IPO: Interpretable Prompt Optimization for Vision-Language Models

Yingjun Du [1 *] , **Wenfang Sun** [2 *] , Cees G.M. Snoek [1]

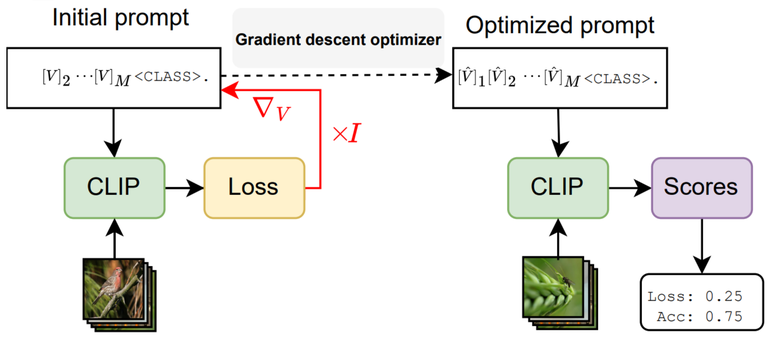
[1]University of Amsterdam, [2] University of Science and Technology of China

Prompt learning for VLM

Existing prompt optimization methods use gradient descent to learn adjustable prompts.

😭 Overfits to training classes

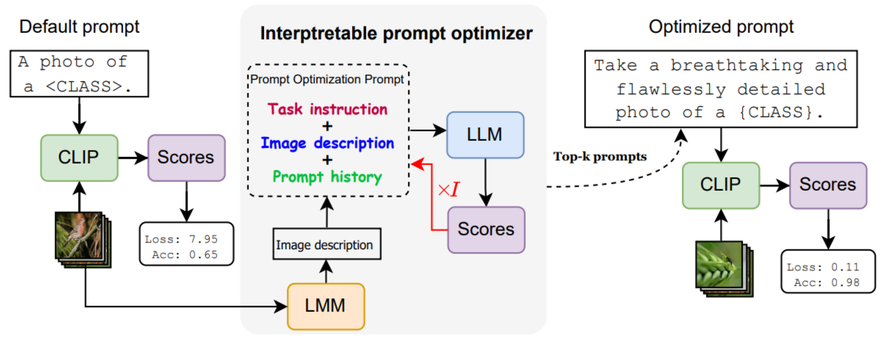
😭 Learned prompts are not human-interpretable



Gradient-based prompt optimization

IPO

- **IPO uses large language models (LLMs) to dynamically generate textual prompts.**
- **IPO introduces a Prompt Optimization Prompt to guide LLMs and stores historical prompts with performance metrics.**
- **A large multimodal model (LMM) generates image descriptions to improve the interaction between text and visual modalities.**



Interpretable prompt optimization

Experiments

- ***Comparison with SOTA***

ViT-B/16	Base	Novel	H
CLIP	69.34	74.22	71.70
CoOp	72.08	66.71	69.29
CoCoOp	72.85	72.17	72.51
MaPLe	70.85	71.57	71.21
PromptSRC	**73.38**	71.47	72.41
CoPrompt	70.44	70.11	70.27
IPO	71.76	**77.00**	**74.29**

(a) Average over 11 datasets.

ViT-B/16	Base	Novel	H
CLIP	72.43	68.14	70.22
CoOp	73.20	67.43	70.20
CoCoOp	73.90	69.07	71.40
MaPLe	74.03	68.73	71.28
PromptSRC	73.27	68.87	71.00
CoPrompt	73.97	**70.87**	**72.39**
IPO	**74.09**	69.17	71.54

(b) ImageNet

ViT-B/16	Base	Novel	H
CLIP	96.84	94.00	95.40
CoOp	90.63	85.20	87.83
CoCoOp	96.37	93.13	94.72
MaPLe	96.40	94.10	95.24
PromptSRC	97.30	**95.57**	96.43
CoPrompt	**97.60**	**95.57**	**96.57**
IPO	96.53	95.39	95.95

(c) Caltech101

ViT-B/16	Base	Novel	H
CLIP	91.17	97.26	94.12
CoOp	93.73	96.23	94.96
CoCoOp	93.47	96.27	94.85
MaPLe	90.83	96.00	93.34
PromptSRC	93.73	97.33	95.50
CoPrompt	92.37	96.37	94.33
IPO	**94.48**	**97.93**	**96.43**

(d) OxfordPets

ViT-B/16	Base	Novel	H
CLIP	63.37	74.89	68.65
CoOp	61.80	68.33	64.90
CoCoOp	65.27	73.73	69.24
MaPLe	66.00	73.67	69.62
PromptSRC	**67.93**	73.73	**70.71**
CoPrompt	64.17	71.50	67.64
IPO	63.83	**75.45**	69.16

(e) StanfordCars

ViT-B/16	Base	Novel	H
CLIP	72.08	77.80	74.83
CoOp	83.97	67.10	74.59
CoCoOp	75.57	77.00	76.28
MaPLe	77.10	76.97	77.03
PromptSRC	**85.57**	74.83	**79.84**
CoPrompt	72.90	72.93	72.91
IPO	74.17	**79.65**	76.81

(f) Flowers102

ViT-B/16	Base	Novel	H
CLIP	**90.10**	91.22	90.66
CoOp	87.90	88.03	87.96
CoCoOp	88.73	89.60	89.16
MaPLe	89.13	90.67	89.89
PromptSRC	88.30	91.03	89.64
CoPrompt	88.40	90.60	89.49
IPO	89.78	**91.59**	**90.67**

(g) Food101

ViT-B/16	Base	Novel	H
CLIP	27.19	36.29	31.09
CoOp	27.77	27.60	27.68
CoCoOp	29.77	31.23	30.48
MaPLe	28.33	29.00	28.66
PromptSRC	10.93	6.73	8.33
CoPrompt	10.10	4.87	6.57
IPO	**31.43**	**36.32**	**33.70**

(h) FGVCAircraft

ViT-B/16	Base	Novel	H
CLIP	69.36	75.35	72.23
CoOp	71.47	72.47	71.97
CoCoOp	73.67	75.50	74.57
MaPLe	74.33	76.37	75.34
PromptSRC	75.60	77.07	76.33
CoPrompt	**76.37**	**78.77**	**77.55**
IPO	72.25	77.53	74.80

(i) SUN397

ViT-B/16	Base	Novel	H
CLIP	53.24	59.90	56.37
CoOp	60.80	47.53	53.35
CoCoOp	58.70	52.70	55.54
MaPLe	58.20	54.17	56.11
PromptSRC	**63.17**	55.60	59.14
CoPrompt	62.77	60.40	**61.56**
IPO	55.45	**62.47**	58.75

(j) DTD

ViT-B/16	Base	Novel	H
CLIP	56.48	64.05	60.03
CoOp	69.13	50.33	58.25
CoCoOp	**71.13**	62.87	66.75
MaPLe	50.20	51.20	50.70
PromptSRC	73.27	67.00	70.00
CoPrompt	59.27	51.60	55.17
IPO	64.97	**82.13**	**72.54**

(k) EuroSAT

ViT-B/16	Base	Novel	H
CLIP	70.53	77.50	73.85
CoOp	72.50	63.57	67.74
CoCoOp	74.73	72.80	73.75
MaPLe	74.83	76.43	75.62
PromptSRC	**78.13**	78.37	**78.25**
CoPrompt	76.93	77.73	77.33
IPO	72.43	**79.35**	75.73

(l) UCF101

- ***Impact of LLM***

Models	Base	Novel	H
Phi2-2.7B	71.15	75.43	73.22
PaLM 2-L	71.32	75.93	73.55
PaLM 2-L-IT	71.13	76.16	73.56
Phi3-7B	71.43	76.68	73.96
GPT-3.5-turbo	**71.76**	**77.00**	**74.29**

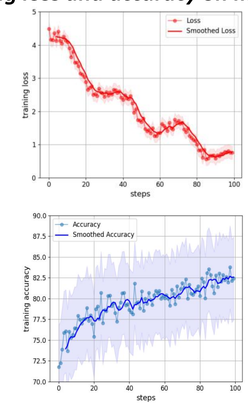
- ***Impact of LMM***

Models	Base	Novel	H
FUYU-8B	70.98	75.45	73.14
BLIP-2	71.95	76.52	73.16
Qwen-VL-Chat-9.6B	71.23	**77.08**	74.03
LLaVA-Llama-3-8B	**73.17**	75.24	74.19
MiniCPM-V-2-2.8B	71.76	77.00	**74.29**

- ***Interpretable prompts generated by IPO***

Dataset	Best Prompt
ImageNet	Take a high-quality photo of a <CLASS>.
Caltech101	Categorize the <CLASS> shown in the image.
OxfordPets	Take a well-composed photo of a <CLASS> with optimal lighting, focus, and minimal distractions. Capture the pet's unique characteristics, including expression and posture, to ensure a clear and distinct image.
StanfordCars	Describe the distinguishing characteristics of the <CLASS> in the image.
Flowers102	Identify the unique visual features of the <CLASS> flower accurately.
Food101	Identify the primary ingredient in the <CLASS> and describe its texture, color, and presentation.
FGVCAircraft	Capture a comprehensive range of well-lit, high-resolution images of an <CLASS> from various angles, meticulously showcasing its specific design features with perfect clarity and precision for unparalleled accuracy in aircraft.
SUN397	A photo of a <CLASS>, a type of large-scale scene.
DTD	Classify the intricate <CLASS> texture.
EuroSAT	Analyze the <CLASS> vehicles in the satellite image with state-of-the-art algorithms for precise classification and optimal efficiency.
UCF101	Capture a high-quality, well-lit image of a person flawlessly demonstrating the <CLASS> action with impeccable visual representation to achieve unmatched.

- ***Training loss and accuracy on ImageNet***



Prompt Optimization Prompt design

❑ Input

Instruction 1

You need to perform image classification on the large-scale visual recognition dataset based on visual features.
Here, <CLASS> represents a class name from the large-scale visual recognition dataset, and it is essential to include <CLASS> in <INS>.
Below is a description of some features of the flowers in the image:

Image descriptions

Image 0: The image showcases a close-up view of the texture on an object with distinct stripes, buttons and stitching details.

Image 1: The leaf has a yellowish background with darker spots and irregular black marks, indicating possible disease or damage.

(... more image examples. . .)

Instruction 2

Below are some previous prompts with their scores. The scores consist of loss and accuracy.
The loss value is equal to or greater than 0, while the accuracy varies from 0 to 100.

Prompt history

text prompt:
a photo of a <CLASS>.
scores:
loss: 1.6435546875, accuracy: 54.166664123535156

text prompt:
Identify the intricate details and unique patterns of the <CLASS> texture scores:
loss: 1.3935546875, accuracy: 70.83332824707031
(... more previous prompts and scores . . .)

Instruction 3

Generate a prompt different from all the prompts <INS> above, with a lower loss and higher accuracy than all the prompts <INS> above. The prompt should begin with <INS> and end with </INS>. The prompt should be concise, effective, and generally applicable to all problems above.

❑ Output

Below are the prompts created according to your guidelines:

(1) <INS> Analyze the intricate texture of <CLASS> with precision and attention to detail. </INS> (2) <INS> Thoroughly analyze and classify the intricate <CLASS> texture with meticulous precision. </INS>

(3) <INS> Analyze, categorize, and classify the intricate <CLASS> texture with unparalleled precision and exceptional attention to detail. </INS> (... more output prompts examples. . .)