Chain of Thought Prompt Tuning for Vision-Language Models

Overview:

The study addresses the inability of vision-language models to generalize unseen visual concepts, mimicking the quick adaptation of the human brain. Motivated by the intention to emulate a human's chain-of-thought reasoning method in solving new tasks, particularly in the vision domain, the authors designed an innovative framework. This framework is said to have effectively modeled the reasoning process and has tested substantially on tasks like image classification, image-text retrieval, and visual question answering. The research contributions claim a successful adaptation of chain of thought in language models in vision domain, in addition, the framework's superiority was also proved through various ablation experiments. An overall improvement in performance on several tasks was evident through the experimental results from 18 different datasets. The ablation study detailed the role of different chain lengths in image classification and deduced that a chain length of 3 yielded the optimal result. Despite the results, the study acknowledges the variable efficiency of the chain of prompts across different tasks and datasets as a limitation.

Detailed Method:

The section illustrates a method to improve image recognition processes through a chain of prompt representations. This approach, comprising Prompt Chaining, Self Adaptive Chain Controller, and Meta-Nets Chaining, primarily aims to break down the understanding of an image content into a sequential reasoning process.

In Prompt Chaining, a set of prompts are connected in a chain where each prompt is fed from the previous one and passes information to the next, simulating a human-like step-by-step cognitive reasoning process. This approach also aids the model to deduct and infer new visual concepts making it more robust for tackling new scenarios as shown in Figure 2.

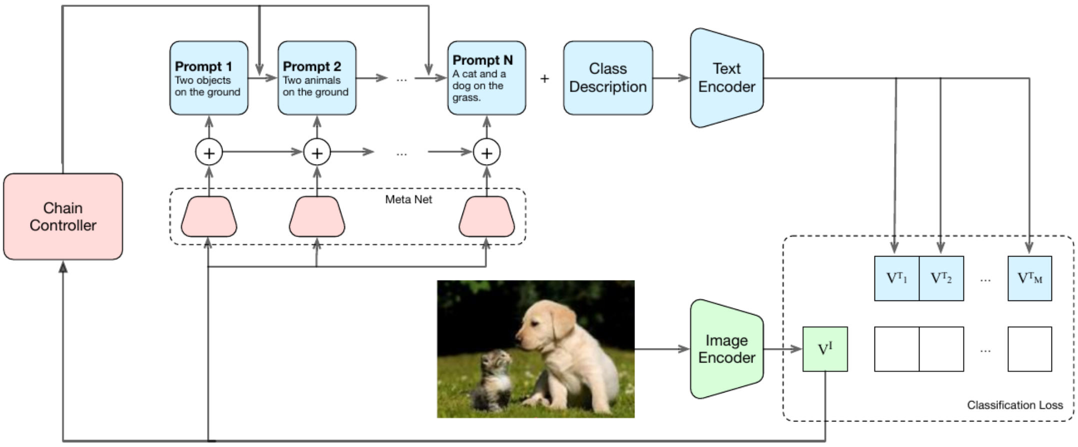

Figure 2. The multi-modal chain of thought prompt tuning framework. We build a series of chained prompts and a sequence of Meta-Nets. We also adopt a dynamic chain controller to control the weights on the chain based on the inputs. We use the final prompt (prompt N) for prediction.

.

Understanding that different image tasks may require various forms of reasoning, the Self Adaptive Chain Controller dynamically controls the chain dependant on the input. A light-weighted chain controller with a straightforward structure is utilized to adapt to various scenarios and maintain stability.

Lastly, adopting the concept of Meta-Net from CoCoOp, a chain of networks is designed in Meta-Nets Chaining. This chain models a step-by-step reasoning process where a different subset of visual information is utilized at each reasoning step with a meta-net, allowing better preservation of information and avoiding gradient vanishing. Each visual representation also acts as a bias term to help adjust the word embeddings of prompt before being fed into the text encoder.

To sum up, this method improves image recognition through a sequence of prompt representations which mimic human cognitive reasoning process and hence, retain an inherent reasoning ability for new scenarios, dynamically control the chain, and utilize different visual representations at each step to strengthen the stability of the architecture.