Montreal, July 8–12

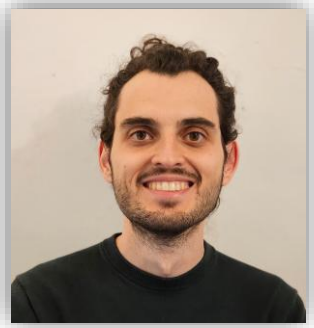
DLMI2024

Deep learning for medical imaging school

Hands-on session

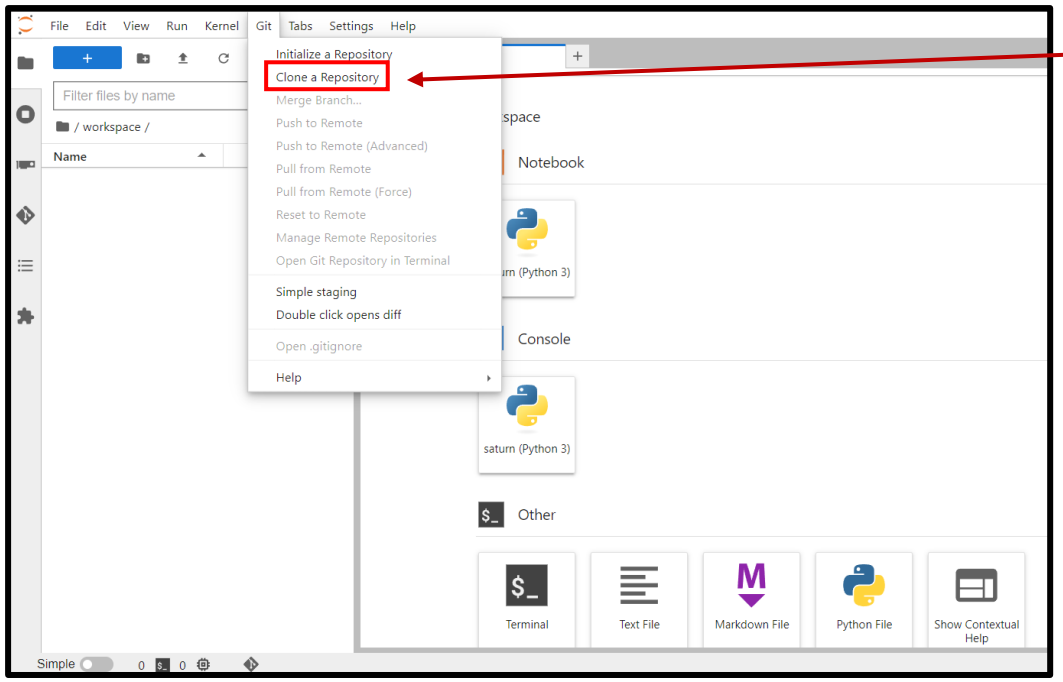
Foundation Models

Julio Silva-Rodríguez

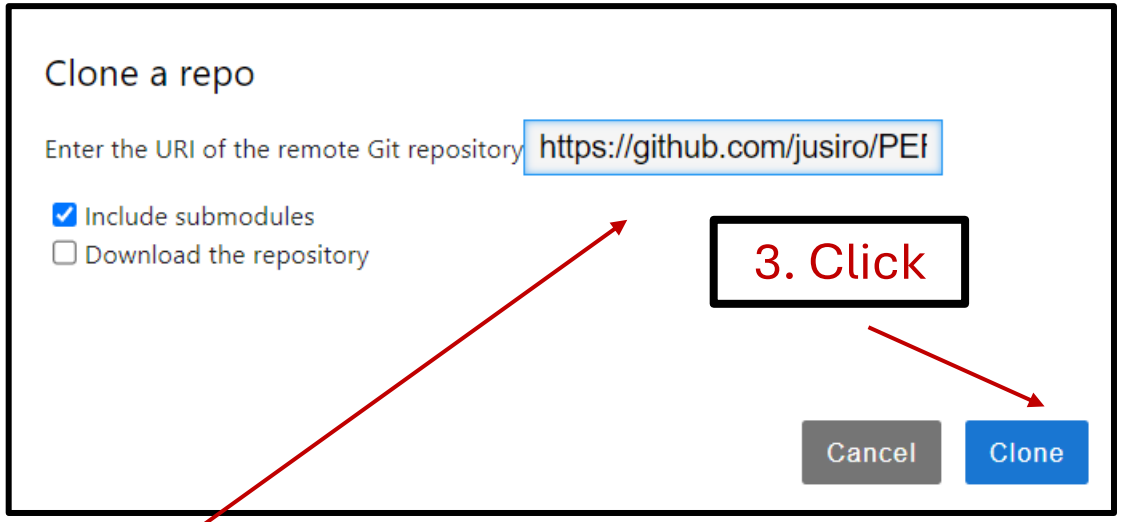


Jose Dolz

Code Download



1. Git - Clone

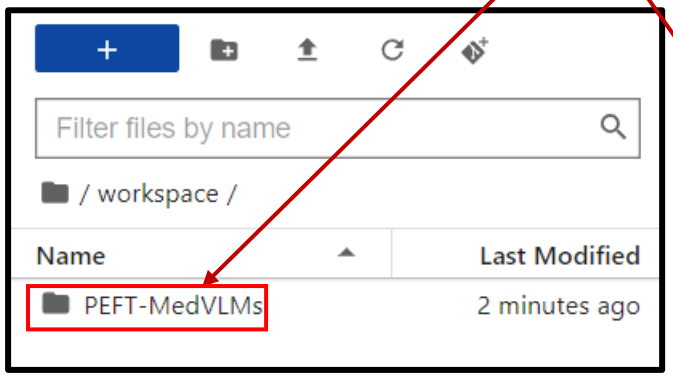


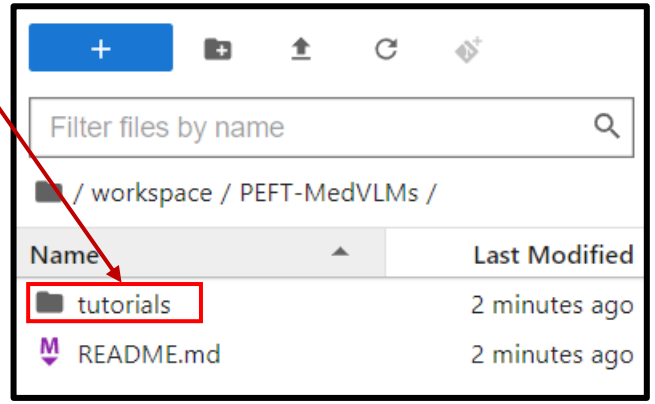
3. Click

2. Add URL

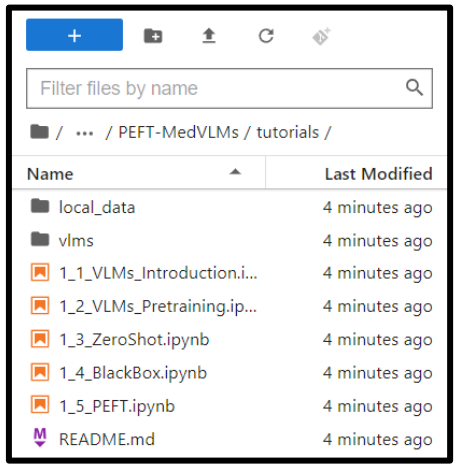
https://github.com/jusiro/PEFT-MedVLMs

4. Click





Ready to start!



Datasets

1. Go to the link

5 minutes

https://data.mendeley.com/datasets/9xxm58dvs3/2

Mendeley Data

SICAPv2 - Prostate Whole Slide Images with Gleason Grades Annotations

Published: 22 October 2020 | Version 2 | DOI: 10.17632/9xxm58dvs3.2

Contributor: Julio Silva-Rodríguez

Description

A database containing prostate histology whole slide images with both annotations of global Gleason scores and path-level Gleason grades.

Data associated with the paper:

Silva-Rodríguez, J., Colomer, A., Sales, M. A., Molina, R., & Naranjo, V. (2020). Going deeper through the Gleason scoring scale : An automatic end-to-end system for histology prostate grading and cribriform pattern detection. Computer Methods and Programs in Biomedicine, 195. https://doi.org/10.1016/j.cmpb.2020.105637

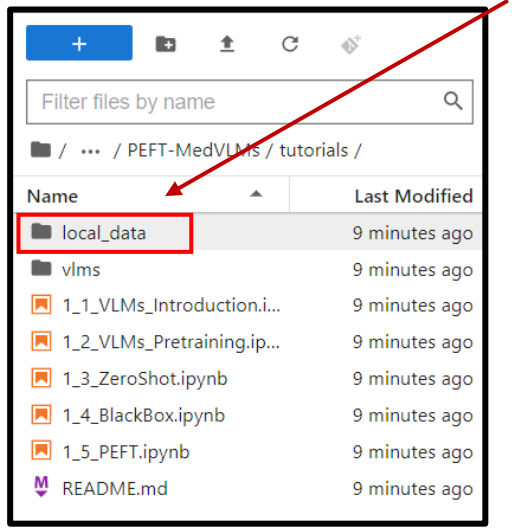
Download All 2.01 GB

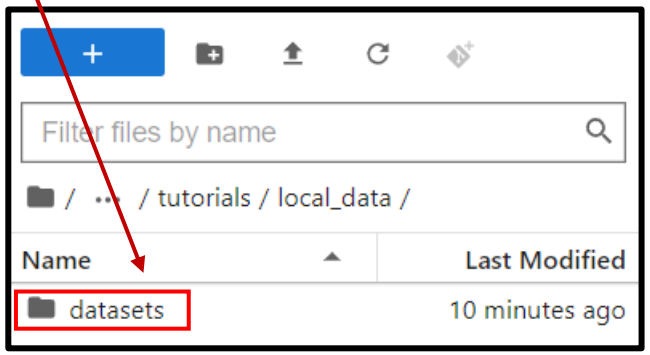
2. Download

Files

SICAPv2.zip 2.01 GB

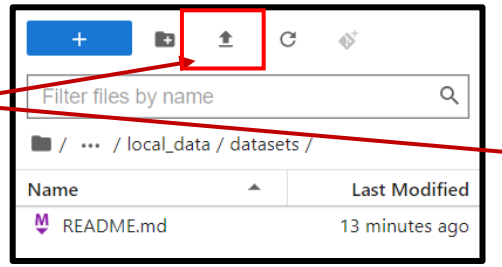
3. Go to dataset folder

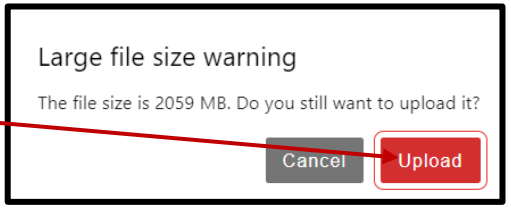




4. Upload

SICAPv2

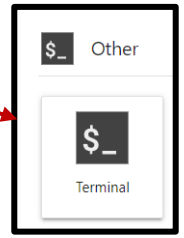




Uploading...

5 minutes

5. Unzip

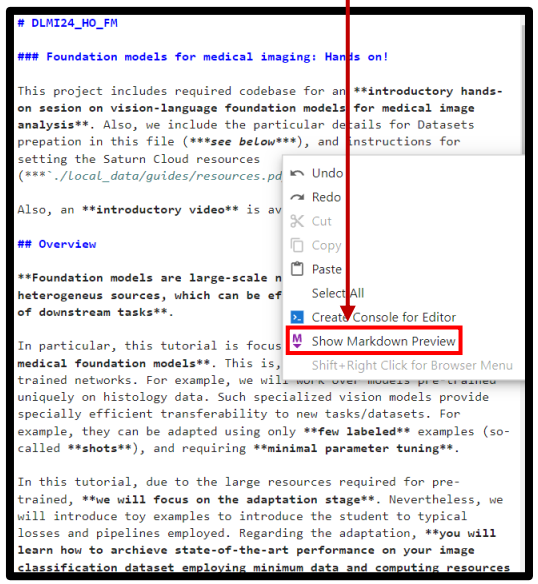


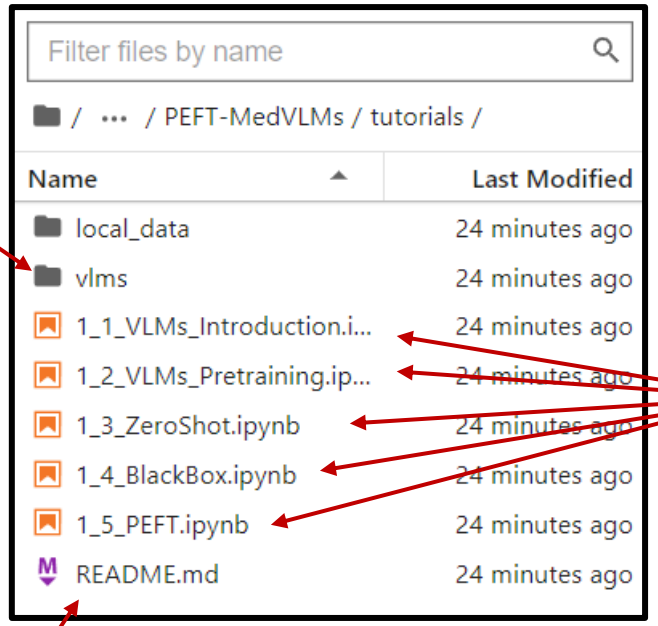

```
~/workspace/PEFT-MedVLMs/tutorials/local_data/datasets$ unzip SICAPv2.zip
```

HANDS-ON

3. Auxiliary functions

Right click in the README + “Show…”



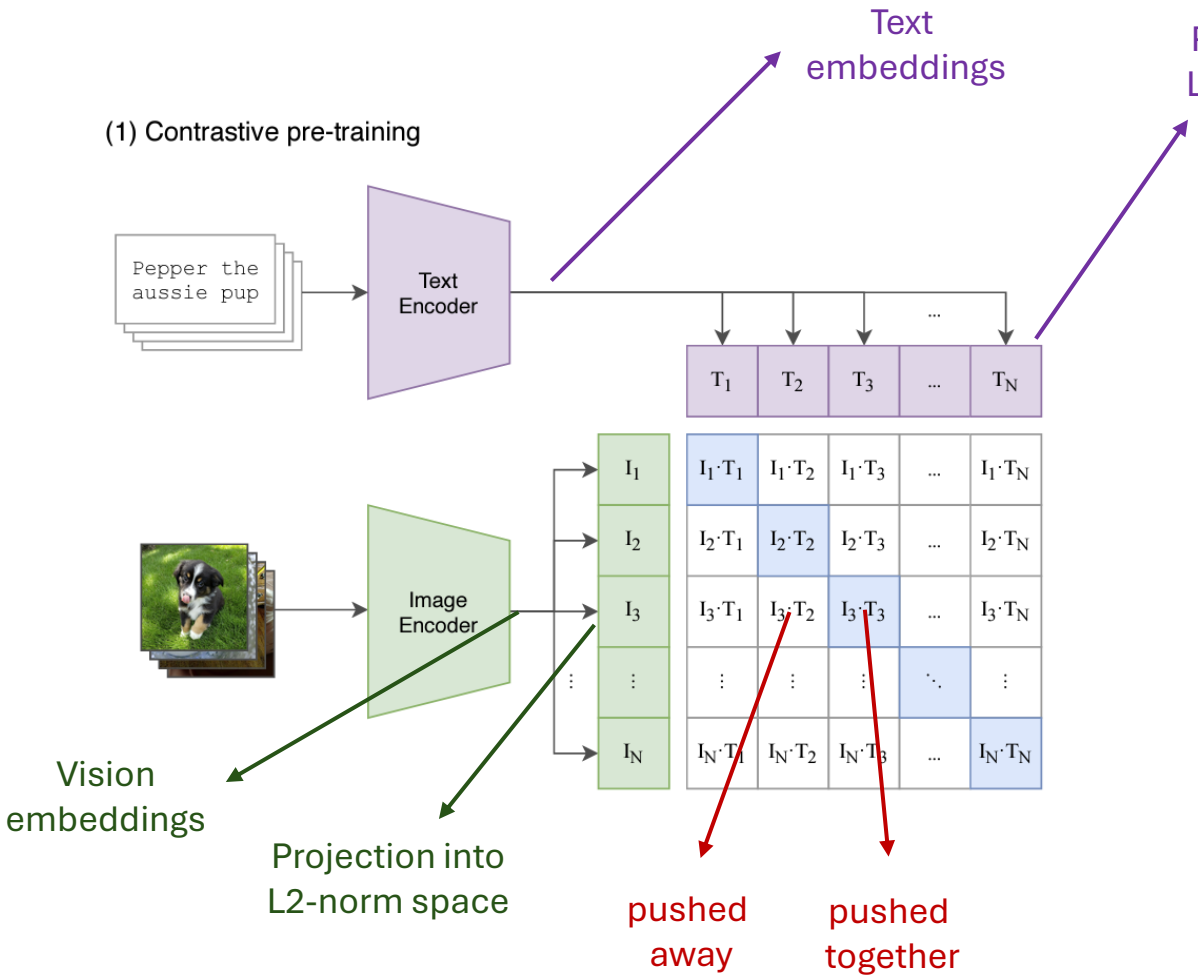


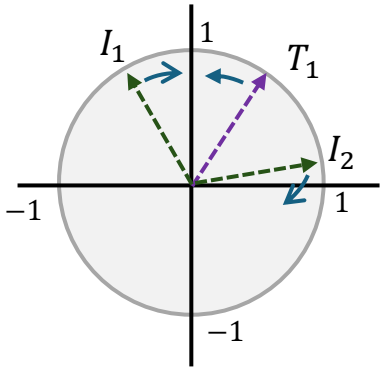
2. Tutorials

1. README

Introduction to Vision-Language Models and PEFT

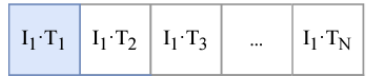
CONTRASTIVE VISION-LANGUAE PRE-TRAINING (CLIP)





Dot product between l2-norm vectors is cosine similarity (angle between vectors)

Softmax + Cross-Entropy (over rows)



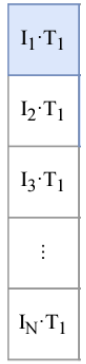
CLIP Loss

- The image-to-text contrastive loss

$$\mathcal{L}_{i2t} = -\sum_{i\in\mathcal{B}} \log \frac{\exp(\tau \boldsymbol{u}_i \boldsymbol{v}_i)}{\sum_{j\in\mathcal{B}} \exp(\tau \boldsymbol{u}_i \boldsymbol{v}_j)} \quad (5)$$

- The text-to-image contrastive loss

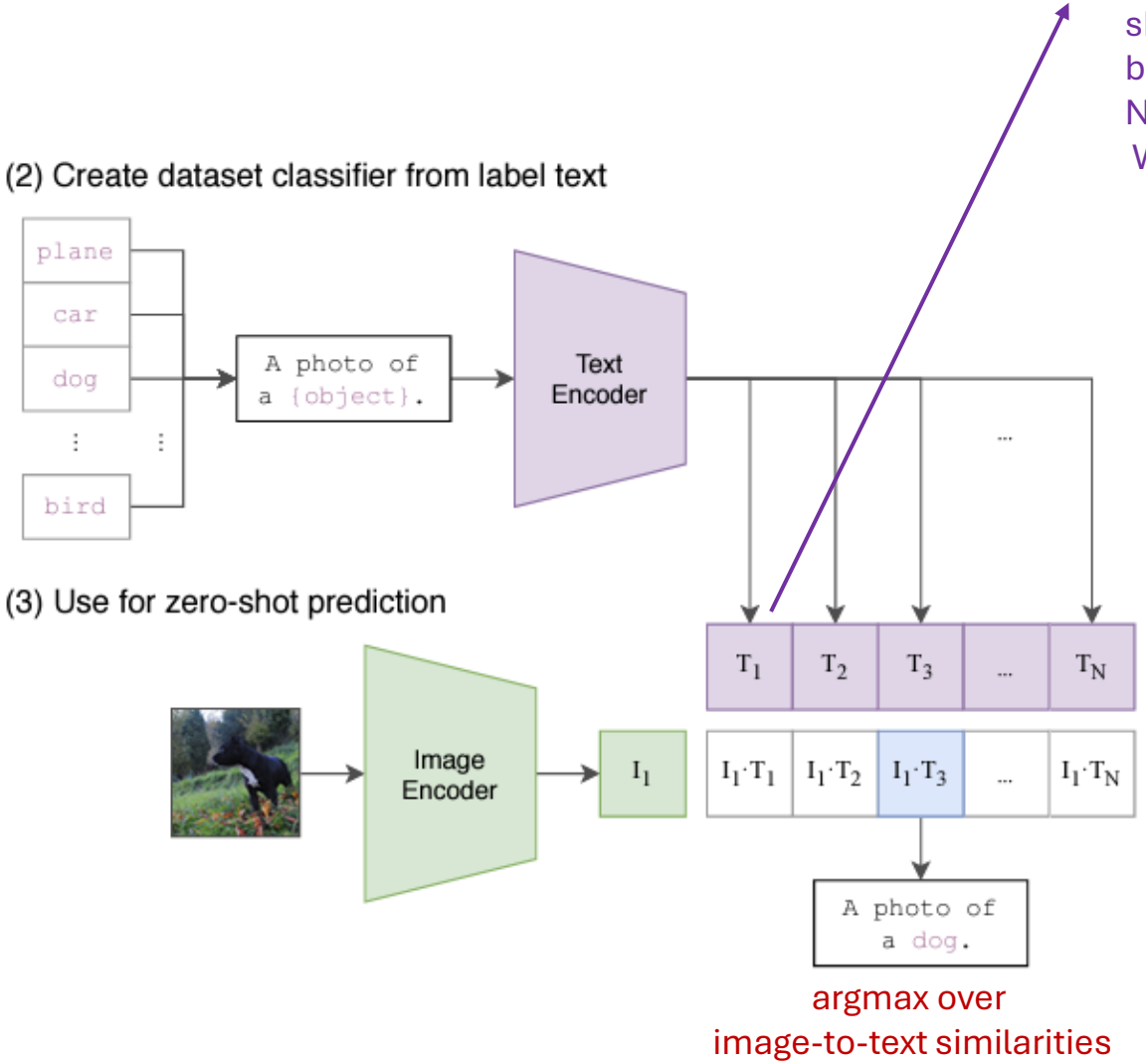
$$\mathcal{L}_{t2i} = -\sum_{j\in\mathcal{B}} \log \frac{\exp(\tau \boldsymbol{u}_j \boldsymbol{v}_j)}{\sum_{i\in\mathcal{B}} \exp(\tau \boldsymbol{u}_i \boldsymbol{v}_j)} \quad (6)$$



Softmax + Cross-Entropy (over columns)

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning.

ZERO-SHOT PREDICTIONS

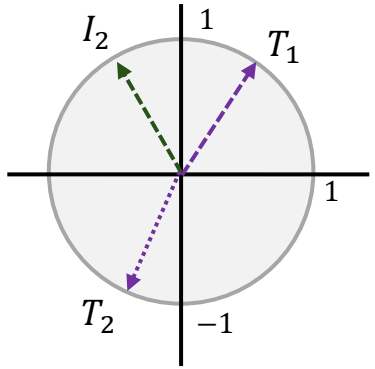


argmax over
image-to-text similarities

Note: Text embeddings for target categories are also called **class prototypes**, or **zero-shot prototypes**.
They do not require image samples to compute this reference embedding, but only text, and that is why they are called "zero-shot". Images with similar representations will be more likely to belong to this category.
Note that they are equivalent to a Linear output layer!
W (classes, features).

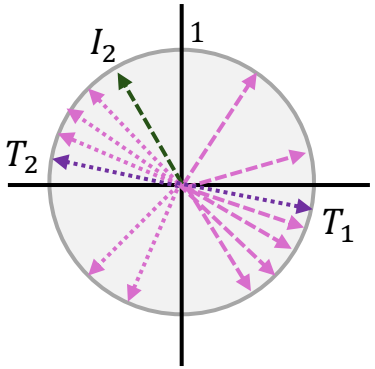
A. Single Prompt

"a photo of [CLS]"



B.Prompt Ensemble

"a photo of [CLS]"
"a sketch of [CLS]"
"small animal with black hair"
"has four legs"
"two animals playing"
…

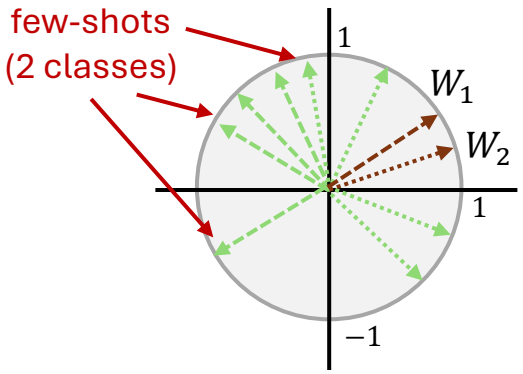


Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning.

BLACK-BOX ADAPTERS

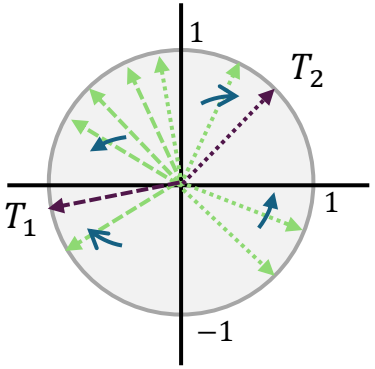
- Work over pre-computed vision features - They are backbone-agnostic.
- May profit zero-shot prototypes for the target tasks.
- They are backbone-agnostic.
- Very efficient, do not even require GPU.
- Potentially, they do not require access to pre-trained weights (similar to ChatGPT).

A. Linear Probing



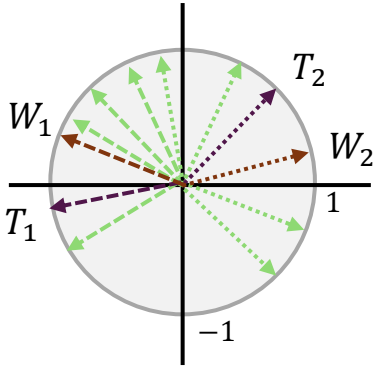
Train W1,W2 to
minimize cross-entropy
We train class prototypes

B. CLIP-Adapter



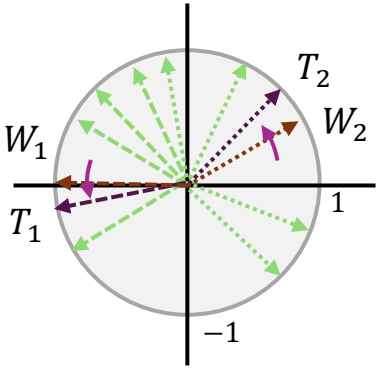
Train an mlp to modify vision features
keep class prototypes as zero-shot

C. ZS-Linear Probe



Train W1,W2 to
minimize cross-entropy
Initialize then to zero-shot

D. Class-Adaptive Linear Probe (CLAP)



Train W1,W2 to
minimize cross-entropy
Constraint them to remail close to zero-shot

Gao, P., Geng, S., Zhang, R. et al. (2024). CLIP-Adapter: Better Vision-Language Models with Feature Adapters. Int J Comput Vis.
Silva-Rodriguez, J., Hajimiri, S., Ben Ayed, I., Dolz, J. (2024). A Closer Look at the Few-Shot Adaptation of Large Vision-Language Models. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

PARAMETER-EFFICIENT FINE-TUNING

- Tran a subset of parameters to modify deep features.
- Two types: selective, and additive.
- More efficient than full-finetuning, and more flexible than black-box Adapters.
- If carefully designed, they can avoid catastrophic forgetting.

A. Affine-Layer Norm

$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta$$

We only tune these from the whole encoder

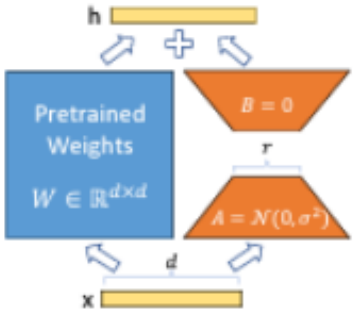
B. Bias Tuning

$$y = xA^T + b.$$

$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta$$

C. Low-Rank Adapters

Usually applied in ViTs to k,q,v layers of MultiHeadAttention



Add and tune a residual connection with low-rank weights. Important!
Note B=0 when t=0

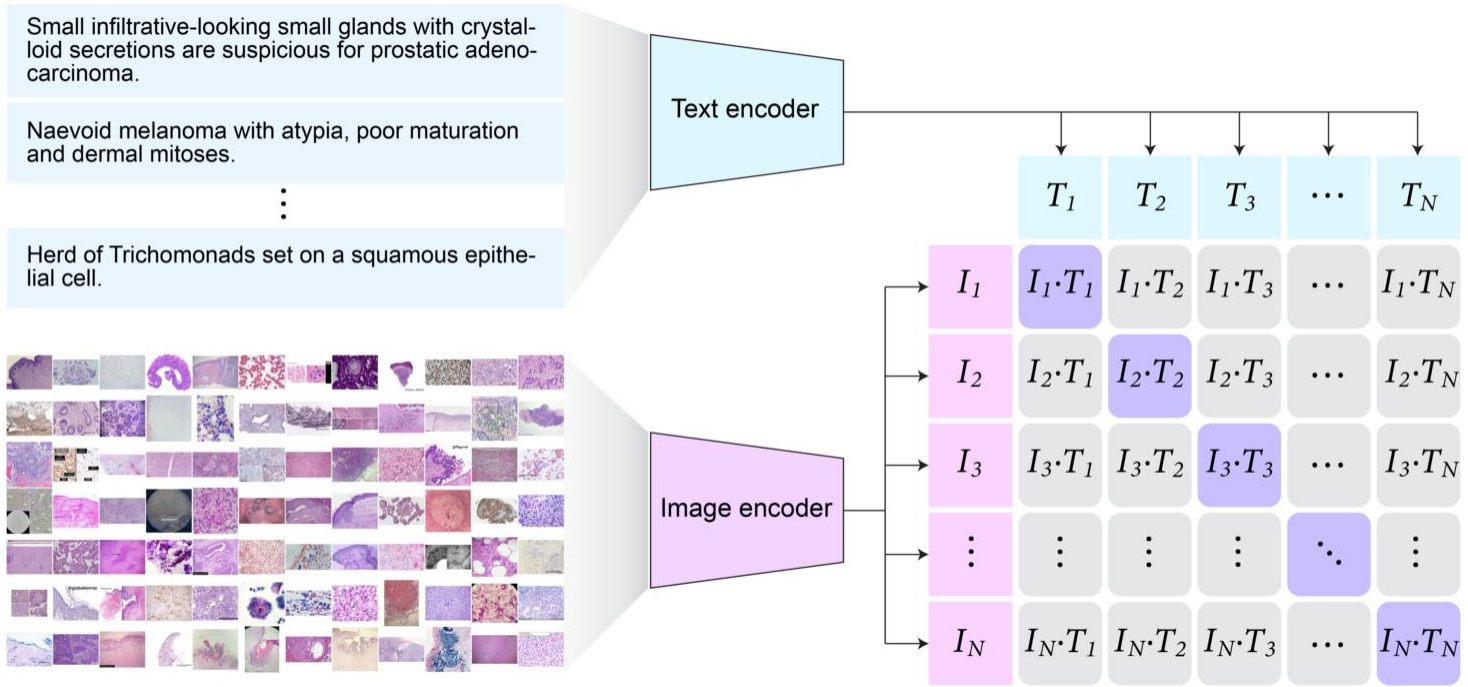
Frankle, J., Schwab, D. J., Morcos, A. S. (2021). Training batchnorm and only batchnorm: On the expressive power of random features in cnns. International Conference on Learning Representations (ICLR).

Ben-Zaken, E., Ravfogel, S., Goldberg, Y. (2021). Bitfit: Simple parameter efficient fine-tuning for transformer-based masked language-models. Association for Computational Linguistics.

Hu, E. J., et al., (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR).

MEDICAL VLMs - PLIP

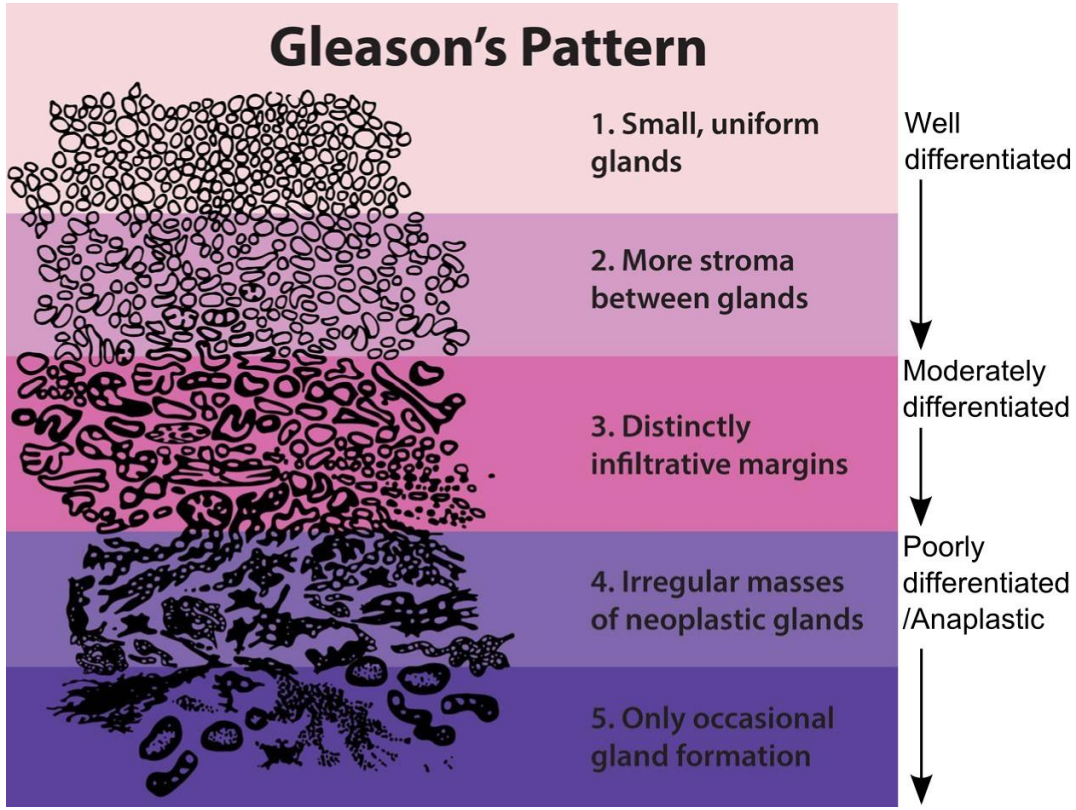
Twitter data!



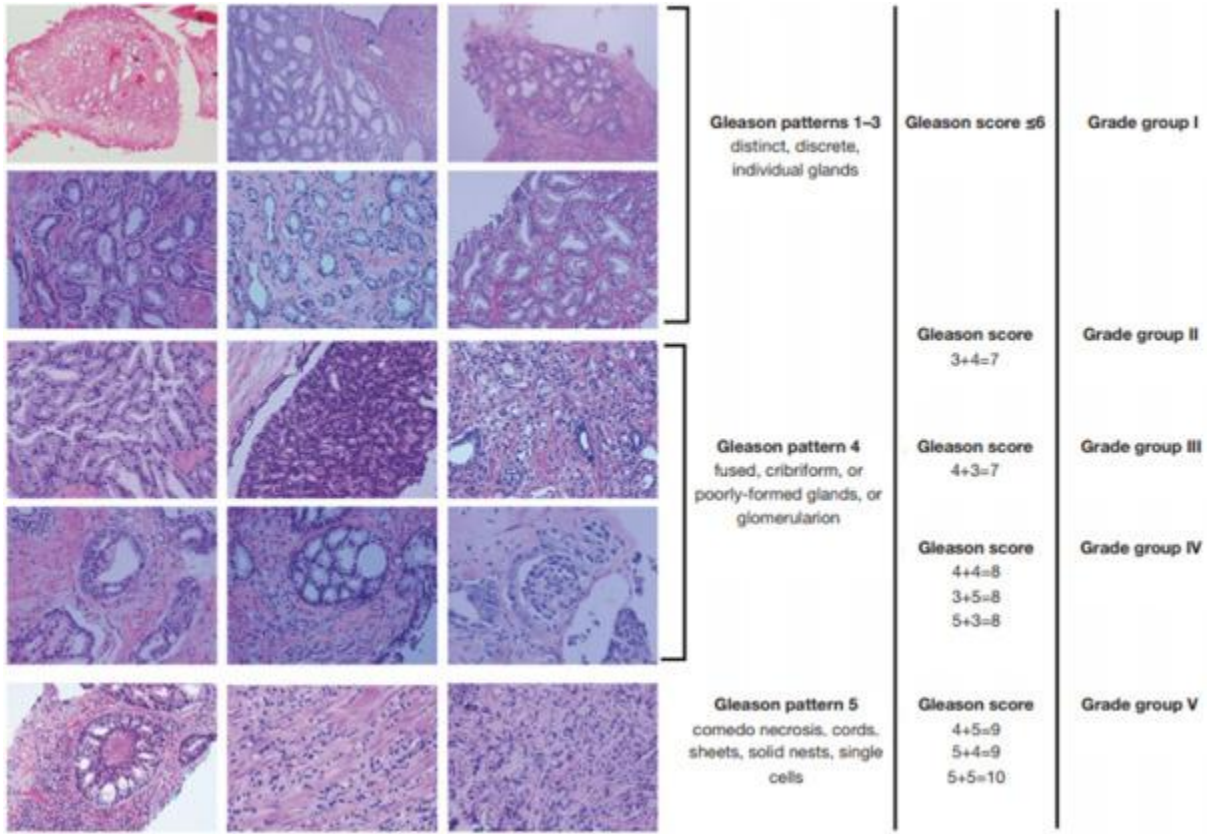
Specialized on histology images

Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. (2023). A visual-language foundation model for pathology image analysis using medical Twitter. Nature Medicine.

APPLICATION: GLEASON GRADING



Gleason grading system - Wikipedia



Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study. Informatics in Medicine Unlocked.

Thanks!