

Title of Project:

(English)	Promoting Zero-Shot Learning Ability for Vision Language Model
-----------	--

(Chinese)	提升視覺語言模型的零樣本學習能力
-----------	------------------

Vision language models (VLMs) like CLIP have shown incredible zero-shot learning ability. For a new learning task, the pre-trained VLMs can perform well without being fine-tuned. Specifically, in the new task, VLMs usually first generate pseudo-labels for samples and then regard those pseudo-labels as supervision signal to guide the prompt learning to make the input suitable for the VLMs, such that the VLMs achieve zero-shot learning on this new task. Although this pipeline has demonstrated effectiveness, it suffers from the following **three major limitations**. **First**, the pseudo-labels generated by the VLMs inevitably have labeling errors as the training domain and the downstream domain in the new task is different. The error pseudo-label will degrade the zero-shot learning ability of VLMs. **Second**, VLMs consist of multiple modalities, and the multimodality gap in the VLMs hinders the reliability of the constructed zero-shot learning model. **Finally**, the current pipeline is sensitive to the adopted prompt.

In this project, we aim to promote the zero-shot learning ability of VLMs. **(1) We will improve the pseudo-labeling strategy.** Specifically, instead of selecting a single pseudo label for each instance, we propose to choose a set of candidate pseudo labels for each instance. This strategy makes the correct label more likely to be assigned to the instance but will also bring additional label ambiguity. Therefore, we will formulate this strategy as a multi-objective optimization problem that reduces the candidate label set size and improves the quality of the candidate label simultaneously. Although this strategy can enhance the pseudo-label quality, the ground-truth label may still be located outside the candidate label set (OCLS); we will further investigate how to solve this OCLS problem. **(2) We will try to reduce the modality gap in VLMs.** First, we will formulate the representation of the embeddings in different modalities as a distribution rather than a point to enhance the robustness of the representation. Then, we will construct a modality-wise classifier to communicate with each other to achieve mutual learning. As different modalities may be skilled in predicting different instances, we will propose an uncertainty-aware classifier fusion strategy to achieve better prediction. **(3) We will investigate a more effective and robust zero-shot learning paradigm for VLMs.** First, to reduce the influence of pseudo labels, we will leverage self-paced learning to first exploit the easy samples and then include the harder ones to make the zero-shot model more robust. Then, we will ameliorate prompt learning scheme to make the zero-shot model more robust and stable.

This project will benefit the fundamental research on visual language models and machine learning. It will highly reduce the annotation cost for many industrial applications like healthcare and automotive. It will also help local community life related to visual language models like healthcare and education.