Project 5 : CLIP-Based Freestyle Project: The Power of Vision and Language CS 323

November 12, 2023

What you will do

CLIP (Contrastive Language-Image Pretraining) by Radford et al. (2021) represents a groundbreaking advancement in Vision-Language understanding, significantly impacting the field. CLIP has also found applications that extend into diverse domains such as video, audio, 3D, and unstructured data, surpassing expectations and pushing the boundaries of interdisciplinary applications. In this project, you are tasked to explore the capabilities of CLIP and unleash your creativity by designing a project that demonstrates the applicability of CLIP. Unlike previous projects, this project is freestyle. This means that you get to design the project on your own. This also means no template code will be provided. We have listed some suggestions about the broad topics to work on but you are free to choose any task/topic you want to design and work on.

FAQs

What is the primary requirement for project selection?

Your project must be CLIP-based. Your project should prominently feature the utilization of the pre-trained CLIP model, showcasing its capabilities in addressing the chosen problem or creative concept.

Can we collaborate with other classmates?

No

What are we required to submit?

A project report (pdf) and the supporting code implementation. You will be needed to present your project and you will be asked thorough questions about what/how you did.

How complex does the project need to be?

Use notebooks from previous projects as your guide of what your project should entail and how it should be structured. You are encouraged to keep the complexity of the project relatively similar to your previous projects. You are not expected to produce a novel revolutionary vision-language understanding work. The project can be an implementation of a well-known idea or a particular paper.

What are some suggested areas to look into?

Some suggested areas to look into are *Zero-shot image segmentation*, *Zero-shot 3D segmentation*, *Zero-shot object detection*, and *CLIP-based art generation*. Of course, these are not comprehensive and you are free to work on any CLIP-based project you want. Simplest yet effective ideas are often the greatest.

How should we structure our project report? What are some important components to include?

Your project report should clearly indicate the problem statement and include details about the solution you constructed/implemented and its evaluation. Evaluation is very important. Use the metrics often used in recent papers related to the problem you work on. Qualitative results are also important. You must provide examples of the results you achieved. You *must* list details about the challenges faced and insights gained during the implementation. Ensure comprehensive documentation.

What is the zero-tolerance plagiarism policy?

You are expected to implement the main component of your project on your own. For some minor components, for example, metrics computation, you are allowed to re-use the code from online sources. However, all the core parts must be implemented by yourself. In case any plagiarism is detected, for example, copy-pasting of code from GitHub, a blog post, from your classmate, or any other source, you will be graded ZERO and flagged for plagiarism to KAUST administration. Referencing and taking inspiration is not plagiarism. You are in fact encouraged to do research and take ideas from online sources. However, you should list down ALL the resources you referenced in your report. Any missing reference will be treated as a plagiarism source.

Do you have any recommendations on where to look for ideas/papers/resources?

Use arXiv, Google Scholar, and publications from previous conferences like CVPR, NeurIPS, ICCV, and ECCV. Search for queries like 'CLIP based', 'Unsupervised CLIP', 'Zero shot CLIP', 'CLIP segmentation' etc. One of the good

surveys on CLIP is by Zhang et al. (2023). X (twitter) can be a surprisingly good source of ideas as well.

Huggingface hosts a pre-trained CLIP model (`https://huggingface.co/docs/transformers/model_doc/clip`)

Evaluation Criteria

- Effective use of CLIP
- Technical implementation and correctness
- Documentation of process/experience, code, experiments performed, evaluation, and results.
- Presentation (report and interview)

References

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2023.