BIODS 276 / CS 286 Advanced Topics in Computer Vision and Biomedicine
Fall 2024

Assignment 1 - Extending Representation Learning Methods

In this assignment, you will write a 1500-2000 word analysis paper, using the NeurIPS LaTeX template. The goal of the assignment is to synthesize and think critically about the topics covered in class, through reading and analyzing recent papers. The assignment must be completed individually; you may not work in groups.

Please select one of the following analysis categories: "Extending representation learning to more complex modality setups" or "Diving into representation learning in a biomedical application", and choose two research papers in that category based on the provided guidelines. This course has two analysis assignments, one focused on representation learning models and the other on generative models. Both assignments will have the same analysis categories, and whichever category you choose for this assignment, you must choose the other for the second assignment.

Option 1. Extending representation learning to more complex modality setups:
In lecture, we have focused on vision-only and vision-language *representation learning* in a "standard" setting: 2D images are in a standard resolution (less than 1000x1000 pixels), and the text is a short caption string (less than 50 words). So now, dive deeply into representation learning outside the standard setting, such as non-standard 2d images or vision paired with other (non-language) modalities. Some examples include video, 3D vision-only, high-resolution 2d image-only, joint image-audio, joint video-language, etc.

Choose one such setting and identify two research papers that propose approaches for your selected modality setup. The key requirement is that your chosen modality setup should not be in the 'standard setting', but should still involve visual pixel data.

Option 2. Diving into representation learning in a biomedical application:
We have surveyed examples of representation learning in different biomedical contexts. Now, dive more deeply into representation learning for a specific biomedical application or use case of your choice. Identify two research papers that propose representation learning approaches for your chosen application, using large vision-only or multimodal (vision-X) models. These papers may be ones discussed in class or others you discover independently. You are also not limited to 2D image or image-text models. For example, you may select more complex modality setups such as those described in the "Representation learning beyond 2D images" prompt.

Prompt

Given the two selected papers in whichever category you have chosen for this assignment, compare and contrast their approaches, answering these key questions:

1. Briefly give an introduction to the context of your setting and its motivation. Explain why you chose your 2 papers.

2. What are the most significant downstream tasks that use the representations?
3. Describe the data and what is distinct about it in this context:
    a. For option 1, explain how the data modality of choice has been represented in deep learning, including in your chosen papers as well as in prior work (if applicable). For example, images are represented as pixel values in CNNs, and patches of pixels in vision transformers.
    b. For option 2, explain what is challenging about the data content compared to natural images. For example in lung tumor classification in X-ray, the classification of healthy / disease can depend on a very small part of the image, while for cat/dog classification, that is not the case.
4. For each paper, how do they motivate their methodology? Put another way, what is their key insight?
5. Describe the strengths and weaknesses of each paper.
6. If one paper is stronger than another, why? If it is not clear that one paper is stronger than the other, why?
7. Summarize and synthesize the key conclusions of your analysis.
8. What are important or promising directions for future work?

You may reference additional papers if they help support an argument, but the two chosen papers should have primary focus for the comparative analysis. Your reports will be assessed based on whether you address this list of questions.

Guidelines:
The analysis paper should be written in the format of standard machine learning papers, abridged as stated following to adapt to the assignment goals. The paper should contain the following sections:

- **Prompt Choice, Selected Two Papers (write out the full citation)**
- **(10%) Introduction (~0.5 page).** Prompt question 1.
- **(20%) Problem background (~0.5-1page)** Prompt questions 2-3.
- **(50%) Analysis (2-3 pages).** Prompt questions 4-6. This should be by far the most significant section of your paper. Feel free to break it down into further subsections if you like. Please reference the prompt questions for further details.
- **(20%) Conclusion and future work (0.5 pages).** Prompt questions 7-8.

The report should be 1500-2000 words and in the Neurips latex format (see here and here), not including references. You are encouraged to include figures.

Choosing two papers
Below are suggested papers, but you are welcome to select your own. If you choose your own paper, please make a private post on Ed and tell us the papers so that we can check that it's in scope. We'll get back to you within 48hrs, please plan accordingly with respect to the assignment deadline.

You can use papers that were used in class if they were only reviewed briefly (4 total slides or less).

You should focus on methods for representation learning. Note that a lot of papers may include a representation learning phase, followed by some other phase. You can include those papers, but your analysis should emphasize the representation learning parts.

Suggested topics & papers for “Extending to more complex modality setups”:

- Video
    - “Is Space-Time Attention All You Need for Video Understanding?”
    - “VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training”
    - "Bevt: Bert pretraining of video transformers"
    - “Revisiting Feature Prediction for Learning Visual Representations from Video” (V-JEPA)
- Video-language
    - “VideoPrism: A Foundational Visual Encoder for Video Understanding”
    - “InternVideo: General Video Foundation Models via Generative and Discriminative Learning”
- High-res 2d images: whole-slide images in tissue pathology
    - “A whole-slide foundation model for digital pathology from real-world data”
    - “Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning”
- Multiple modalities simultaneously:
    - “ImageBind: One Embedding Space To Bind Them All”
    - “VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text””
- 3D vision
    - “SimCVD: Simple Contrastive Voxel-Wise Representation Distillation for Semi-Supervised Medical Image Segmentation”
    - “Contrast with Reconstruct: Contrastive 3D Representation Learning Guided by Generative Pretraining”

Other setting could include: 3D voxel, video-audio, 3D meshes, human keypoints

Suggested topics & papers for “Diving into a biomedical application”:

- General biomedical representation learners
    - “BiomedCLIP: a multimodal biomedical foundation model pre-trained from fifteen million scientific image-text pairs”
    - “Making the most of text semantics to improve biomedical vision–language processing”
    - “GLoRIA: A Multimodal Global-Local Representation Learning Framework forLabel-efficient Medical Image Recognition”

- Surgical video
    - “A vision transformer for decoding surgeon activity from surgical videos”
    - “General surgery vision transformer: A video pre-trained foundation model for general surgery”
- Microscopy: sub-cellular
    - “Self-supervised deep learning encodes high-resolution features of protein subcellular localization”
    - “Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting.”
- Chest Xray
    - “Learning to exploit temporal structure for biomedical vision-language processing”
    - “Multi-granularity cross-modal alignment for generalized medical visual representation learning”
- Histopathology / tissue
    - “A visual--language foundation model for pathology image analysis using medical Twitter”
    - “A visual-language foundation model for computational pathology”
- Histopathology and spatial transcriptomics
    - “Multimodal contrastive learning for spatial gene expression prediction using histology images”
    - “Accurate Spatial Gene Expression Prediction by Integrating Multi-Resolution Features”