# Generic Prompts for Robust Vision-Language Models


Mohammed Azz

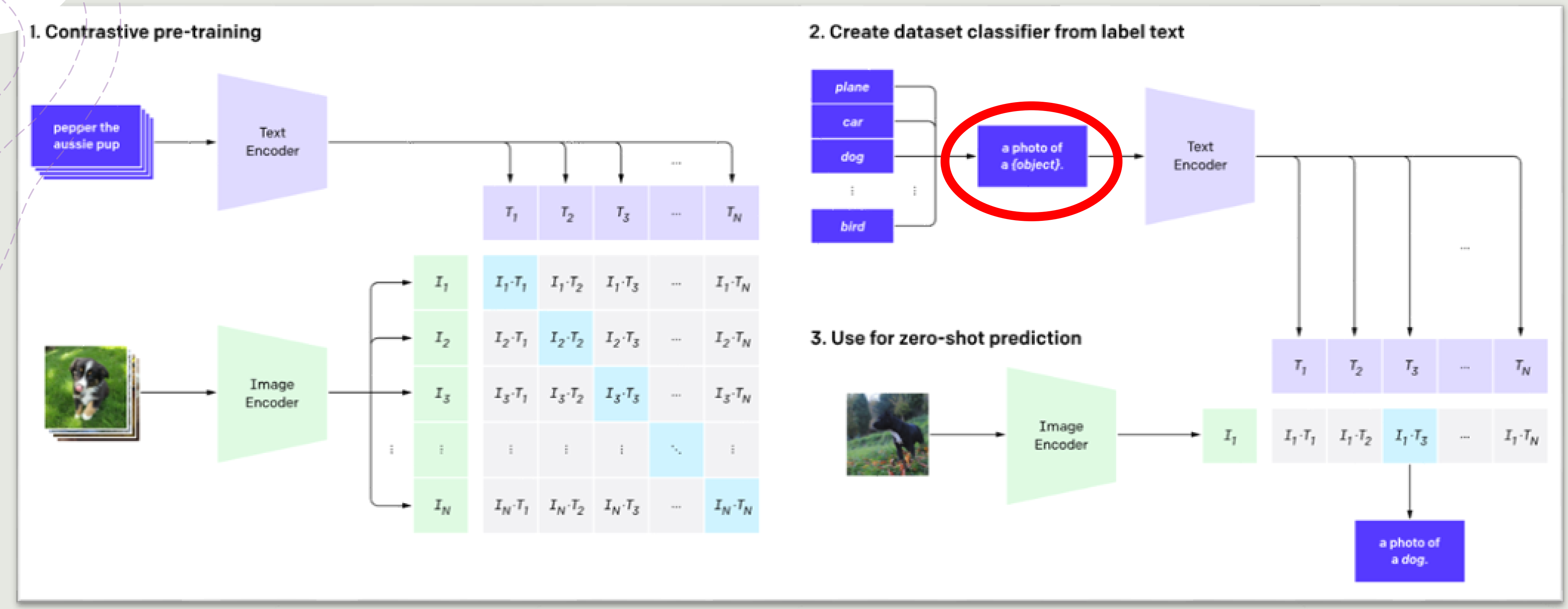
Raza Imam

Dmitry Demidov

# 1. Introduction & Problem Statement

# Introduction

- Rise of Vision-Language models.
- Encoder training required for the models
- CLIP
- Advantage:
  - Improved Performance over older classification methods.
  - Pre-trained models can be used for zero shot classification.
- Disadvantages:
  - Inference: results rely heavily on the introduced text prompts




# How CLIP Works

+ Notes: Fine-tuning the pre-trained CLIP model often leads to worse results due to hardware/process limitations (CLIP is trained with potentially the largest pool of images with a self-supervised process).

# Motivation

- What the issue with the CLIP:
  - Manual prompt preparation, time inefficient.
  - Unreliable results.
  - Limited transferability to other datasets.
- Can we provide a prompt engineering cost-effective way to produce good results with limited training?

# Related work

- Baseline:
  - A photo of "label"
- Manual prompt generation with fine-tuned selection(*):
  - Introduce many prompt templates. (Requires a lot of manual work)
  - Fine-tune the best subset selection of those templates.
- Optimized prompt Generation (CoOp)(**):
  - With a template, it finetunes the words of the template using a dictionary.
  - Results are unreliable and result in random words and symbols.


***(*)Zhou et al. Learning to prompt for vision-language models. International Journal of Computer Vision (2022).***
***(**)Zhou et al. Conditional prompt learning for vision-language models (2022).***

# Related work



**Table 4** The nearest words for each of the 16 context vectors learned by CoOp, with their distances shown in parentheses N/A means non-Latin characters.

| # | ImageNet | Food101 | OxfordPets | DTD | UCF101 |
|---|---|---|---|---|---|
| 1 | potd (1.7136) | lc (0.6752) | tosc (2.5952) | boxed (0.9433) | meteorologist (1.5377) |
| 2 | that (1.4015) | enjoyed (0.5305) | judge (1.2635) | seed (1.0498) | exe (0.9807) |
| 3 | filmed (1.2275) | beh (0.5390) | fluffy (1.6099) | anna (0.8127) | parents (1.0654) |
| 4 | fruit (1.4864) | matches (0.5646) | cart (1.3958) | mountain (0.9509) | masterful (0.9528) |
| 5 | ,... (1.5863) | nytimes (0.6993) | harlan (2.2948) | eldest (0.7111) | fe (1.3574) |
| 6 | ° (1.7502) | prou (0.5905) | paw (1.3055) | pretty (0.8762) | thof (1.2841) |
| 7 | excluded (1.2355) | lower (0.5390) | incase (1.2215) | faces (0.7872) | where (0.9705) |
| 8 | cold (1.4654) | N/A | bie (1.5454) | honey (1.8414) | kristen (1.1921) |
| 9 | stery (1.6085) | minute (0.5672) | snuggle (1.1578) | series (1.6680) | imam (1.1297) |
| 10 | warri (1.3055) | ~ (0.5529) | along (1.8298) | coca (1.5571) | near (0.8942) |
| 11 | marvelcomics (1.5638) | well (0.5659) | enjoyment (2.3495) | moon (1.2775) | tummy (1.4303) |
| 12 | .: (1.7387) | ends (0.6113) | jt (1.3726) | lh (1.0382) | hel (0.7644) |
| 13 | N/A | mis (0.5826) | improving (1.3198) | won (0.9314) | boop (1.0491) |
| 14 | lation (1.5015) | somethin (0.6041) | srsly (1.6759) | replied (1.1429) | N/A |
| 15 | muh (1.4985) | seminar (0.5274) | asteroid (1.3395) | sent (1.3173) | facial (1.4452) |
| 16 | .# (1.9340) | N/A | N/A | piedmont (1.5198) | during (1.1755) |

# Possible Approaches

- Related work-specific areas of work:
  - General prompts: Introduce a more diverse list of templates for selection (Manual work, limited robustness).
  - Prompt optimization: Filter symbols and unreasonable constructions (Unpredictable results, requires heavy training).

- Make class specific prompts:
  - Gives potential uniqueness to each class (possibly more class specific during inference).
  - Incorporates more information about the class.
  - Requires access to prompts of each class.

# 2. Methodology

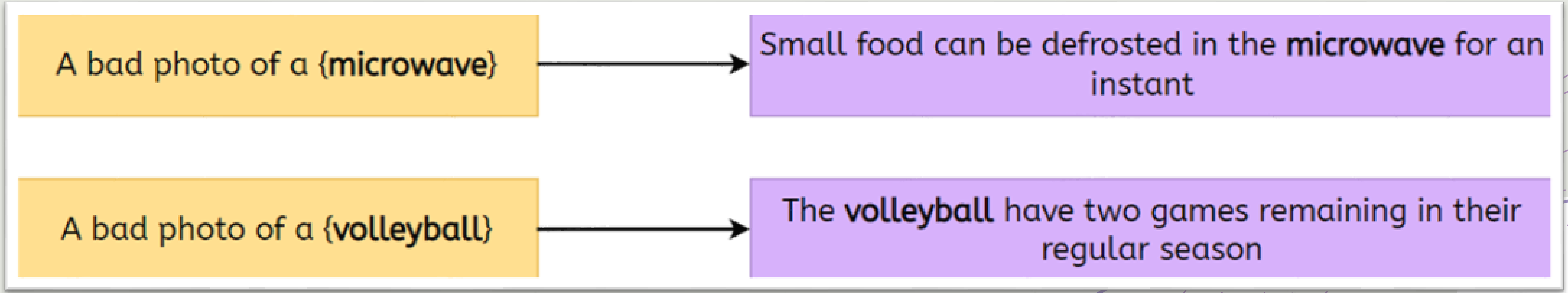
Prompt Engineering

# Generic Prompts

- **Replace** hard labels with meta-level descriptions to improve robustness and generalization of CLIP.
- Use automatically gathered labels to provide more **generic** and descriptive prompts to the model.
- Avoid human-intensive techniques such as **manually** prepared sentences or learnable tokens.

# Generic Prompts

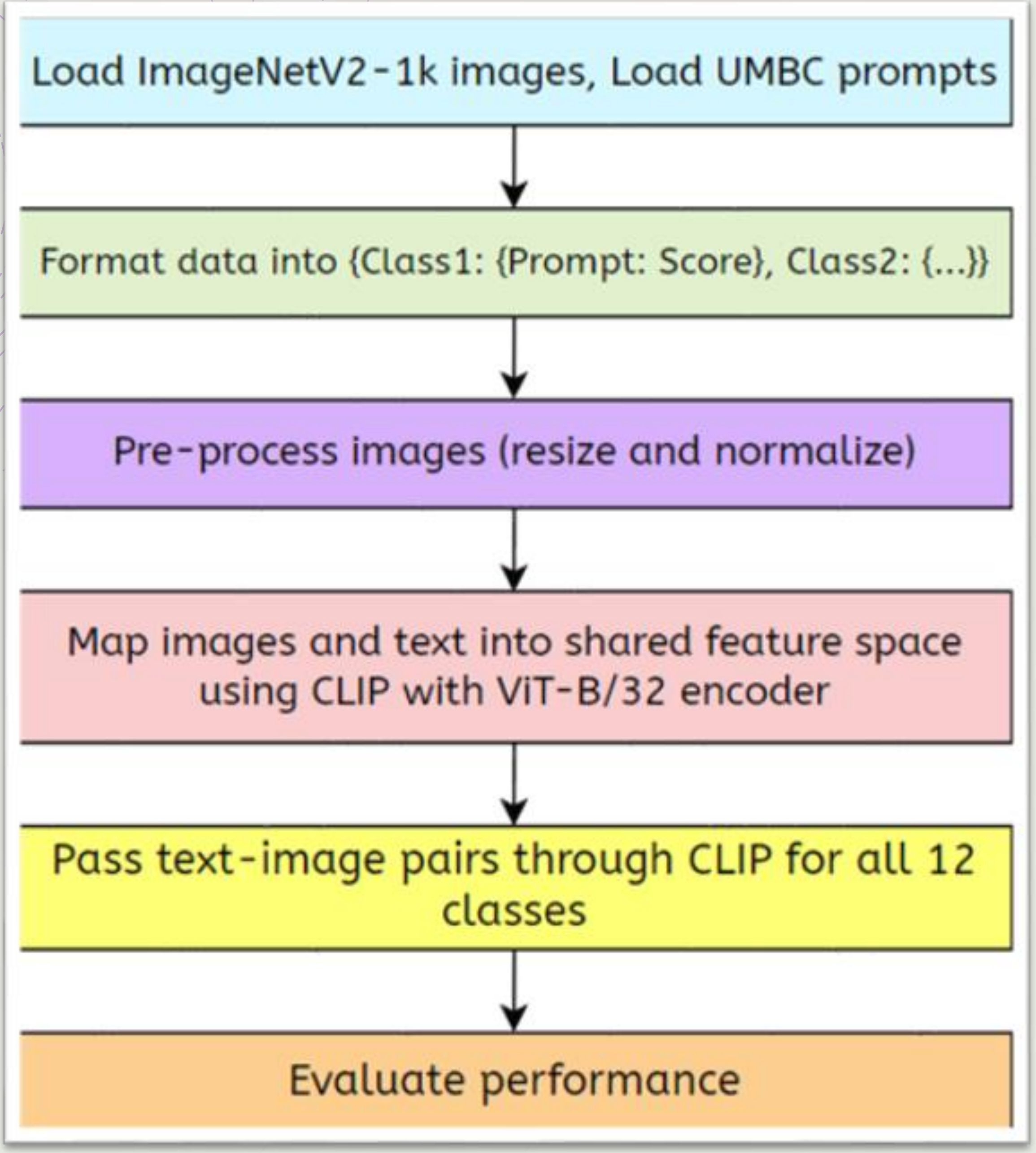
+ **Replace** hard labels with meta-level descriptions to improve robustness and generalization of CLIP.
+ Use automatically gathered labels to provide more **generic** and descriptive prompts to the model.
+ Avoid human-intensive techniques such as **manually** prepared sentences or learnable tokens.



# Data Source

The UMBC WebBase corpus is a dataset:

- derived from 2007 by the Stanford WebBase project;
- containing English paragraphs with over **3 billion** words;
- consists of text from **100 million** web pages from more than **50,000** websites;
- extracted from textual content from HTML tags;
- compressed, it is about 13GB in size.




# How we're making prompts

**ImageNetV2**-1k validation classes & **UMBC** web base corpus of paragraphs.

Select **12** classes → load prompts into this format: {Banana:{"Prompt1" : 24.50, "Prompt2 : 28.4}, Cucumber:{...}, ...}.

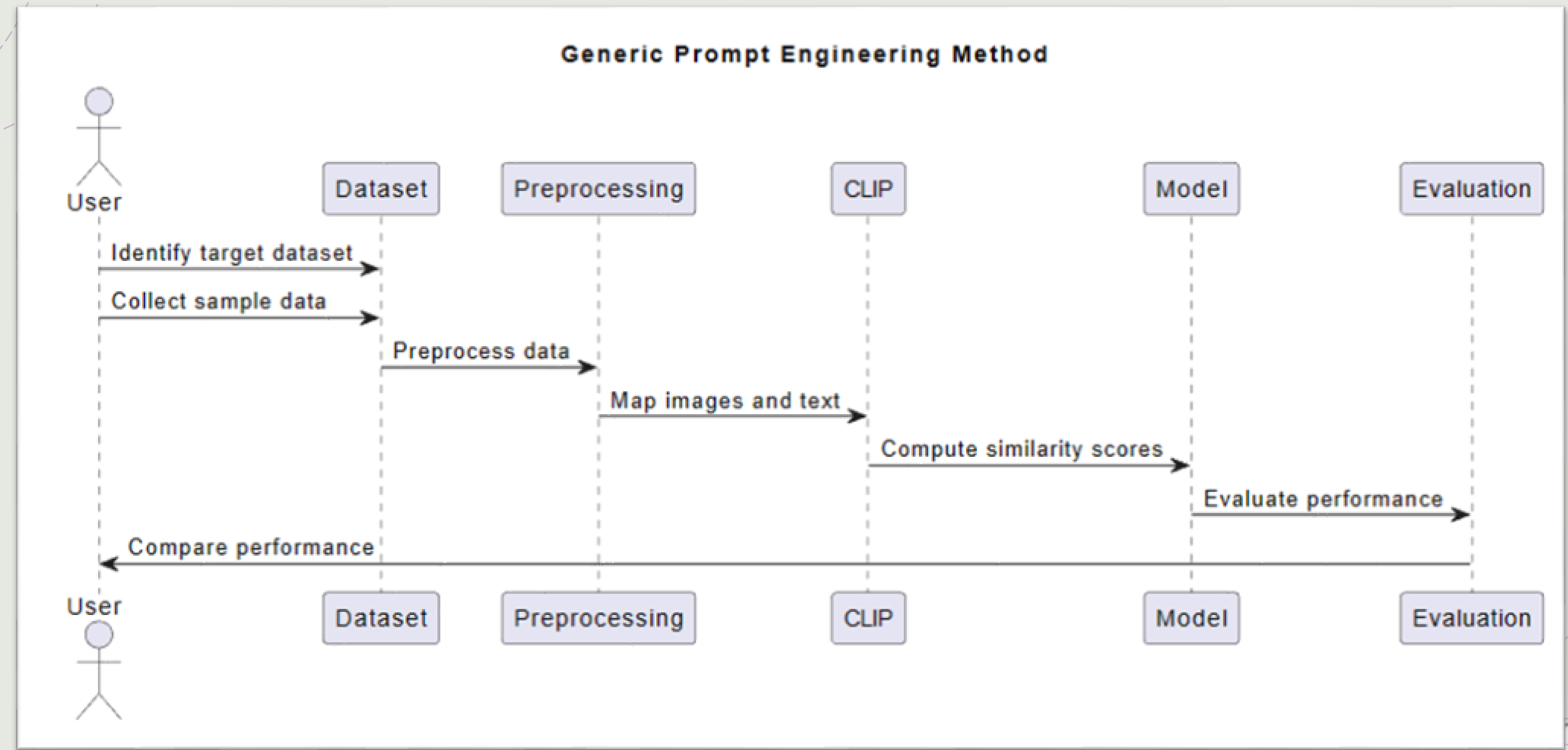
For each class, we provide **100** or 1000 prompts, which are paired with a single image of that class.

We pre-process images by resizing them to 224x224 and normalizing them.

We use pre-trained CLIP with ViT-B/32 image encoder to map both images and texts into a shared feature space.

We pass all text-image pairs to CLIP for all 12 classes, resulting in 1'200 runs when using 100 prompts per class and 12'000 runs when using 1000 prompts per class.

# Transferability



# Robustness

**Generalization to new data:**

- Generic prompts generated show higher transferability and generalization to new datasets.
- Generated prompts can be used to achieve good performance on different datasets without the need for manual adaptation.
- Potentially captures more general features of the images, making it more robust to changes in data distribution and able to handle diverse datasets.

# 3. Experiments & Results

# Setup

+ Model: Pre-trained CLIP (ViT-B/32)
+ Dataset: ImageNet v2 (Evaluation Set)

Hyperparameters:

- Length of prompts
- Prompt's number per class
- Filtering rules for final prompts

| Method | ImageNet Validation Subset (12 classes) | | | | ImageNet Validation Set | | Random Images | |
|---|---|---|---|---|---|---|---|---|
| | 12 classes only | | 12 classes among 1000 | | 1000 classes | | 12 classes only | |
| | *Top-1* | *Top-5* | *Top-1* | *Top-5* | *Top-1* | *Top-5* | *Top-1* | *Top-5* |
| CLIP (ImageNet prompts) | *92.50* | *100.00* | *60.83* | *79.17* | *55.93* | *83.40* | *100.00* | *100.00* |
| CLIP (A photo of {}) | 93.00 | 100.00 | 52.50 | 75.83 | 51.92 | 78.80 | **100.00** | 100.00 |

# Quantitative Results: 12 classes

| Method | ImageNet Validation Subset (12 classes) | | | | ImageNet Validation Set | | Random Images | |
|---|---|---|---|---|---|---|---|---|
| | 12 classes only | | 12 classes among 1000 | | 1000 classes | | 12 classes only | |
| | ***Top-1*** | ***Top-5*** | ***Top-1*** | ***Top-5*** | ***Top-1*** | ***Top-5*** | ***Top-1*** | ***Top-5*** |
| CLIP (ImageNet prompts) | *92.50* | *100.00* | *60.83* | *79.17* | *55.93* | *83.40* | *100.00* | *100.00* |
| CLIP (A photo of {}) | 93.00 | 100.00 | 52.50 | 75.83 | 51.92 | 78.80 | **100.00** | 100.00 |
| Ours (100p, 1 word) | **94.17** | 100.00 | 35.83* | 76.67* | - | - | 98.33 | 100.00 |
| Ours (100p, 1 word, A photo of {}) | **94.17** | 100.00 | **63.33*** | 85.00* | - | - | 98.33 | 100.00 |

| Method | ImageNet Validation Subset (12 classes) | | | | ImageNet Validation Set | | Random Images | |
|---|---|---|---|---|---|---|---|---|
| | 12 classes only | | 12 classes among 1000 | | 1000 classes | | 12 classes only | |
| | *Top-1* | *Top-5* | *Top-1* | *Top-5* | *Top-1* | *Top-5* | *Top-1* | *Top-5* |
| CLIP (ImageNet prompts) | *92.50* | *100.00* | *60.83* | *79.17* | *55.93* | *83.40* | *100.00* | *100.00* |
| CLIP (A photo of {}) | 93.00 | 100.00 | 52.50 | 75.83 | 51.92 | 78.80 | **100.00** | 100.00 |
| Ours (100p, 1 word) | **94.17** | 100.00 | 35.83* | 76.67* | - | - | 98.33 | 100.00 |
| Ours (100p, 1 word, A photo of {}) | **94.17** | 100.00 | **63.33*** | 85.00* | - | - | 98.33 | 100.00 |
| Ours (1000p, 1 word) | 92.50 | 100.00 | 37.50* | 80.00* | - | - | 98.33 | 100.00 |
| Ours (1000p, 1 word, A photo of {}) | 93.33 | 100.00 | **63.33*** | **88.33*** | - | - | **100.00** | 100.00 |

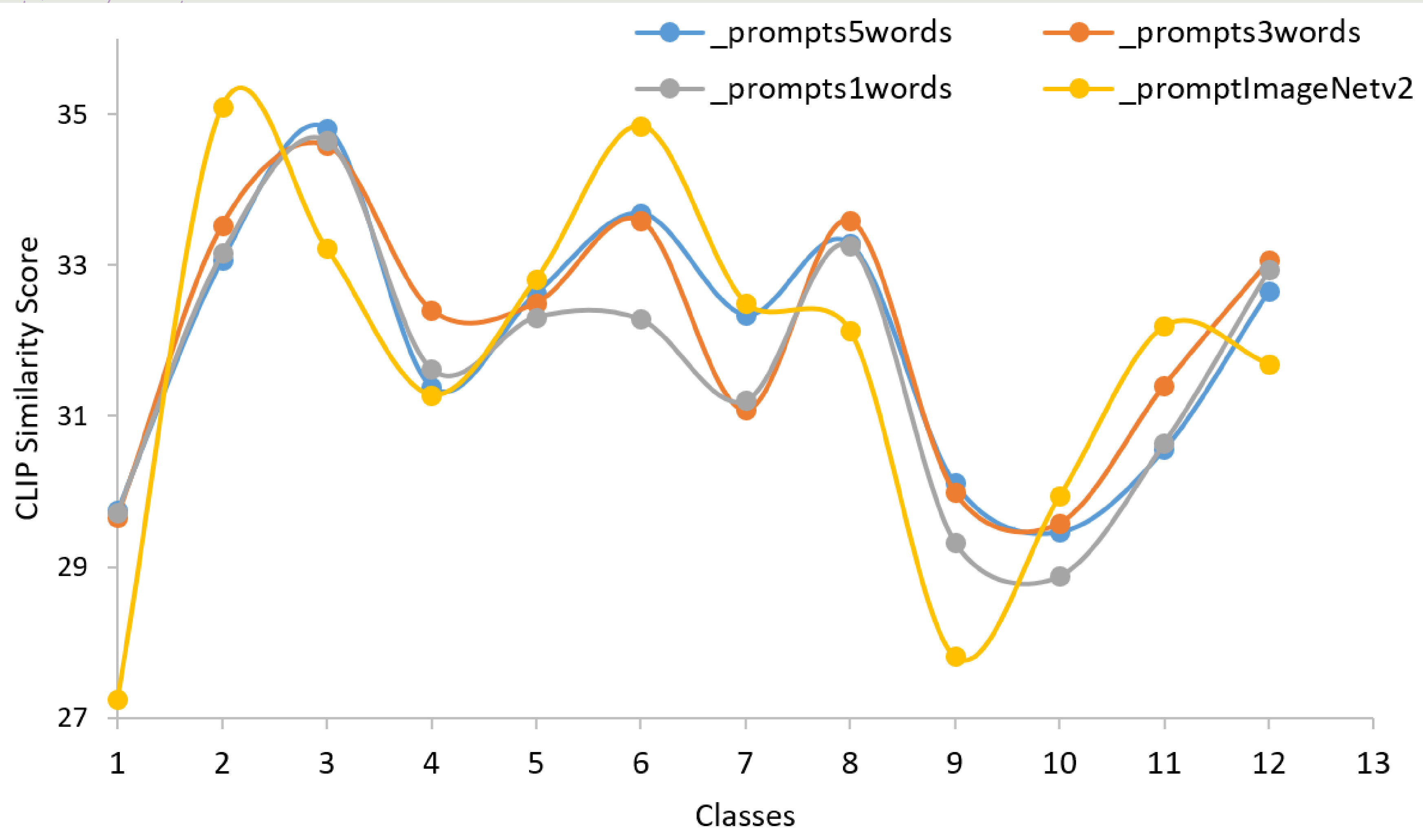
| Method | ImageNet Validation Subset (100 classes) | | | | ImageNet Validation Set | |
|---|---|---|---|---|---|---|
| | 100 classes only | | 100 classes among 1000 | | 1000 classes | |
| | *Top-1* | *Top-5* | *Top-1* | *Top-5* | *Top-1* | *Top-5* |
| CLIP (ImageNet prompts) | *84.9* | *97.2* | *59.3* | *85.6* | *55.93* | *83.40* |
| CLIP (A photo of {}) | 80.8 | **96.1** | 55.6 | 81.3 | 51.92 | 78.80 |

# Quantitative Results: 100 classes

| Method | ImageNet Validation Subset (100 classes) | | | | ImageNet Validation Set | |
|---|---|---|---|---|---|---|
| | 100 classes only | | 100 classes among 1000 | | 1000 classes | |
| | ***Top-1*** | ***Top-5*** | ***Top-1*** | ***Top-5*** | ***Top-1*** | ***Top-5*** |
| CLIP (ImageNet prompts) | *84.2* | *97.3* | *58.3* | *85.7* | *55.93* | *83.40* |
| CLIP (A photo of {}) | 80.1 | **95.8** | 54.2 | 81.4 | 51.92 | 78.80 |
| Ours (100p, 1 word) | 78.0 | 94.9 | 42.4* | 76.0* | - | - |
| Ours (100p, 1 word, A photo of {}) | **80.4** | 95.3 | **57.9*** | **84.0*** | - | - |

| Method | ImageNet Validation Subset (100 classes) | | | | ImageNet Validation Set | |
|---|---|---|---|---|---|---|
| | 100 classes only | | 100 classes among 1000 | | 1000 classes | |
| | ***Top-1*** | ***Top-5*** | ***Top-1*** | ***Top-5*** | ***Top-1*** | ***Top-5*** |
| CLIP (ImageNet prompts) | *84.2* | *97.3* | *58.3* | *85.7* | *55.93* | *83.40* |
| CLIP (A photo of {}) | 80.1 | **95.8** | 54.2 | 81.4 | 51.92 | 78.80 |
| Ours (100p, 1 word) | 78.0 | 94.9 | 42.4* | 76.0* | - | - |
| Ours (100p, 1 word, A photo of {}) | **80.4** | 95.3 | **57.9*** | **84.0*** | - | - |
| Ours (100p, 2 words) | 77.9 | 95.4 | 42.0* | 73.9* | - | - |
| Ours (100p, 2 words, A photo of {}) | 79.3 | 95.2 | **59.4*** | **84.8*** | - | - |

# Qualitative Results




Maximum similarity score for each class

# 4. Conclusion

# Conclusion

- The proposed idea indeed works;
- The optimal **prompt's length** is **3-10 words**;
- The optimal prompt number **per class** is **100-1000 prompts**;
- The results **can be further improved** by tweaking and polishing the search pipeline;
- The solution **can work incredibly faster** if optimized for multi-core processing;
- The code will be released.

# Future work

# Future Work

- Use all 1000 ImageNet classes;
- Do better class names pre-processing (e.g., search by more common synonyms to find sentences easier);
- Filter sentences in a better way (e.g., remove "and/or", etc.);
- Check on other datasets (e.g., CIFAR, Fine-Grained datasets, etc.);
- Pick Top-N prompts (e.g., top-10 %);
- Consider a publication out of the project.

# Thank you for attention