# پرسش ۲ – استفاده از Vision Transformer برای طبقه‌بندی تصاویر

در این پرسش با کاربرد تبدیل‌کننده‌ها در تصویر آشنا خواهید شد و مقاله‌ای را در این رابطه پیاده‌سازی خواهید کرد.

## ۲-۱. آشنایی با تبدیل‌کننده تصاویر

در توسعه‌های اخیر حوزه بینایی ماشین، افزایش محسوسی در استفاده از ساختارهای مبتنی بر ترنسفورمر مشاهده شده است. این ساختارها عملکردی بهتر از ساختارهای شبکه عصبی کانولوشنی (CNN) ارائه می‌دهند؛ اما از سوی دیگر، هزینه محاسباتی آنها برای آموزش از ابتدا بسیار زیاد است. از آنجایی که این مدل‌ها در حوزه بینایی ماشین به تازگی معرفی شده‌اند، نیاز به مطالعه قابلیت‌های یادگیری انتقالی آنها و مقایسه آن با CNNها وجود دارد تا بتوانیم ساختار مناسب‌تر را پیدا کرده و هنگام استفاده در مسائل واقعی با مجموعه‌داده‌های کوچک از آن‌ها استفاده کنیم.

این تبدیل‌کننده‌های تصویر با عملکرد بالا با استفاده از صدها میلیون تصویر به عنوان پیش‌آموزش، با یک زیرساخت بزرگ آموزش داده شده‌اند، که به همین دلیل توانایی استفاده مجدد از آن‌ها محدود شده است. مدل DeiT، تبدیل‌کننده‌ی بدون کانولوشن است که فقط با آموزش بر روی ImageNet ایجاد شده. مدل DeiT، از یک استراتژی معلم-شاگرد خاص برای تبدیل‌کننده‌ها استفاده می‌کند. این استراتژی بر یک توکن خلاصه‌سازی[1] تکیه می‌کند که مطمئن شود که شاگرد از طریق مکانیزم توجه از معلم یاد می‌گیرد. این روش خلاصه‌سازی مبتنی بر توکن، به خصوص زمانی که یک شبکه کانولوشنی به عنوان معلم استفاده می‌شود، بهتر عمل می‌کند. دسترسی به مقاله DeiT از طریق پیوند زیر ممکن است:

https://arxiv.org/abs/2012.12877

همانطور که می‌دانید یکی از روش‌های استفاده مجدد از مدل‌ها، fine-tuning است. در مقاله‌ی

Investigating Transfer Learning Capabilities of Vision Transformers and CNNs by Fine-Tuning a Single Trainable Block

روشی پیشنهاد شده که فقط با فاین-تیون کردن وزن‌های آخرین بلاک تبدیل کننده و MLP Head مدل، بتوان مدل را فاین-تیون کرد. از طریق پیوند زیر می‌توانید به مقاله ذکر شده دسترسی داشته باشید:

[1] Distillation

https://arxiv.org/ftp/arxiv/papers/2110/2110.05270.pdf

## ۲-۲. پیاده‌سازی و ارزیابی نتایج

در این بخش ابتدا به پیاده‌سازی مقاله، سپس ارزیابی نتایج خود خواهید پرداخت:

### ۲-۲-۱- لود کردن دیتاست و انجام پیش‌پردازش‌های لازم

**(۱۰ نمره)**

برای پیاده‌سازی این بخش در محیط گوگل کولب لازم است کتابخانه‌های transformers و datasets را با دستور pip install نصب کنید (این کتابخانه‌ها به صورت پیش‌فرض روی محیط کولب نصب نیستند).

حال دیتاست **CIFAR-10** را لود کرده و در صورت لزوم، پیش‌پردازش‌های ذکر شده در مقاله را انجام دهید.

### ۲-۲-۱- شبکه کانولوشنی

**(۳۰ نمره)**

پس از مطالعه‌ی مقاله‌ی بالا، یکی از مدل‌های تماما کانولوشنی را انتخاب کرده و با unfreeze کردن لایه‌های ذکر شده در مقاله، مدل را روی دیتاست **CIFAR-10** فاین-تیون کنید.

نتایج Validation Accuacy و Validation Loss را گزارش کنید.

### ۲-۲-۲- شبکه ViT (تبدیل‌کننده تصویر)

**(۶۰ نمره)**

یکی از مدل‌های تماما ترنسفورمری ذکر شده در مقاله را انتخاب کرده و با unfreeze کردن لایه‌های ذکر شده در مقاله، مدل را روی دیتاست **CIFAR-10** فاین-تیون کنید.

می‌توانید از آموزش‌های Hugging Face برای نحوه فاین-تیون کردن مدل تبدیل‌کننده تصویر خود استفاده کنید.

نتایج Validation Accuacy و Validation Loss را گزارش کنید. نتایج خود را با نتایج مقاله مقایسه کنید.