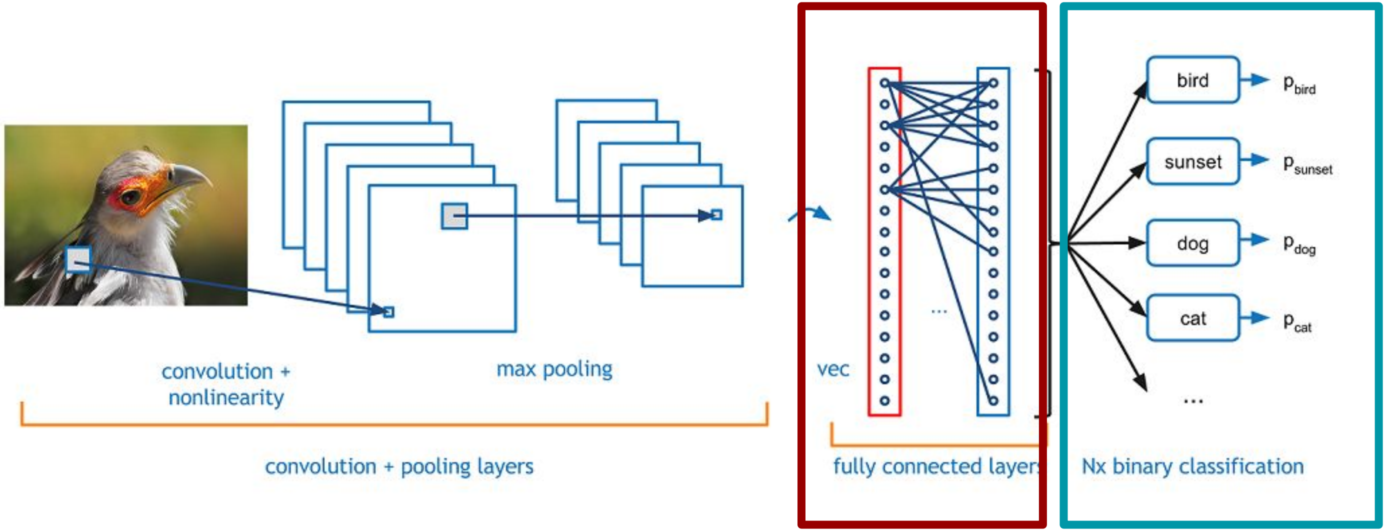
Land Cover Classification Using Foundation Models

Haiyang Jiang, Jingnan Cao

Introduction

Traditional Vision System Limitations

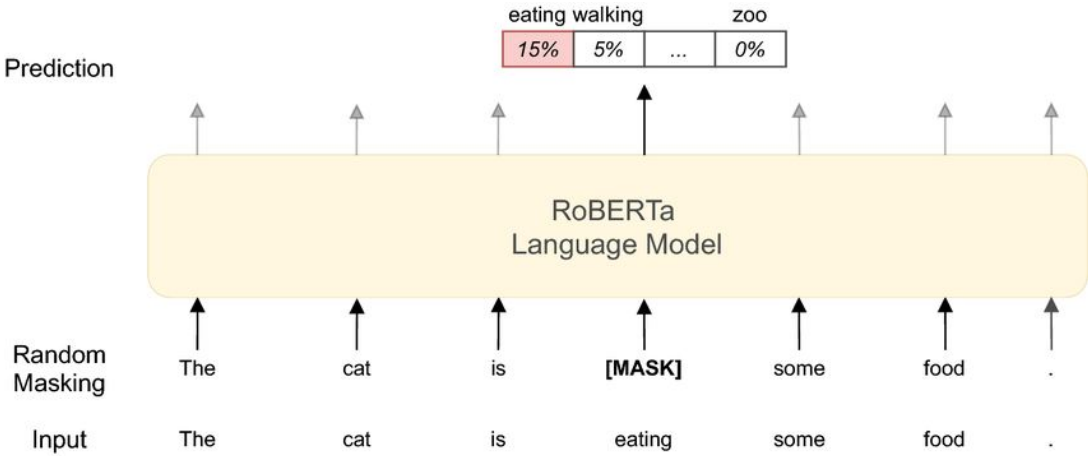


Predefined classifier
-> Static output
-> Lack of dynamic

Predefined object classes
-> Fixed set solution
-> Hard to add additional label
-> Limited generability

NLP Revolution by BERT & GPT

Directly learn from **raw text** by task-agnostic objectives (autoregressive / masked language modeling)



-> Can learn from rich online resources -> **Scalability**

-> Boarder source of supervision from the text -> **Self-supervised**

-> Ability to learn **generalized** feature representation

-> Strong zero-shot **generalization** to downstream tasks

Language vs. Vision

Natural Language

- Self supervision (LM)
- Large training data
- Zero-shot transferability

Images

- Supervised learning
- Not that large training data (ImageNet)

Idea: enabling better transferability by connecting vision tasks by languages guiding

Related Work

Leanring Image Representation from Natural Language

By reviewing historical works, reasons why this idea haven't achieved impressive performance:

1. Image-text paired dataset not large enough

   -> WIT WebImageText: **400M image-text pairs**

   -> Comparable with WebText for GPT-2

2. Model not powerful enough

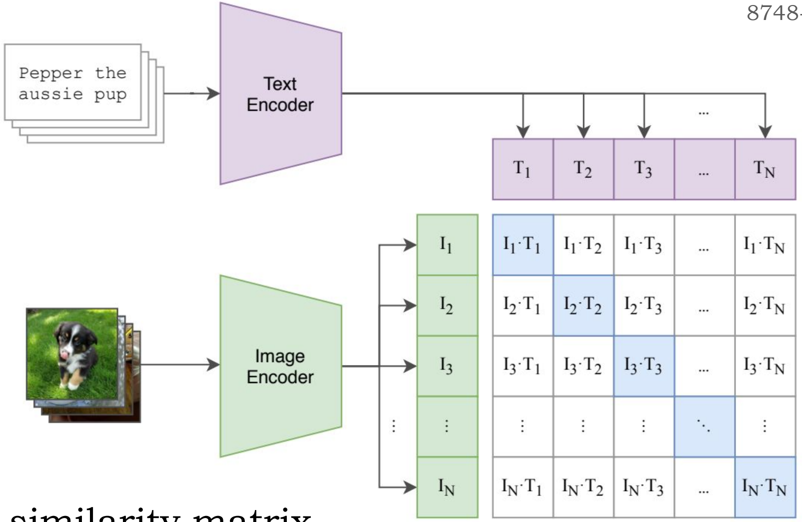
   -> **ResNet / Vision Transformer** as vision encoder

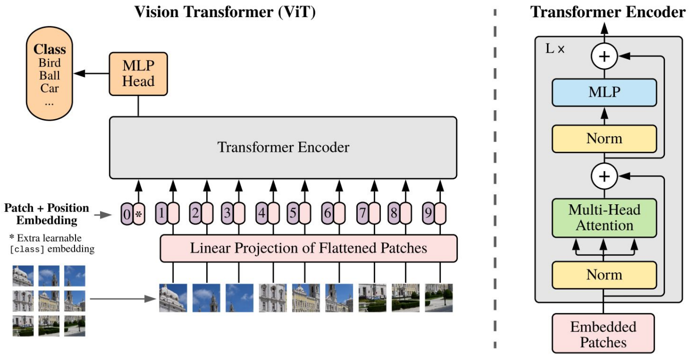
   -> **Transformer** as text encoder

CLIP: Contrastive Language-Image Pretraining

Train on WebImageText(WIT): a newly constructed dataset of **400 million** (image, text) pairs on the Internet

Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.





$$\min\left(\sum_{i=1}^{N}\sum_{j=1}^{N}(I_i \cdot T_j)_{i \neq j} - \sum_{i=1}^{N}(I_i \cdot T_i)\right)$$

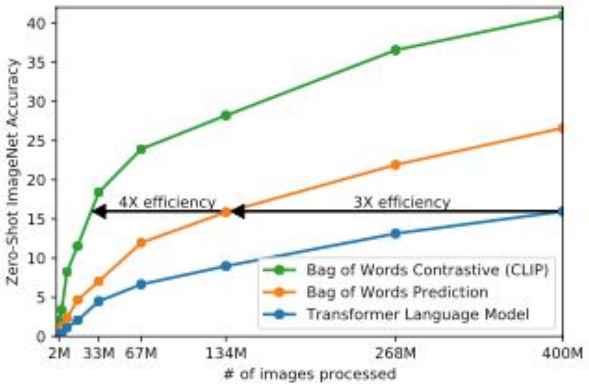
N*N similarity matrix

Maximizing **N** pairs similarity (positive pair)

Minimizing **N^2-N** pairs similarity (negative pair)

batch-size = 32,768 **VERY BIG!**

Training Objective



Crucial Problem: for scaled data,

training efficiency is the key to success

Start from predicting **caption** of the image (similar in generative model)

-> Wide variety of possible answer -> **Hard & high computation cost**

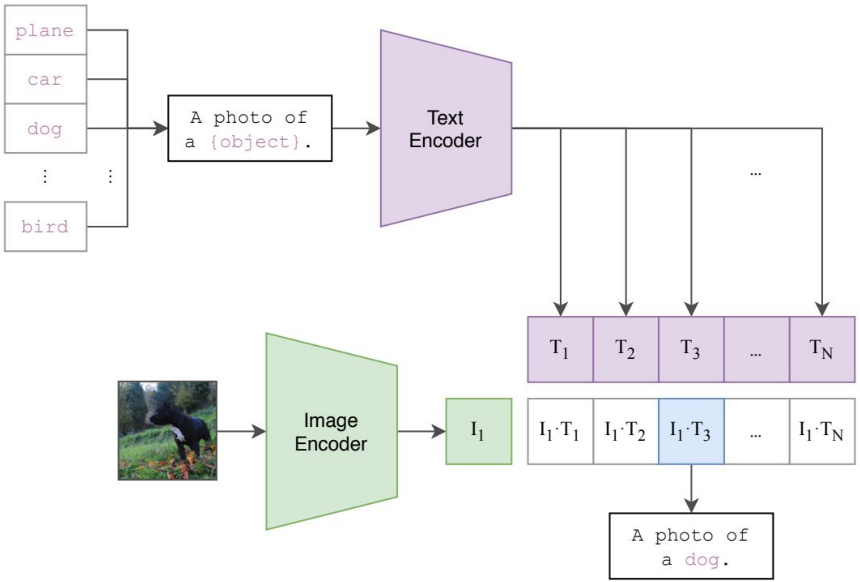
Captioning: question answering

Classification: multiple choice

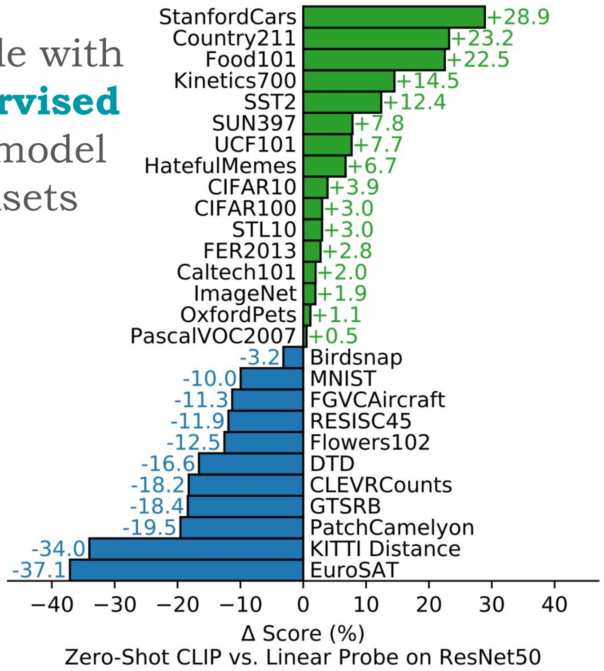
Pairing: simple yes or no -> **Contrastive learning**

computation cost reducing

Zero-Shot Image Classification



Comparable with **fully supervised** ResNet50 model on 27 datasets



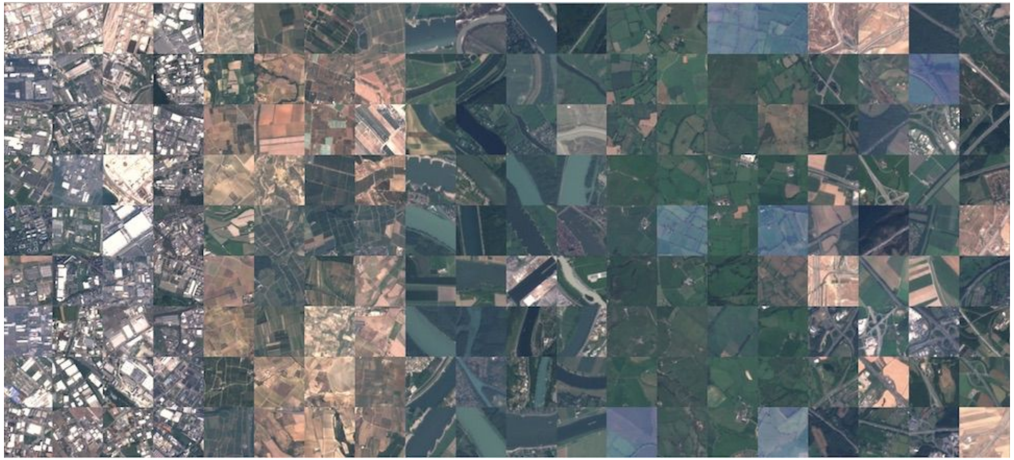
Class + Template = Sentence

Each image -> N classes similarity -> Get the max as prediction

Task

Zero-shot / Few-shot Classification on EuroSAT

Using Foundation Model CLIP for land cover classification on the EuroSAT dataset to classify satellite images into various land cover types



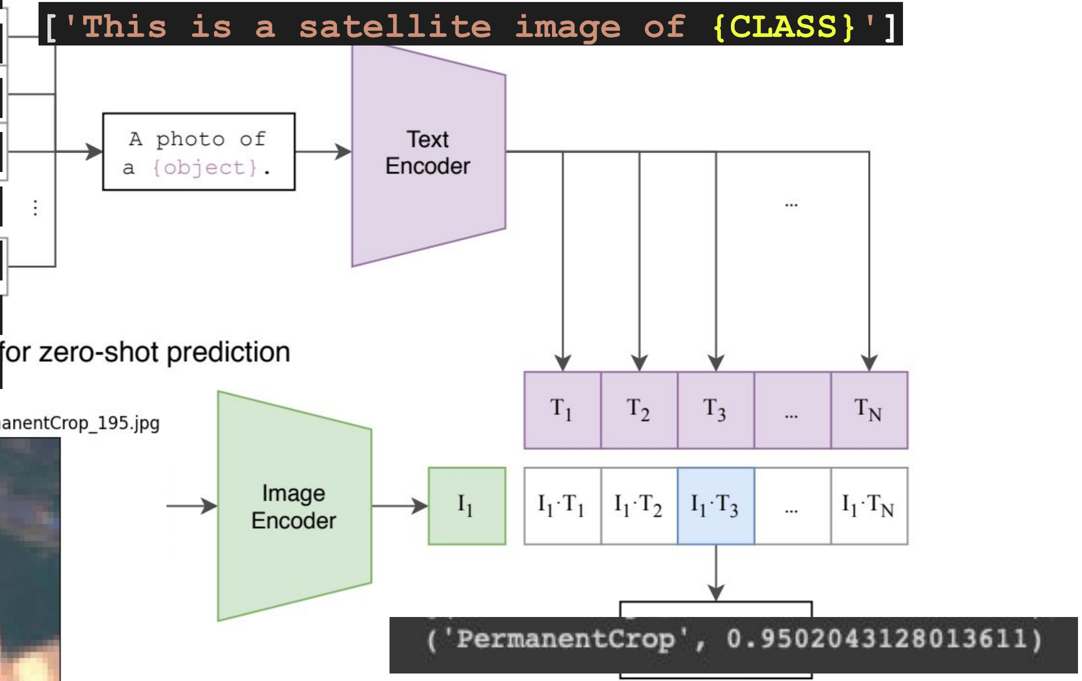
EuroSAT covers **13 spectral bands** and consisting out of **10 classes** with in total **27,000 labeled** and geo-referenced images

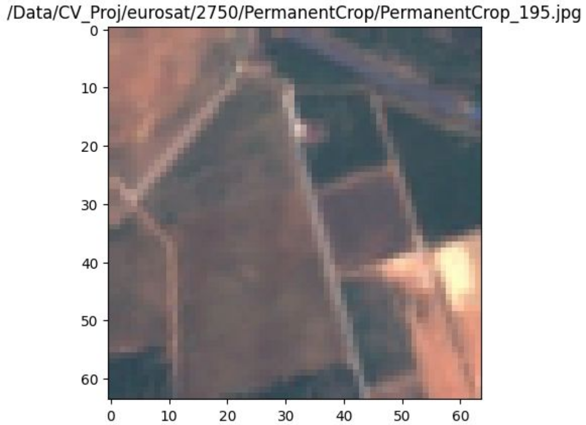
```
['AnnualCrop', 'Forest', 'HerbaceousVegetation', 'Highway',
'Industrial','Pasture', 'PermanentCrop', 'Residential', 'River', 'SeaLake']
```

Method

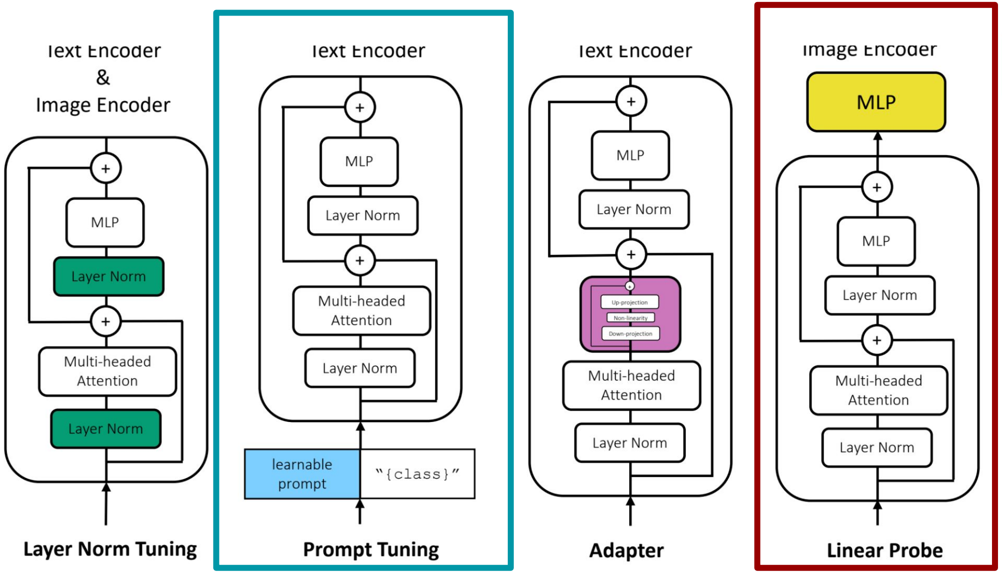
Zero-shot Classification on EuroSAT

```
['AnnualCrop',
'Forest',
'HerbaceousVegetation',
'Highway',
'Industrial',
'Pasture',
'PermanentCrop',
'Residential',
'River',
'SeaLake']
```



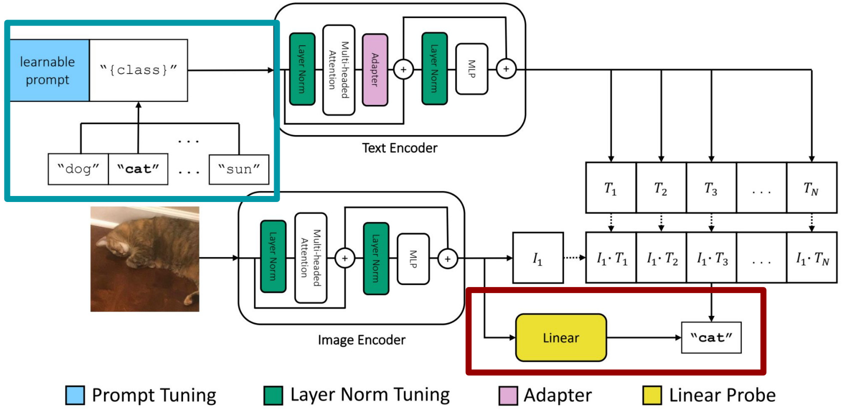


Few-Shot Learning Enhancement

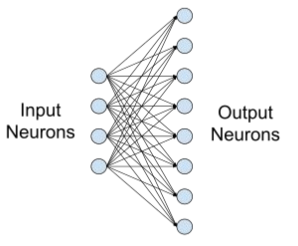


No possible for Foundation Model !

No possible for Foundation Model !

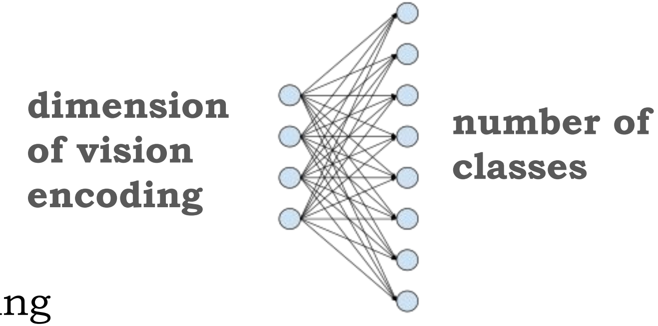


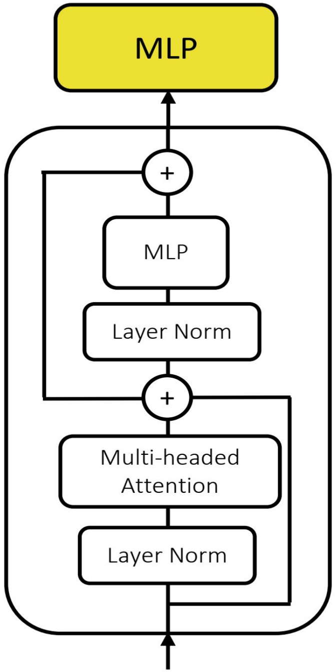

```
imagenet_templates = [
    'a bad photo of a {}.',
    'a photo of many {}.',
    'a sculpture of a {}.',
    'a photo of the hard to see {}.',
    'a low resolution photo of the {}.',
    'a rendering of a {}.',
    'graffiti of a {}.',
    'a bad photo of the {}.',
    'a cropped photo of the {}.',
    'a tattoo of a {}.',
    'the embroidered {}.',
    'a photo of a hard to see {}.',
    'a bright photo of a {}.',
    'a photo of a clean {}.',
    'a photo of a dirty {}.',
    'a dark photo of the {}.',
    'a drawing of a {}.',
    'a photo of my {}.',
```



Kim K, Laskin M, Mordatch I, et al. How to adapt your large-scale vision-and-language model[J]. 2021.

Few-Shot Learning Enhancement
- Linear Probe





1. 16 shots learning
   16 samples / class for training

2. Full dataset feature learning
   7:3 Training vs. Validating

Both test on whole dataset

```
classifier = SimpleClassifier(input_dim, num_classes).to(device)
optimizer = optim.Adam(classifier.parameters(), lr=1e-3)
loss_fn = nn.CrossEntropyLoss()
```

```
num_epochs = 100
validation_interval = 10  # Validate after every 10 epochs
save_interval = 20  # Save after every 20 epochs
early_stopping_patience = 10  # Number of epochs to wait for loss improvement
best_val_accuracy = 0  # Track best validation accuracy
no_improvement_epochs = 0  # Track epochs without improvement for early stopping
```

Few-Shot Learning Enhancement - Prompt Learning CoOp

Hand-crafted prompt engineering limitation:

1. Require domain expertise
2. Extremely time-consuming
3. Slight change will cause big difference

Automatically learn the contextual vectors in continuous space

-> Use cross-entropy classification loss to evaluate a context vector towards a class

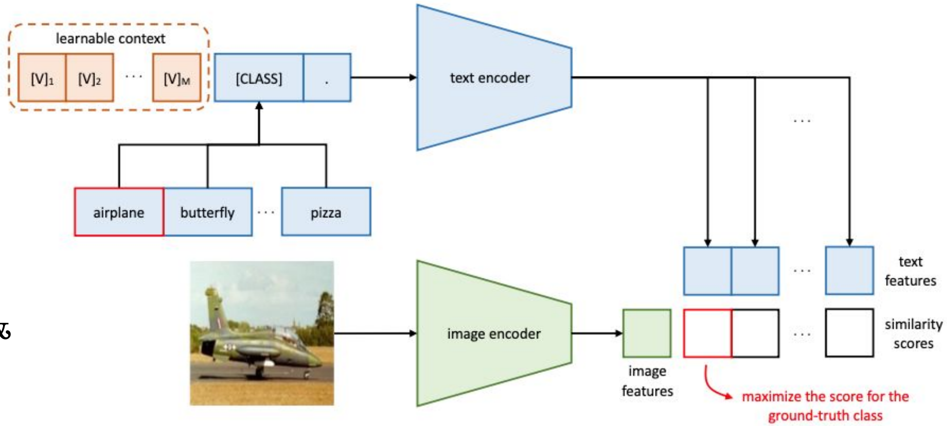
-> Gradient from text encoder all the way to original context vectors

-> Data-efficient, beat hand-creafted prompts & linear probe with 1-2 shots

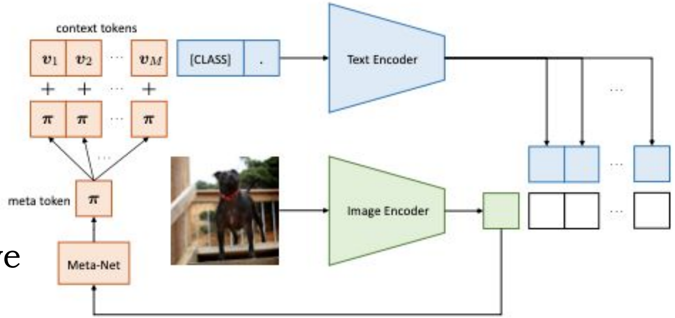
Flowers102	Prompt	Accuracy
	a photo of a [CLASS].	60.86
	a flower photo of a [CLASS].	65.81
	a photo of a [CLASS], a type of flower.	66.14
	$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**94.51**

(b)

EuroSAT	Prompt	Accuracy
	a photo of a [CLASS].	24.17
	a satellite photo of [CLASS].	37.46
	a centered satellite photo of [CLASS].	37.56
	$[V]_1 [V]_2 \ldots [V]_M$ [CLASS].	**83.53**



Few-Shot Learning Enhancement
- Prompt Learning CoOp & CoCoOp

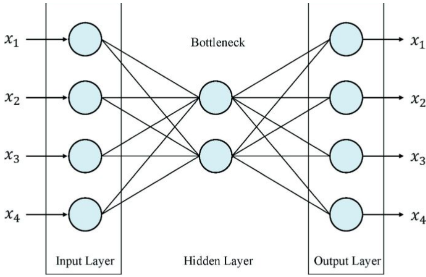


$$p(y = i|\boldsymbol{x}) = \frac{\exp(\cos(g(\boldsymbol{t}_i), \boldsymbol{f})/\tau)}{\sum_{j=1}^{K} \exp(\cos(g(\boldsymbol{t}_j), \boldsymbol{f})/\tau)},$$

Simple cross entropy objective

$$\boldsymbol{t} = [\mathrm{V}]_1[\mathrm{V}]_2 \ldots [\mathrm{V}]_M[\mathrm{CLASS}],$$

1. Unified context for all classes
-> More generalized context



$$\boldsymbol{t}_i = [\mathrm{V}]_1^i[\mathrm{V}]_2^i \ldots [\mathrm{V}]_M^i[\mathrm{CLASS}]_i,$$

2. Class-specificed context
-> might benefit fine-grained dataset such as StandfordCars
-> more parameters -> need more shots
-> limited generalization to new class

$$\boldsymbol{t}_i(\boldsymbol{x}) = \{\boldsymbol{v}_1(\boldsymbol{x}), \boldsymbol{v}_2(\boldsymbol{x}), \ldots, \boldsymbol{v}_M(\boldsymbol{x}), \boldsymbol{c}_i\}$$

$$\boldsymbol{\pi} = h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

image-based token

$$\boldsymbol{v}_m(\boldsymbol{x}) = \boldsymbol{v}_m + \boldsymbol{\pi}$$

merge image token and context vector

3. Context conditioned on image
Above prompt: poor performance on unseen class -> static context
-> from class-oriented to instance-oriented
-> dynamic context based on image
-> better generalized on unseen class
-> much more parameters to train

Few-Shot Learning Enhancement

- Prompt Learning CoOp & CoCoOp

$$t = [\mathrm{V}]_1[\mathrm{V}]_2 \dots [\mathrm{V}]_M[\mathrm{CLASS}],$$

1. Unified context for all classes
16 context + 1 class = 17 words
-> learn tensor [1*16*77]

$$t_i = [\mathrm{V}]_1^i[\mathrm{V}]_2^i \dots [\mathrm{V}]_M^i[\mathrm{CLASS}]^i,$$

2. Class-specificed context
(16 context + 1 class) * 10 classes = 170 words
-> learn tensor [10*16*77]

$$t_i(x) = \{v_1(x), v_2(x), \dots, v_M(x), c_i\}$$

$$\pi = h_\theta(x)$$

image-based token

$$v_m(x) = v_m + \pi$$

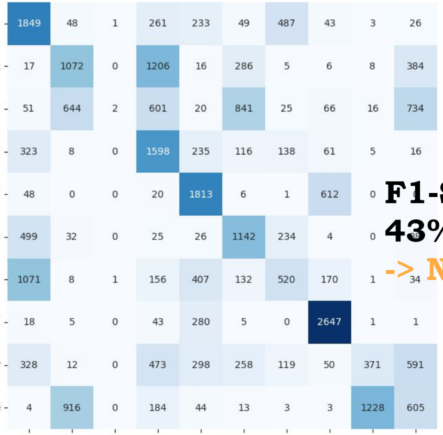
merge image token and context vector

3. Context conditioned on image
Starts from initialized “a photo of a {CLASS}”
4 context + 1 class = 5 words
-> learn tensor [1*4*77]
+
Meta-Net (two-layer bottleneck)

Since introduce a network, to keep training efficiency reduce the context length to 4

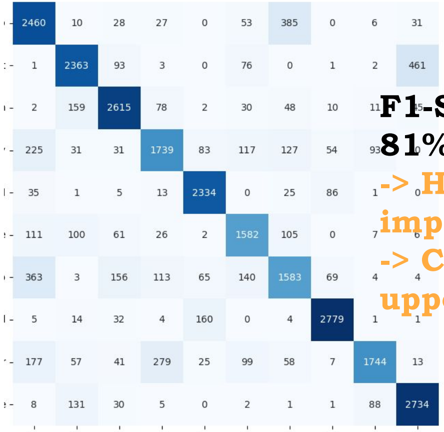
Result

Zero-shot CLIP



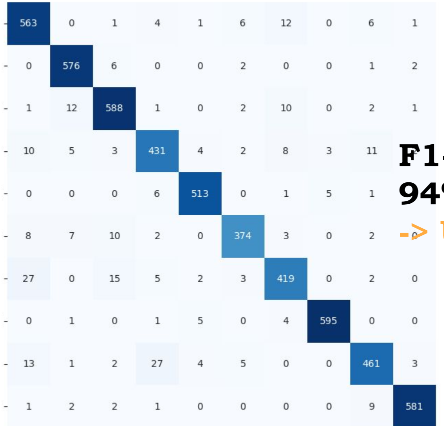
F1-Score:
43%
-> Not good

16-shots Linear Probe



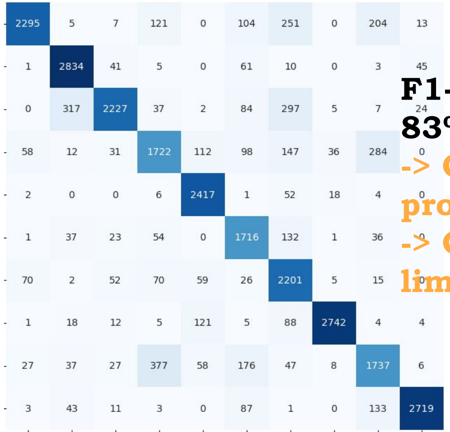
F1-Score:
81% (+38%)
-> Huge improve
-> Close to upper limit

Representation Learning Full Dataset Linear Probe



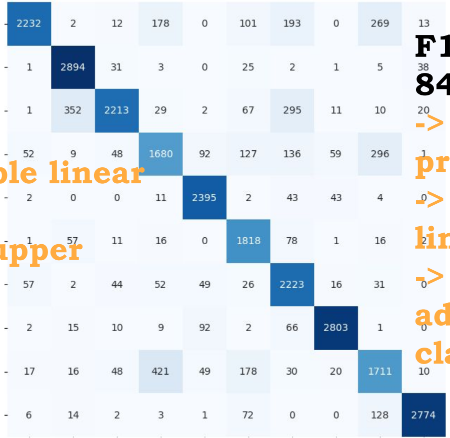
F1-Score:
94% (+51%)
-> Upper limit

16-shots Unified Context Prompt Learning + Zero-shot CLIP



F1-Score:
83% (+40%)
-> Comparable linear probe
-> Close to upper limit

16-shots Class-specified Context Prompt Learning + Zero-shot CLIP



F1-Score:
84% (+41%)
-> Comparable linear probe
-> Close to upper limit
-> No obvious advantage on class-specified

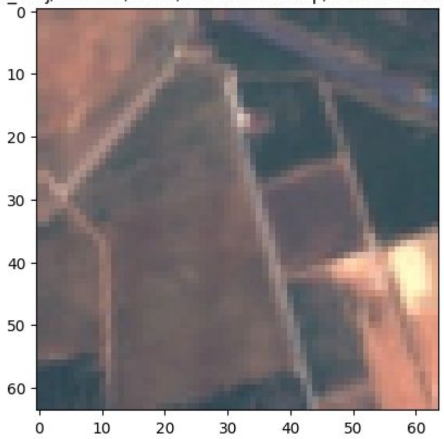
16-shots Conditional Context Prompt Learning + Zero-shot CLIP

Train on 5 classes (½)
Test on all 10 classes

F1-Score:
65% (+22%)
-> Improved generalization compared to baseline

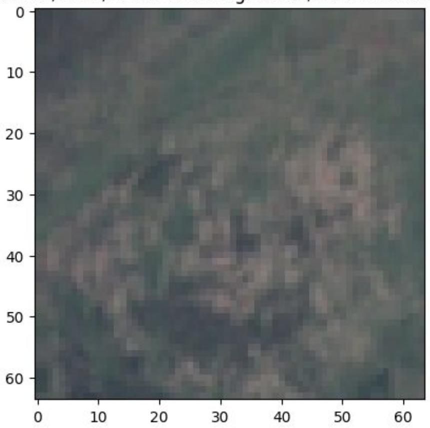
Some Results from Unified Context Prompt

/Data/CV_Proj/eurosat/2750/PermanentCrop/PermanentCrop_195.jpg



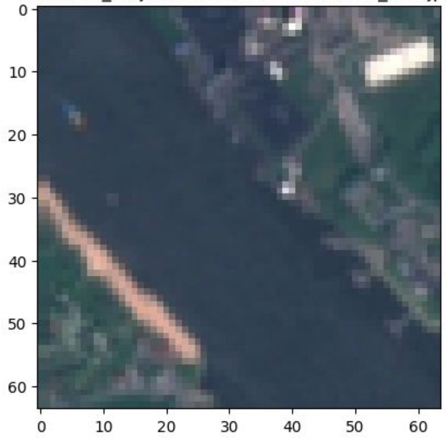

```
[('AnnualCrop', 0.0112),
('Forest', 4e-05),
('HerbaceousVegetation',
0.0004),
('Highway', 0.0322),
('Industrial', 0.0018),
('Pasture', 0.0036),
('PermanentCrop', 0.9502),
('Residential', 0.0002),
('River', 0.0004),
('SeaLake', 5e-06)]
```

/Data/CV_Proj/eurosat/2750/HerbaceousVegetation/HerbaceousVegetation_2069.jpg



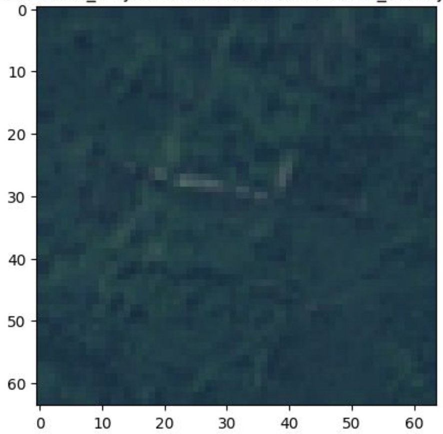

```
[('AnnualCrop', 1e-05),
('Forest', 0.1713),
('HerbaceousVegetation',
0.7921),
('Highway', 0.0010),
('Industrial', 0.0002),
('Pasture', 0.0147),
('PermanentCrop', 0.0164),
('Residential', 0.0030),
('River', 0.0001),
('SeaLake', 0.0012)]
```

/Data/CV_Proj/eurosat/2750/River/River_201.jpg




```
[('AnnualCrop', 0.0002),
('Forest', 0.0032),
('HerbaceousVegetation',
0.0042),
('Highway', 0.0185),
('Industrial', 0.0196),
('Pasture', 0.0365),
('PermanentCrop', 0.0053),
('Residential', 0.0078),
('River', 0.9032),
('SeaLake', 0.0014)]
```

/Data/CV_Proj/eurosat/2750/Forest/Forest_2303.jpg




```
[('AnnualCrop', 0.0001),
('Forest', 0.7211),
('HerbaceousVegetation',
0.0647),
('Highway', 0.0018),
('Industrial', 0.0424),
('Pasture', 0.0999),
('PermanentCrop', 0.0510),
('Residential', 0.0171),
('River', 0.0008),
('SeaLake', 0.0012)]
```

Interpret the Learned Prompt

from Unified Context Prompt

1: ['**decorations**</w>', 'lizards</w>', 'wed</w>', 'dor', 'erin</w>'] ['0.6720', '0.6734', '0.6743', '0.6754', '0.6755']

2: ['**pelo**', 'sculpted</w>', 'lit</w>', 'revol', 'appeared</w>'] ['0.8290', '0.8376', '0.8379', '0.8389', '0.8390']

3: ['jake', 'jw', 'joe', 'kab</w>', 'half'] ['0.7303', '0.7352', '0.7359', '0.7373', '0.7385']

4: ['blames</w>', '**organised**</w>', 'applaud</w>', 'picked</w>', 'implic'] ['0.8232', '0.8240', '0.8244', '0.8251', '0.8258']

5: ['list', '**rotating**</w>', '**represented**</w>', 'strack</w>', 'alline</w>'] ['0.7324', '0.7375', '0.7377', '0.7402', '0.7413']

6: ['knife</w>', 'etu', 'broken', '**foreign**', 'exploding</w>'] ['1.0701', '1.0703', '1.0724', '1.0741', '1.0744']

7: ['sweat', 'stri', 'tall', 'masa', 'yaw'] ['0.9688', '0.9709', '0.9724', '0.9756', '0.9761']

8: ['**linear**</w>', 'optional</w>', 'logical</w>', 'smear</w>', 'phillips</w>'] ['0.8934', '0.8982', '0.9033', '0.9046', '0.9047']

9: ['piss</w>', 'simplified</w>', 'kow</w>', '**modes**</w>', '**calm**'] ['0.8488', '0.8512', '0.8536', '0.8564', '0.8569']

10: ['terri', 'ting', 'newsp', 'cops</w>', '**relocated**</w>'] ['0.9932', '0.9934', '0.9936', '0.9950', '0.9950']

11: ['milb</w>', 'taxpayer</w>', ':|</w>', 'moms', 'nsfw</w>'] ['0.9133', '0.9133', '0.9152', '0.9166', '0.9181']

12: ['tummy</w>', '**residence**</w>', '**retreat**</w>', 'chest</w>', 'anger</w>'] ['0.8954', '0.8981', '0.8997', '0.9000', '0.9003']

13: ['tivity</w>', 'dar</w>', 'ities</w>', '**wood**</w>', '**sight**</w>'] ['0.6240', '0.6274', '0.6319', '0.6323', '0.6347']

14: ['**rooms**</w>', 'officer</w>', '**weather**</w>', '**toward**</w>', '**lighting**</w>'] ['0.6663', '0.6685', '0.6696', '0.6725', '0.6729']

15: ['ilove', 'ignored</w>', 'rig</w>', 'vi</w>', 'gamer</w>'] ['0.7930', '0.7935', '0.7953', '0.7959', '0.7962']

16: ['**addresses**</w>', '**cyber**</w>', 'aring</w>', '´', 'irl</w>'] ['0.8209', '0.8216', '0.8233', '0.8234', '0.8237']

Problem with continuous prompt vector learning: **Hard to interpret the result**

Searching within the vocabulary for words that **closest to the learned vector**, by Euclidean distance

Conclusion

Conclusion

1. CLIP zero-shot on EuroSAT is not good (actually in the paper, among the 27 datasets, EuroSAT is one of the hardest to CLIP)

2. Linear probe can improve the performance significantly with only 16 shots, which proves CLIP already learn the generalized image representation

3. Fully train the linear head on the whole dataset can push the performance to nearly 95%

   -> Fully capable for potential downstream task

4. Prompt learning is an auxiliary approach to deploy CLIP to downstream task, performance similar to linear probe using same shots learning

   Advantage to linear probe:

   a. More dynamic output and generalization potential
   b. Better transparency and explainability

Reference

[1] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.

[2] Kim K, Laskin M, Mordatch I, et al. How to adapt your large-scale vision-and-language model[J]. 2021.

[3] Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

[4] Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.