژگی متناظر با توضیح مربوطه‌اش را داشته باشد، آموزش این دو شبکه باید به گونه‌ای باشد که ماتریس شباهت به سمت ماتریس قطری همگرا شود. این کار به سادگی به وسیله یک تابع هزینه cross entropy انجام می‌شود (شکل ۲).

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

*شکل ۲- شبه‌کد آموزش CLIP*

الف) کد مربوط به آموزش CLIP برای دیتاست flickr8k به صورت ناقص در پیوست ارائه شده‌است. این کد را تکمیل و فرآیند آموزش را




*شکل ۲- فرآیند آموزش CLIP*

[3] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

کامل کنید.

ب) از یک جمله نمونه (مانند:This is a dog ) برای بازیابی عکس‌های مربوط به یک کلاس از دیتاست استفاده کنید. برای این کار با استفاده از شبکه Transformer برای این جمله بردار ویژگی استخراج کنید. این بردار را به فضای مشترک نگاشت کنید. سپس با مقایسه این بردار با بردار ویژگی تمام تصاویر دیتاست، شبیه‌ترین تصاویر به این جمله را نمایش دهید.

مدل اصلی CLIP روی ۴۰۰ میلیون عکس و چند صد کارت گرافیک به صورت موازی برای چند هفته آموزش دیده‌است. استخراج کننده ویژگی تصویر این مدل اما همچنان یک ResNet50 معمولی است! به علت حجم بالای داده‌های آموزشی، قدرت تعمیم پذیری این شبکه به شدت بالا است. از این شبکه می‌توانید بدون آموزش و با وزن‌های کاملا ثابت برای دسته‌بندی تصاویر در دیتاست‌های کاملا جدید استفاده کنید.

ج) از وزن‌های رسمی منتشر شده CLIP برای دسته‌بندی تصاویر دیتاست cifar10 استفاده کنید. برای این کار تمام وزن‌های شبکه کانولوشنی را ثابت نگه‌دارید و فقط یک لایه تمام متصلا روی آن آموزش دهید. (راهنمایی)