NOTES

> Purpose of report:
>
> xfolds:
>
> - give effective making techniques for longer sequence lengths (compare)
> - scaling law to give appropriately sized parameter model
> - to multiple adapters training or to single encoder decoder training (but this with longer sequence)
> - compare results with previous similarly sized models.

MedClip https://huggingface.co/flaviagiammarino/pubmed-clip-vit-base-patch32

PubMedCLIP was trained on the Radiology Objects in COntext (ROCO) dataset, a large-scale multimodal medical imaging dataset. The ROCO dataset includes diverse imaging modalities (such as X-Ray, MRI, ultrasound, fluoroscopy, etc.) from various human body regions (such as head, spine, chest, abdomen, etc.) captured from open-access PubMed articles.

PubMedCLIP was trained for 50 epochs with a batch size of 64 using the Adam optimizer with a learning rate of 10−5. The authors have released three different pre-trained models at this link which use ResNet-50, ResNet-50x4 and ViT32 as image encoders. This repository includes only the ViT32 variant of the PubMedCLIP model.

- Repository: PubMedCLIP Official GitHub Repository
- Paper: Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain?

- Current approaches for this multimodal task adopt deep neural encoders to interpret the image and the question and then pick a corresponding answer. They typically consist of four main components: a visual encoder, question encoder, attention-based fusion of vision and text features, and an answer classifier Vu et al. [2020], Zhan et al. [2020], Nguyen et al. [2019], Pan et al. [2021], Liu et al. [2021b].

> [Ideas]
>
> They already used 2 different Encoders embedding for text and images[specialized].
>
> so now we can just try to make 1 unified encoder and do ablation
>
> [we arent doing ultrasound] so would it be fair comparison. so maybe just use their tokenizer.?

other notes

- find benchmark datasets if you dont use ultraSound , flouroscopy

GPU

The experiments were conducted using
one 32 GB GPU (Nvidia DGX1 8x Tesla V100) in an OKD 4.6 cluster
under the Maastricht University Data Science Research Infrastructure

> To compute the performance in petaflop days (PF-days) for an Nvidia DGX-1 with 8x Tesla V100 GPUs, we can break it down as follows:
>
> 1. Understand PFLOPS for Nvidia DGX-1 with 8x V100:
>
> A Tesla V100 (32 GB) GPU can deliver up to 15.7 teraflops (TFLOPS) of single-precision (FP32) performance.

The DGX-1 has 8 V100 GPUs, so the total peak single-precision performance of the DGX-1 is:

Total TFLOPS=15.7×8=125.6 TFLOPS

Total TFLOPS=15.7×8=125.6 TFLOPS

Convert TFLOPS to petaflops (PFLOPS):

Total PFLOPS=125.61000=0.1256 PFLOPS

Total PFLOPS=1000125.6•=0.1256 PFLOPS

2. Time Calculation for PF-days:

1 PF-day is equal to 1 petaflop sustained for 1 day (i.e., 24 hours or 86,400 seconds).

Now, you can calculate how many PF-days the DGX-1 can achieve in a given time period. For 1 day, the DGX-1 can achieve:

PF-days=0.1256 PFLOPS×1 day=0.1256 PF-days/day

This means the Nvidia DGX-1 with 8x Tesla V100 GPUs can deliver 0.1256 petaflop-days of computation in one day.

from 4m

- All modalities are mapped to sets or sequences of discrete tokens (indices of a vocabulary) through the use of modality specific tokenizer
- captions and bounding boxes are both treated as text and tokenized by wordpiece

- Encoder is a std transformer `but features modality-specific learnable input embedding layers to map token indices to vectors` .
- Non learnable sin-cos position embedding.

- The decoder handles tokens from both dense image-like and sequence-like modalities, with each type requiring a different approach