



预训练模型 II

赵洲

浙江大学计算机学院 副教授

CLIP的研究动机

- CLIP解决传统监督预训练模型的**高标注**和**泛化弱**问题。
- 互联网上存在大量图像文本对，且样本本身差异性大，不仅解决数据高标注，且容易获得泛化能力强模型。

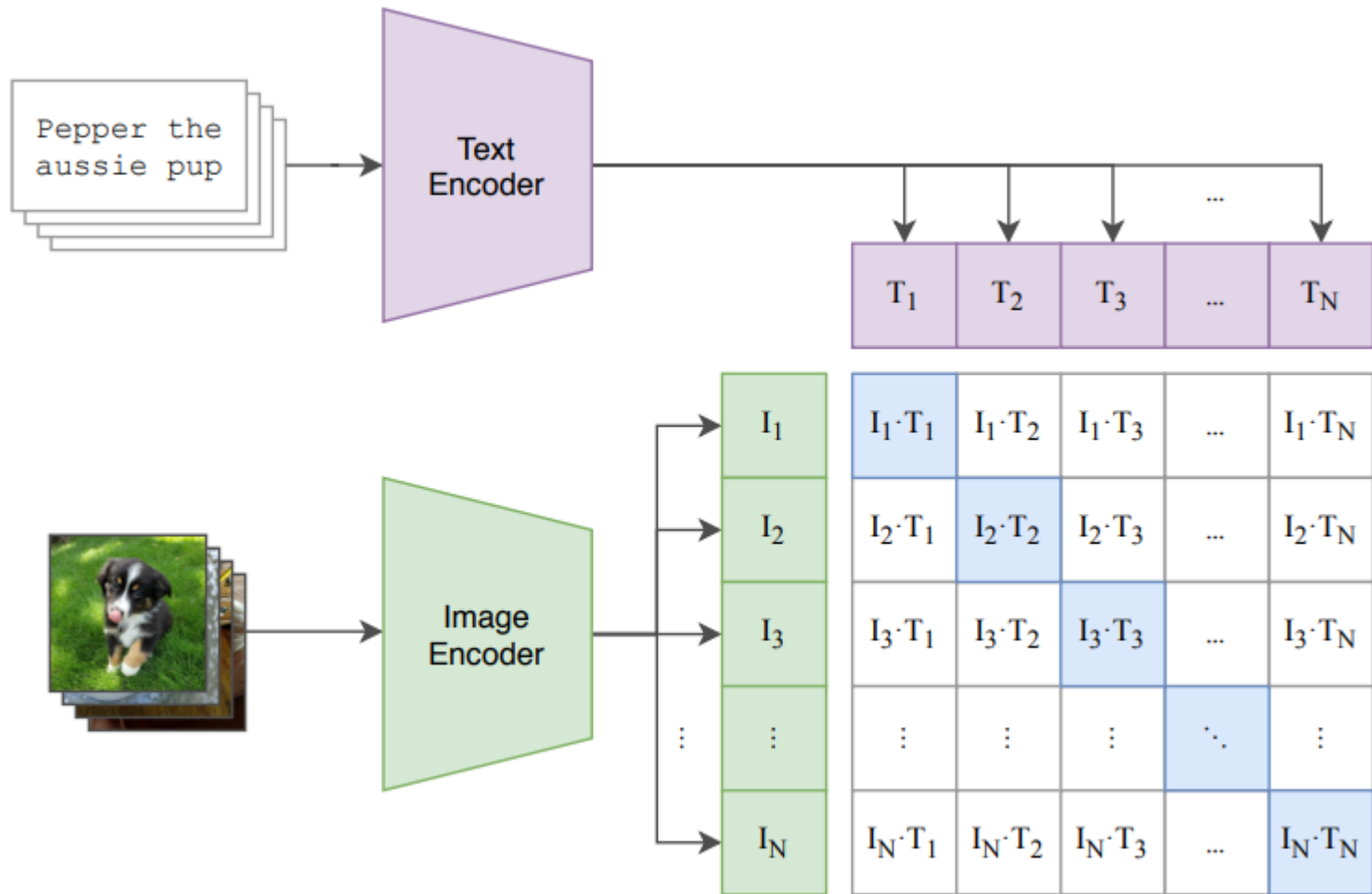


Full ImageNet Pre-Trained
Model

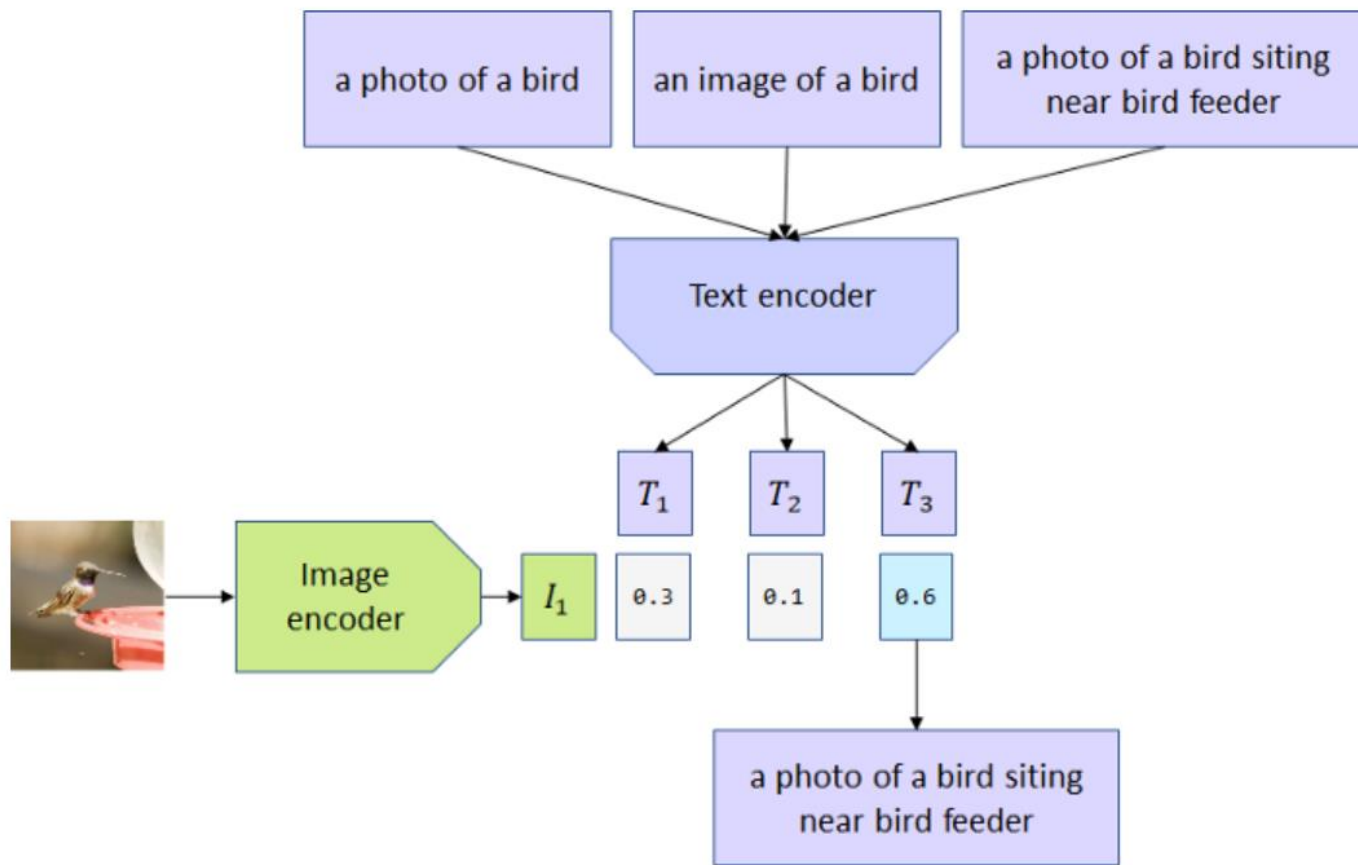


"Panda"

CLIP模型训练



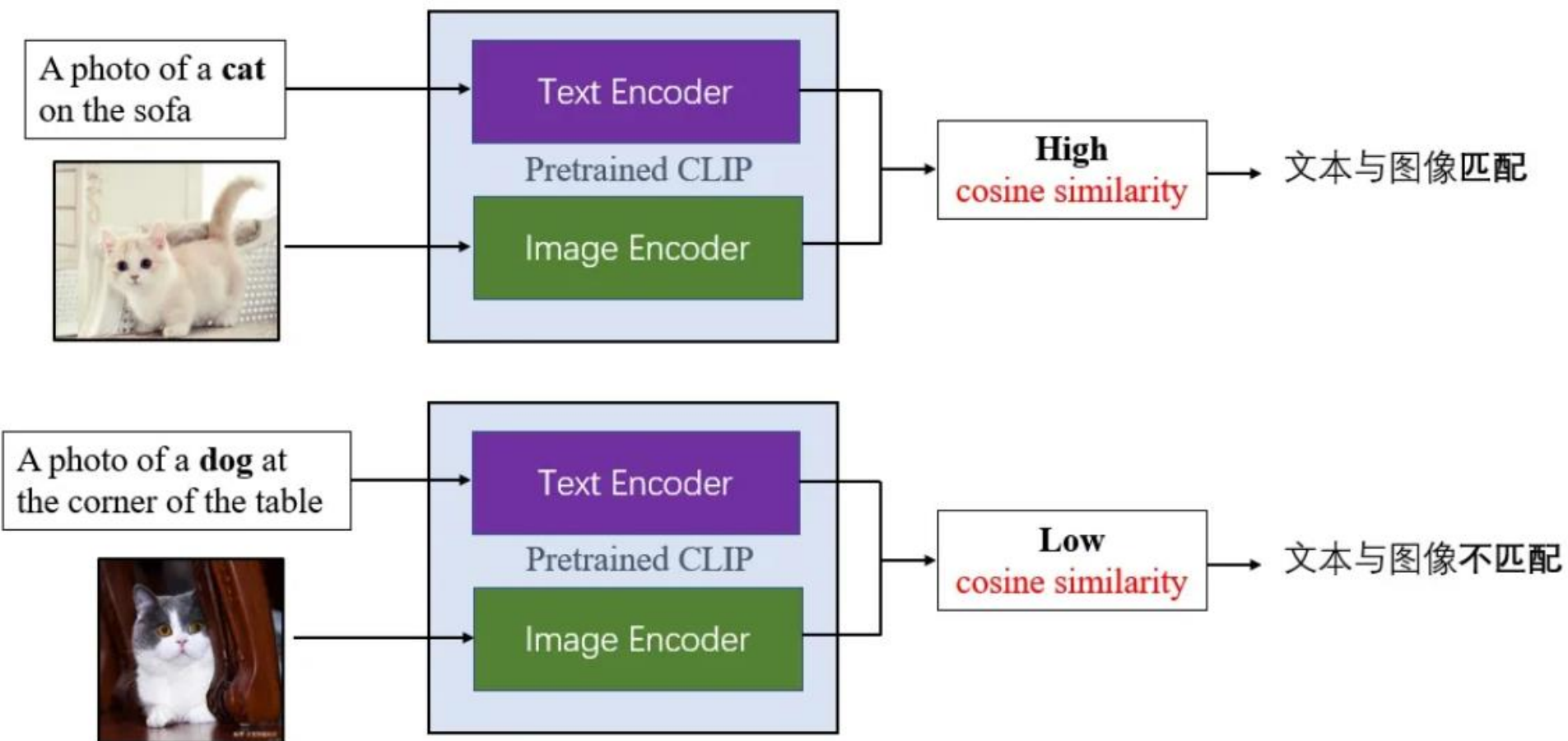
CLIP模型数学原理



■优化目标

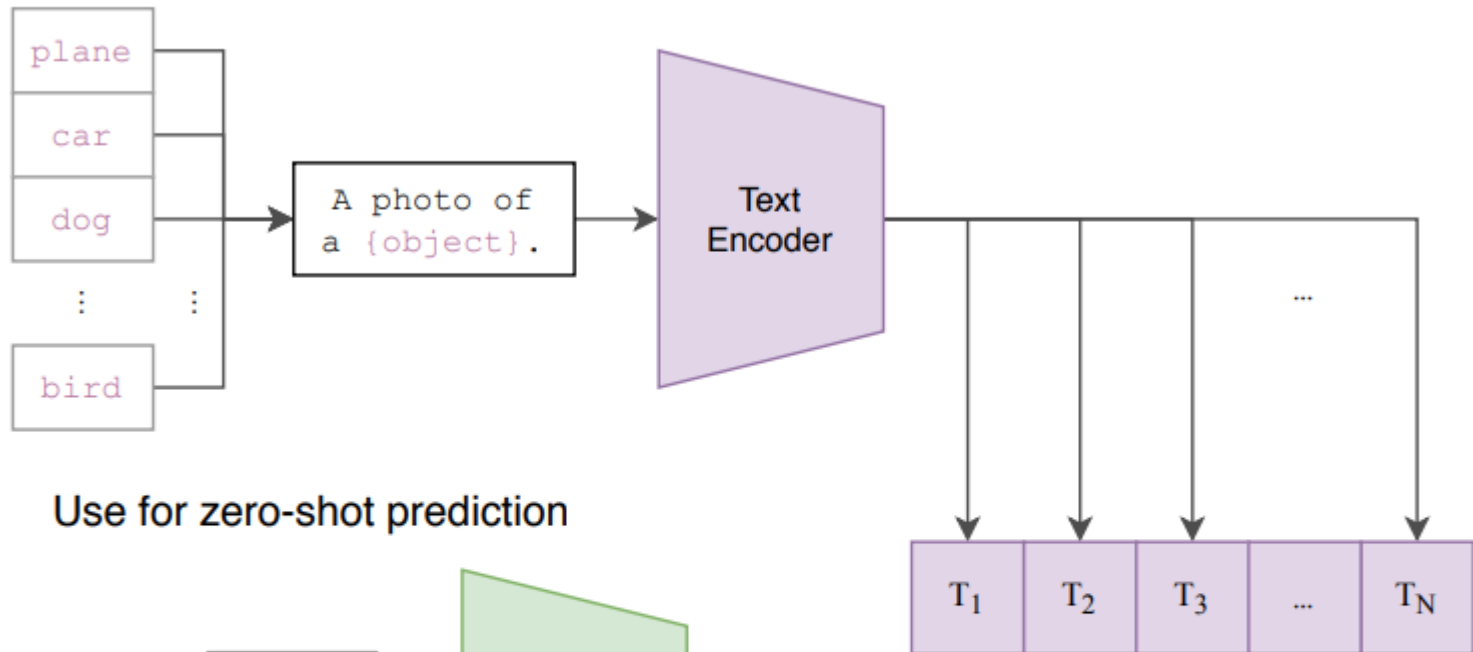
$$\min \left(\sum_{i=1}^N \sum_{j=1}^N (I_i \cdot T_j)_{(i \neq j)} - \sum_{i=1}^N (I_i \cdot T_i) \right)$$

CLIP模型推理

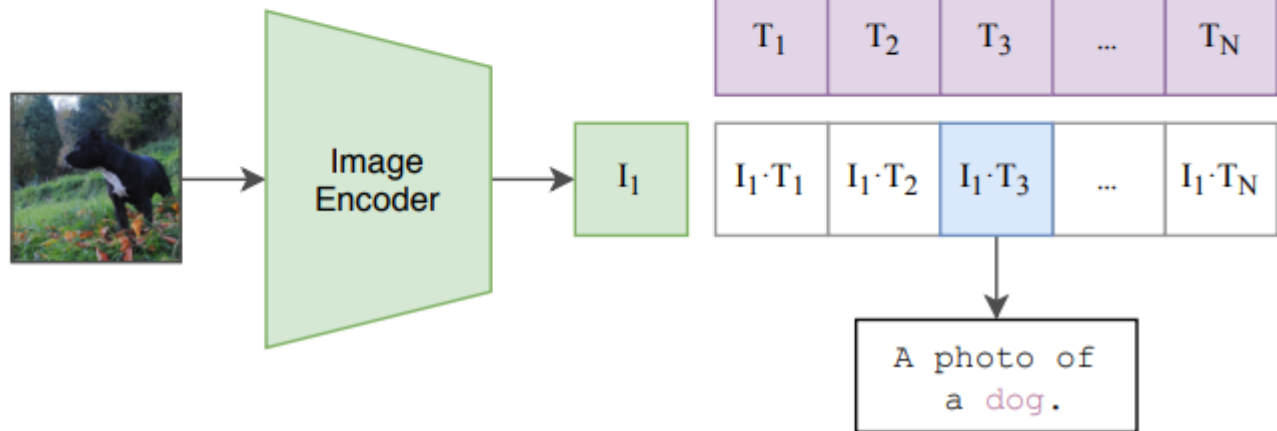


迁移CLIP模型到分类

Create dataset classifier from label text



Use for zero-shot prediction



CLIP分类实验结果

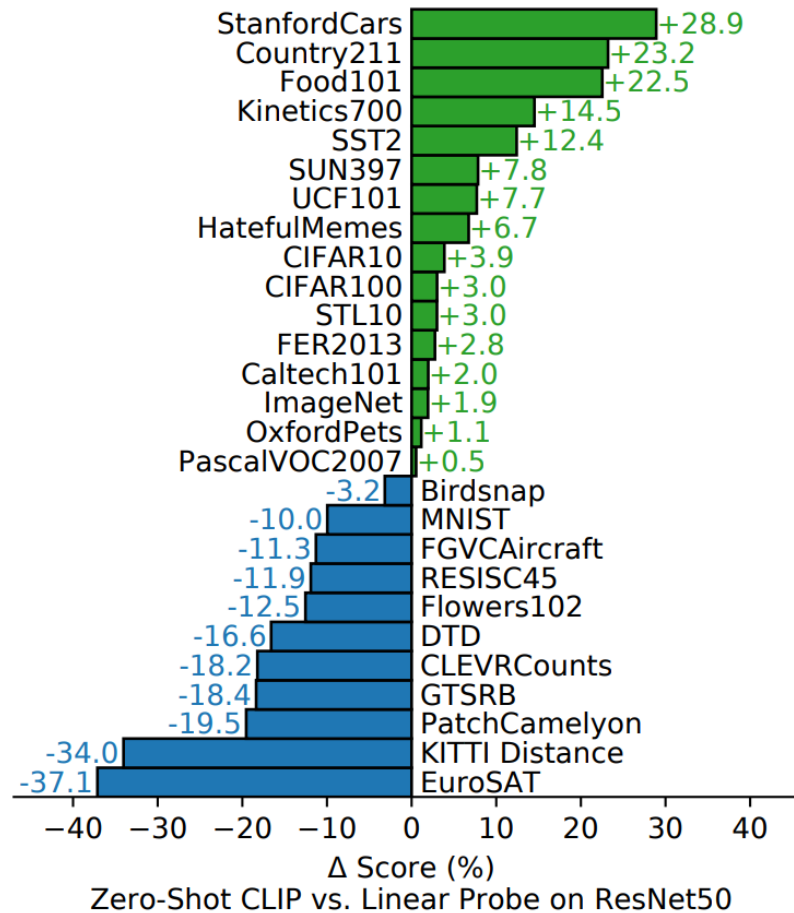


Figure 4. Zero-shot CLIP is competitive with a fully supervised baseline. Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet50 features on 16 datasets, including ImageNet.

few-shot CLIP实验结果

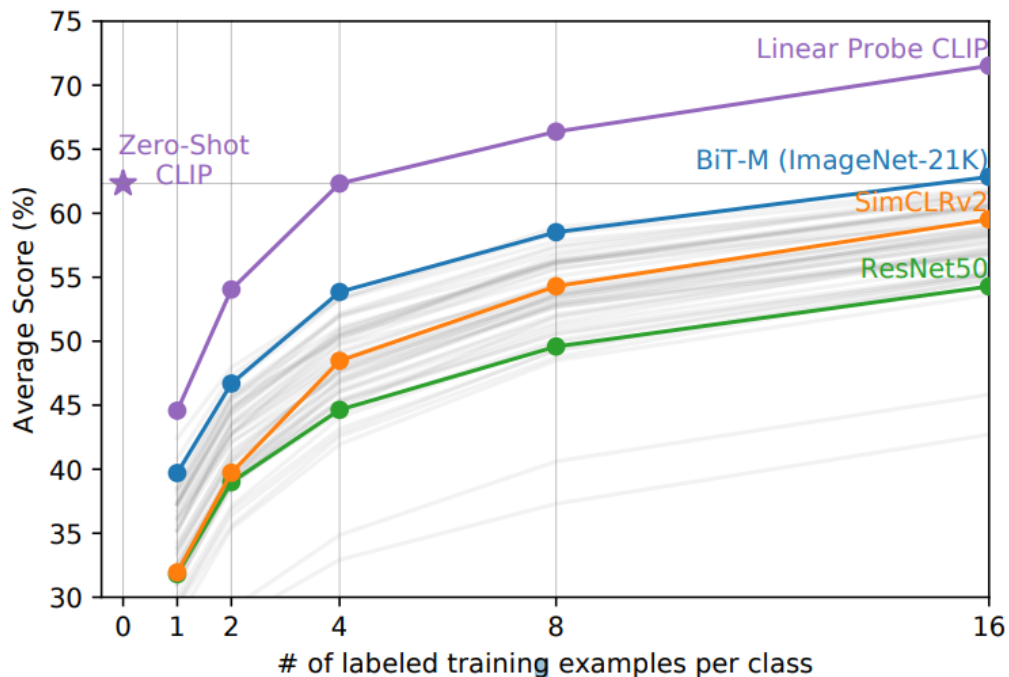








Figure 5. Zero-shot CLIP outperforms few-shot linear probes. Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, the best performing model is highlighted. Light gray lines are other models in the eval suite. The 20 datasets with at least 16 examples per class were used in this analysis.

CLIP泛化实验结果

| | Dataset Examples | ImageNet ResNet101 | Zero-Shot CLIP | Δ Score |
|-----------------|--|-----------------------|-------------------|----------------|
| ImageNet |  | 76.2 | 76.2 | 0% |
| ImageNetV2 |  | 64.3 | 70.1 | +5.8% |
| ImageNet-R |  | 37.7 | 88.9 | +51.2% |
| ObjectNet |  | 32.6 | 72.3 | +39.7% |
| ImageNet Sketch |  | 25.2 | 60.2 | +35.0% |
| ImageNet-A |  | 2.7 | 77.1 | +74.4% |

CLIP模型分类例子

```
# Prepare the inputs
image, class_id = cifar100[3637]
image_input = preprocess(image).unsqueeze(0).to(device)
text_inputs = torch.cat([clip.tokenize(f"a photo of a {c}") for c in cifar100.classes]).to(device)
#cifar 每个类别，输入图片，检索匹配类别

# Calculate features
with torch.no_grad():
    image_features = model.encode_image(image_input)
    text_features = model.encode_text(text_inputs)

# Pick the top 5 most similar labels for the image
image_features /= image_features.norm(dim=-1, keepdim=True)
text_features /= text_features.norm(dim=-1, keepdim=True)
similarity = (100.0 * image_features @ text_features.T).softmax(dim=-1)
values, indices = similarity[0].topk(5)
```

ALIGN的研究动机



“motorcycle front wheel”



“*thumbnail for version as of 21
57 29 june 2010*”



“file frankfurt airport
skyline 2017 05 jpg”



“file london barge race 2 jpg”



“moustache seamless
wallpaper design”



“st oswalds way and shops”

Figure 2. Example image-text pairs randomly sampled from the training dataset of ALIGN. One clearly noisy text annotation is marked in *italics*.

ALIGN模型训练

■ 图片-文本分类

$$L_{i2t} = -\frac{1}{N} \sum_i \log \frac{\exp(x_i^\top y_i / \sigma)}{\sum_{j=1}^N \exp(x_i^\top y_j / \sigma)}$$

■ 文本-图片分类

$$L_{t2i} = -\frac{1}{N} \sum_i \log \frac{\exp(y_i^\top x_i / \sigma)}{\sum_{j=1}^N \exp(y_i^\top x_j / \sigma)}$$

检索结果

Table 1. Image-text retrieval results on Flickr30K and MSCOCO datasets (zero-shot and fine-tuned). ALIGN is compared with ImageBERT (Qi et al., 2020), UNITER (Chen et al., 2020c), CLIP (Radford et al., 2021), GPO (Chen et al., 2020a), ERNIE-ViL (Yu et al., 2020), VILLA (Gan et al., 2020), and Oscar (Li et al., 2020).

| | | Flickr30K (1K test set) | | | | | | MSCOCO (5K test set) | | | | | |
|------------|-----------|-------------------------|-------------|--------------|--------------|-------------|-------------|----------------------|-------------|-------------|--------------|-------------|-------------|
| | | image → text | | | text → image | | | image → text | | | text → image | | |
| Zero-shot | ImageBERT | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| | UNITER | 70.7 | 90.2 | 94.0 | 54.3 | 79.6 | 87.5 | 44.0 | 71.2 | 80.4 | 32.3 | 59.0 | 70.2 |
| | CLIP | 83.6 | 95.7 | 97.7 | 68.7 | 89.2 | 93.9 | - | - | - | - | - | - |
| | ALIGN | 88.0 | 98.7 | 99.4 | 68.7 | 90.6 | 95.2 | 58.4 | 81.5 | 88.1 | 37.8 | 62.4 | 72.2 |
| Fine-tuned | | 88.6 | 98.7 | 99.7 | 75.7 | 93.8 | 96.8 | 58.6 | 83.0 | 89.7 | 45.6 | 69.8 | 78.6 |
| | GPO | 88.7 | 98.9 | 99.8 | 76.1 | 94.5 | 97.1 | 68.1 | 90.2 | - | 52.7 | 80.2 | - |
| | UNITER | 87.3 | 98.0 | 99.2 | 75.6 | 94.1 | 96.8 | 65.7 | 88.6 | 93.8 | 52.9 | 79.9 | 88.0 |
| | ERNIE-ViL | 88.1 | 98.0 | 99.2 | 76.7 | 93.6 | 96.4 | - | - | - | - | - | - |
| | VILLA | 87.9 | 97.5 | 98.8 | 76.3 | 94.2 | 96.8 | - | - | - | - | - | - |
| | Oscar | - | - | - | - | - | - | 73.5 | 92.2 | 96.0 | 57.5 | 82.8 | 89.8 |
| | ALIGN | 95.3 | 99.8 | 100.0 | 84.9 | 97.4 | 98.6 | 77.0 | 93.5 | 96.9 | 59.9 | 83.3 | 89.8 |

分类结果

Table 4. Top-1 Accuracy of zero-shot transfer of ALIGN to image classification on ImageNet and its variants.

| Model | ImageNet | ImageNet-R | ImageNet-A | ImageNet-V2 |
|--------------|-------------|-------------|-------------|-------------|
| CLIP | 76.2 | 88.9 | 77.2 | 70.1 |
| ALIGN | 76.4 | 92.2 | 75.8 | 70.1 |

多模态检索结果

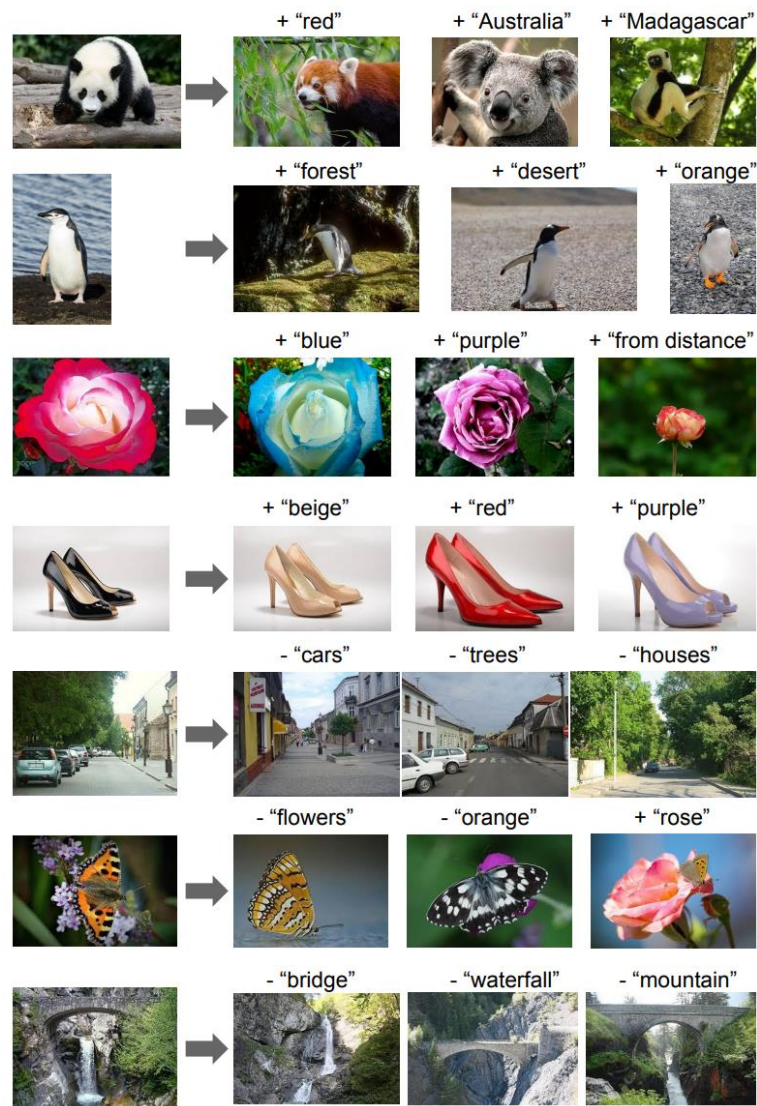


Figure 5. Image retrieval with image \pm text queries. We add (or

细粒度检索结果

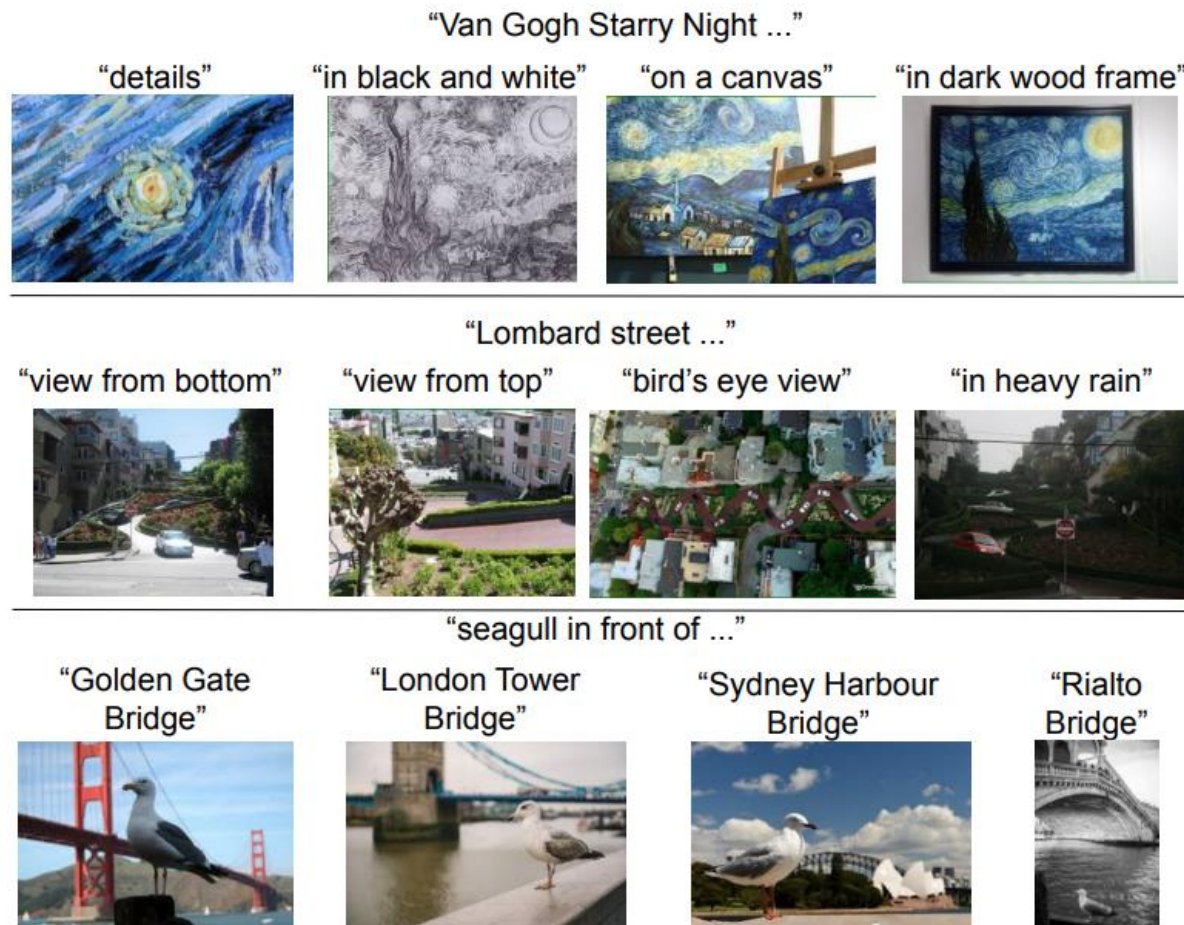
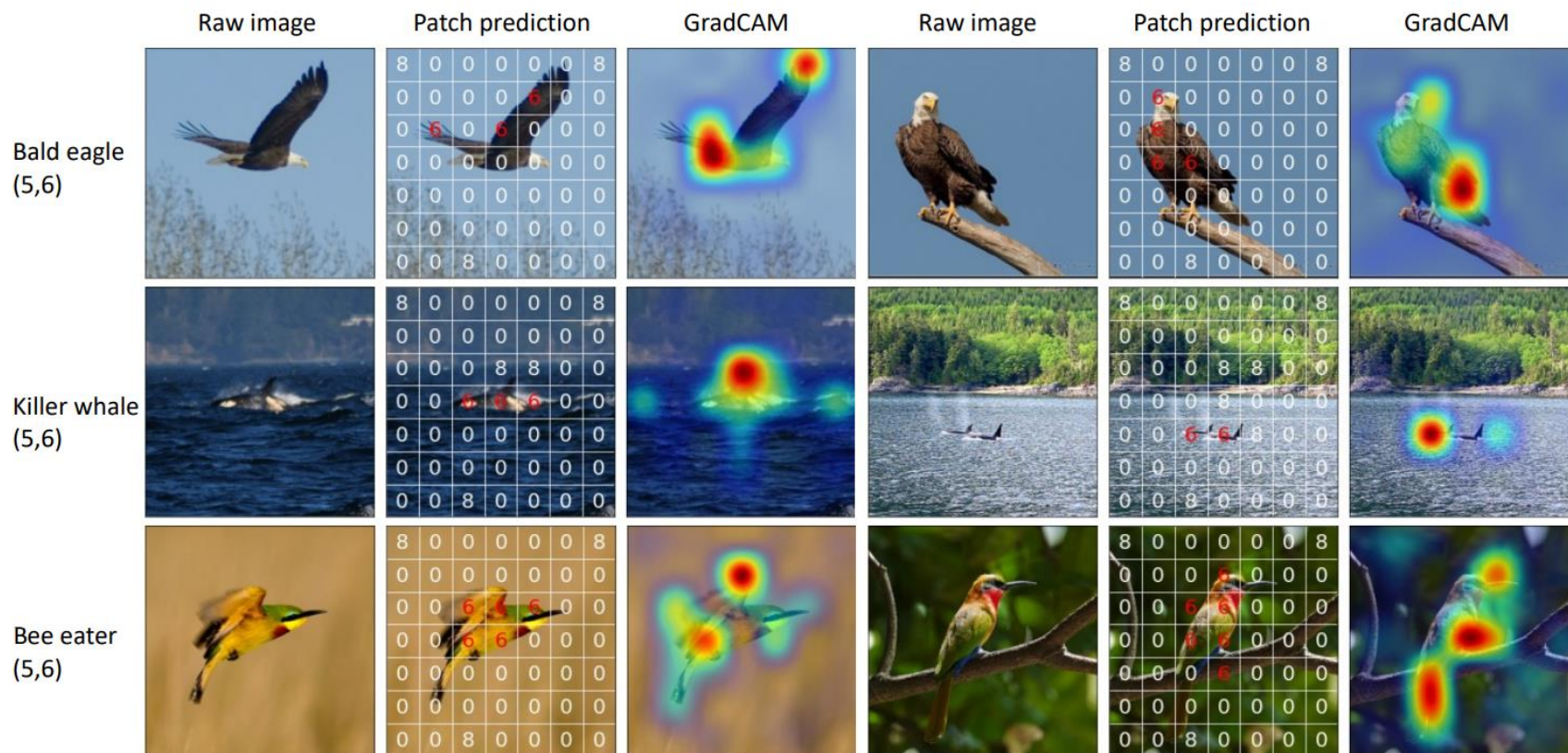


Figure 4. Image retrieval with fine-grained text queries using ALIGN’s embeddings.

FILIP的研究动机

- CLIP基于各模态的全局特征间的相似度进行建模。



FILIP模型训练

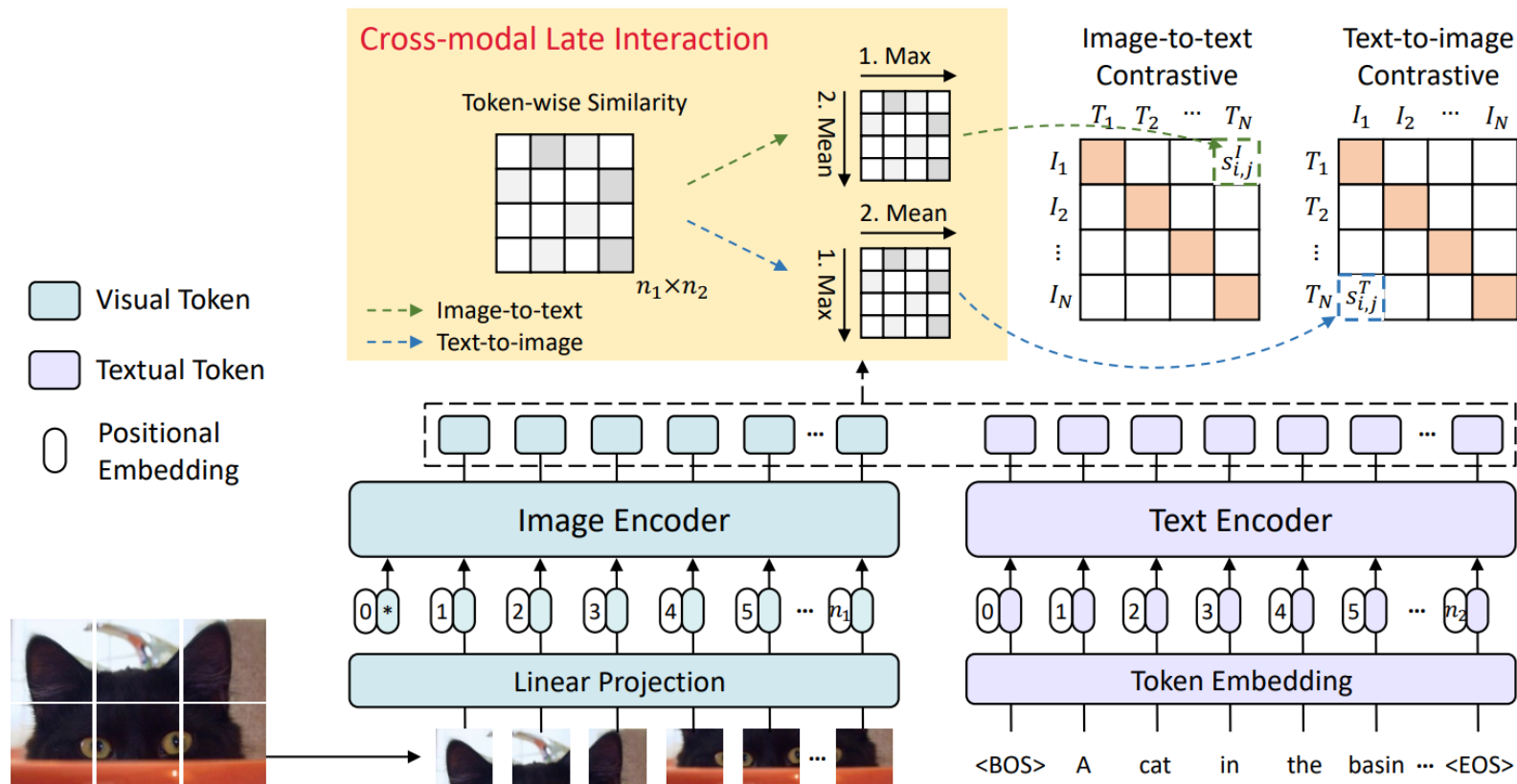


Figure 1: Overall architecture of FILIP, a dual-stream model with Transformer-based image and text encoders. On top of the image and text encoders, the representations of textual tokens and visual tokens are linearly projected to the multi-modal joint space. A novel fine-grained contrastive learning equipped with cross-modal late interaction is proposed, which uses a token-wise maximum similarity between visual and textual tokens.

FILIP数学原理

■ 全局特征

$$\mathcal{L}_k^I(\mathbf{x}_k^I, \{\mathbf{x}_j^T\}_{j=1}^b) = -\frac{1}{b} \log \frac{\exp(s_{k,k}^I)}{\sum_j \exp(s_{k,j}^I)}$$

$$\mathcal{L}_k^T(\mathbf{x}_k^T, \{\mathbf{x}_j^I\}_{j=1}^b) = -\frac{1}{b} \log \frac{\exp(s_{k,k}^T)}{\sum_j \exp(s_{j,k}^T)}$$

$$\mathcal{L} = \frac{1}{2} \sum_{k=1}^b (\mathcal{L}_k^I + \mathcal{L}_k^T)$$

■ 局部特征

$$\max_{0 \leq r < n_2} [f_\theta(\mathbf{x}_i^I)]_k^\top [g_\phi(\mathbf{x}_j^T)]_r$$

$$s_{i,j}^I(\mathbf{x}_i^I, \mathbf{x}_j^T) = \frac{1}{n_1} \sum_{k=1}^{n_1} [f_\theta(\mathbf{x}_i^I)]_k^\top [g_\phi(\mathbf{x}_j^T)]_{m_k^I}$$

$$s_{i,j}^T(\mathbf{x}_i^I, \mathbf{x}_j^T) = \frac{1}{n_2} \sum_{k=1}^{n_2} [f_\theta(\mathbf{x}_i^I)]_{m_k^T}^\top [g_\phi(\mathbf{x}_j^T)]_k$$

FILIP分类结果

Table 1: Top-1 accuracy(%) of zero-shot image classification on 12 datasets. Our FILIP can boost 3~5% accuracy on average.

| | CIFAR10 | CIFAR100 | Caltech101 | StanfordCars | Flowers102 | Food101 | SUN397 | DTD | Aircrafts | OxfordPets | EuroSAT | ImageNet | Average |
|----------------------------------|---------|----------|------------|--------------|------------|---------|--------|------|-----------|------------|---------|----------|-----------------------------|
| CLIP-ViT-B/32 | 91.3 | 65.1 | 87.9 | 59.4 | 66.7 | 84.4 | 63.2 | 44.5 | 21.2 | 87.0 | 49.4 | 63.2 | 65.3 |
| FILIP _{base} -ViT-B/32 | 86.9 | 65.5 | 91.9 | 55.4 | 85.3 | 82.8 | 69.1 | 49.3 | 57.2 | 88.1 | 49.9 | 68.8 | 70.9 ^{+5.6} |
| CLIP-ViT-L/14 | 96.2 | 77.9 | 92.6 | 77.3 | 78.7 | 92.9 | 67.7 | 55.3 | 36.1 | 93.5 | 59.9 | 75.3 | 75.3 |
| FILIP _{large} -ViT-L/14 | 95.7 | 75.3 | 93.0 | 70.8 | 90.1 | 92.2 | 73.1 | 60.7 | 60.2 | 92 | 59.2 | 77.1 | 78.3 ^{+3.0} |

FILIP检索结果

Table 2: Results of zero-shot image-text retrieval on Flickr30K and MSCOCO datasets. The last two rows (marked with *) report the zero-shot results on Flickr30K dataset of model fine-tuned on MSCOCO dataset, following the setting of ALBEF (Li et al., 2021a).

| | Flickr30K | | | | | | MSCOCO | | | | | |
|---------------|---------------|-------------|--------------|---------------|-------------|-------------|---------------|-------------|-------------|---------------|-------------|-------------|
| | image-to-text | | | text-to-image | | | image-to-text | | | text-to-image | | |
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| Unicoder-VL | 64.3 | 85.8 | 92.3 | 48.4 | 76.0 | 85.2 | — | — | — | — | — | — |
| ImageBERT | 70.7 | 90.2 | 94.0 | 54.3 | 79.6 | 87.5 | 44.0 | 71.2 | 80.4 | 32.3 | 59.0 | 70.2 |
| UNITER | 83.6 | 95.7 | 97.7 | 68.7 | 89.2 | 93.9 | — | — | — | — | — | — |
| CLIP | 88.0 | 98.7 | 99.4 | 68.7 | 90.6 | 95.2 | 58.4 | 81.5 | 88.1 | 37.8 | 62.4 | 72.2 |
| ALIGN | 88.6 | 98.7 | 99.7 | 75.7 | 93.8 | 96.8 | 58.6 | 83.0 | 89.7 | 45.6 | 69.8 | 78.6 |
| FILIP | 89.8 | 99.2 | 99.8 | 75.0 | 93.4 | 96.3 | 61.3 | 84.3 | 90.4 | 45.9 | 70.6 | 79.3 |
| ALBEF* | 94.1 | 99.5 | 99.7 | 82.8 | 96.3 | 98.1 | — | — | — | — | — | — |
| FILIP* | 95.4 | 99.8 | 100.0 | 84.7 | 97.0 | 98.7 | — | — | — | — | — | — |

可视化例子

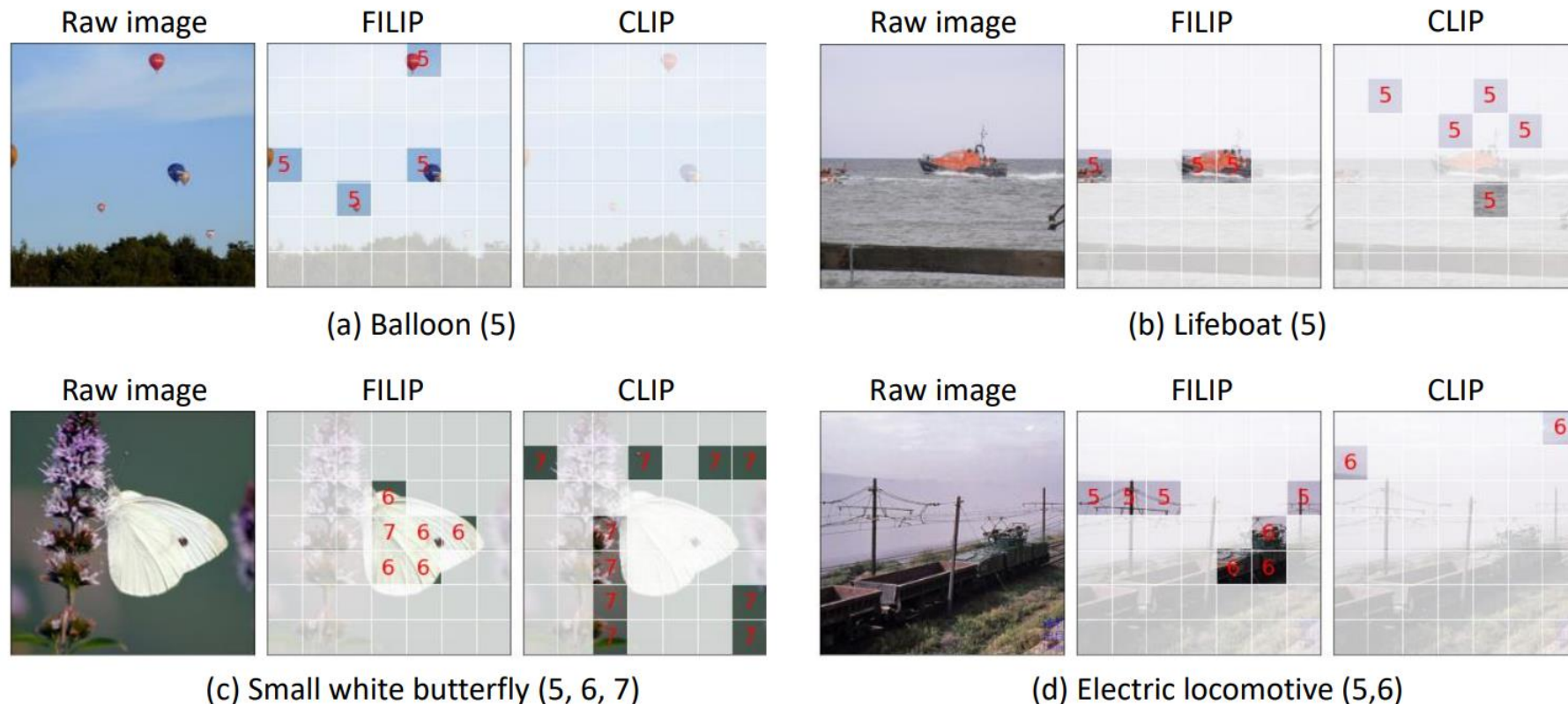
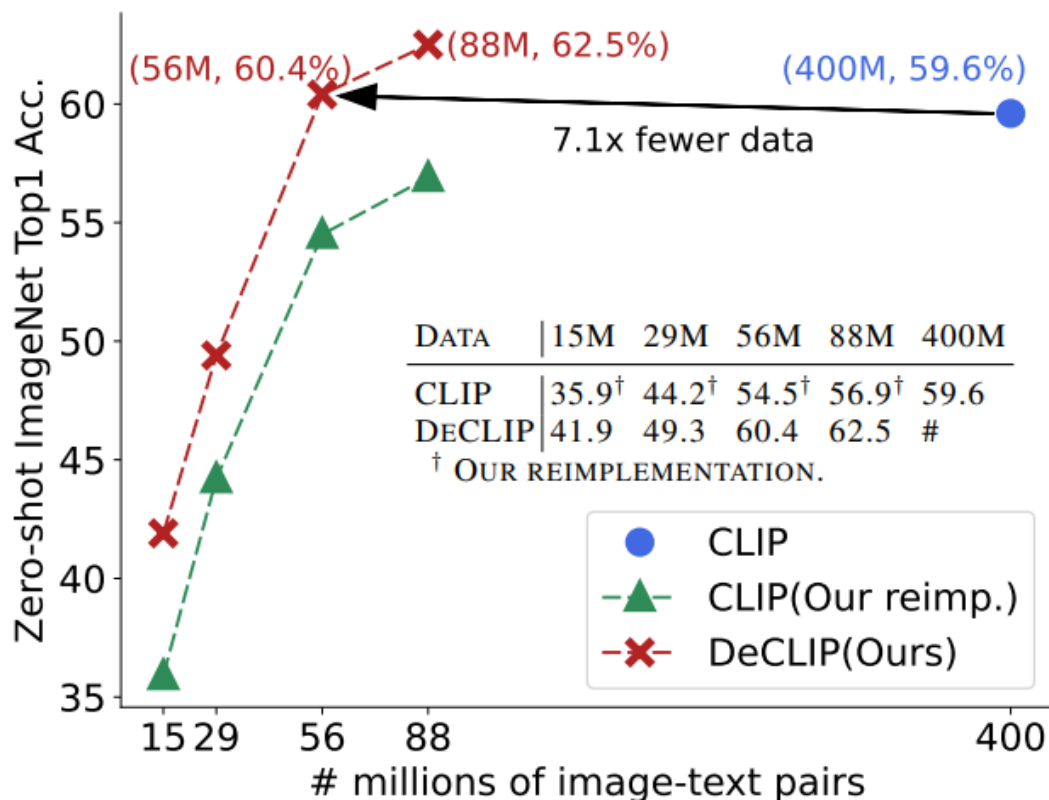


Figure 2: Visualizations of word-patch alignment for 4 classes of the ImageNet dataset and “a photo of a {label}.” is the prompt. Numbers in the parentheses after the class label indicate the location indices of the class label in the tokenized textual sequence. The correct predictions are highlighted by opaque patches with the class label indices in red.

DECLIP的研究动机

- CLIP使用了**4亿**图像-文本对数据，是否存在较少数据夏依旧可以取得不错效果的方法？



最近邻监督 (NNS)



Origin

actor opted for a loose side - parted hairstyle when she attended event .



going to see a lot of vintage tractors this week

NN

actor wore her hair in face - framing layers during the show .

vintage at tractors a gathering

DECLIP模型训练

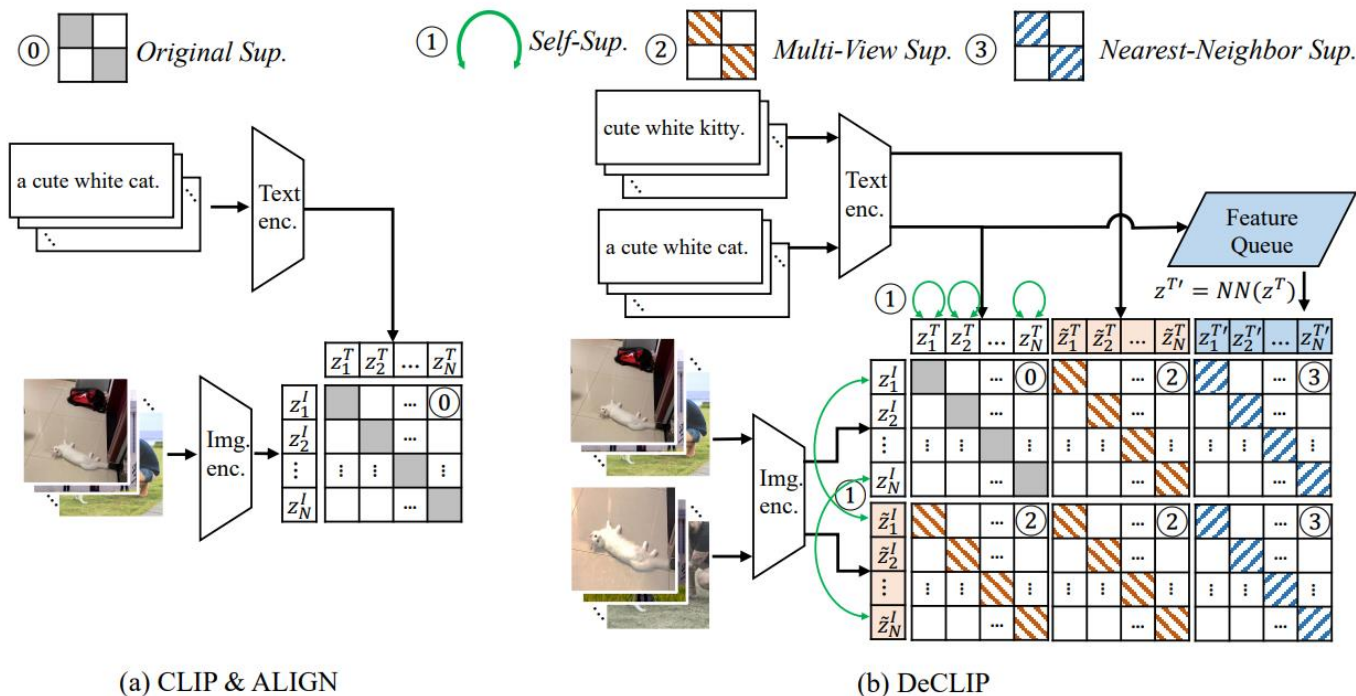
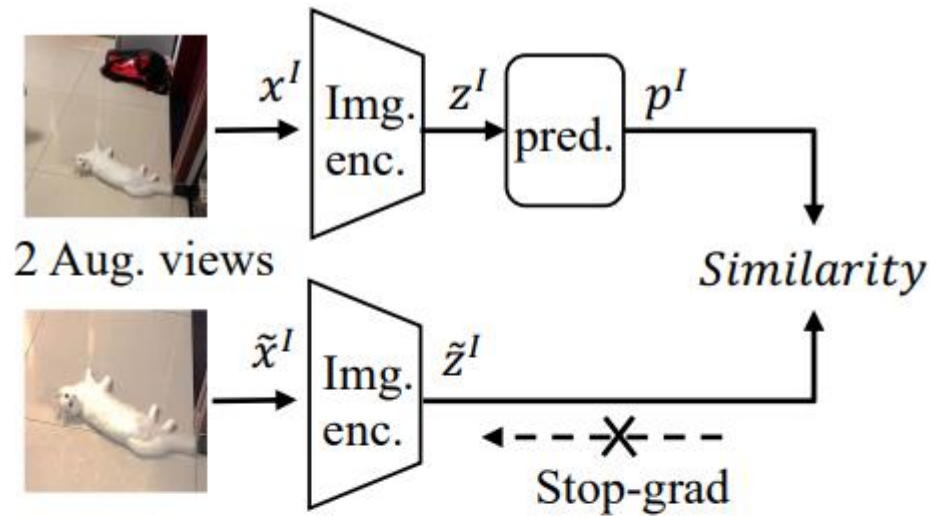
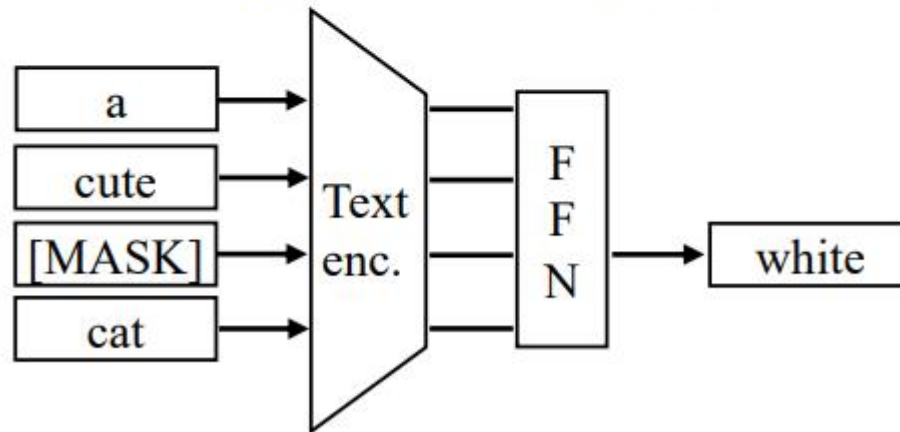


Figure 4: (a) CLIP and ALIGN jointly train an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. (b) Our DeCLIP overview. ① means **Self-Supervision(SS)**. For image SS, we maximize the similarity between two augmented views of the same instance. For text SS, we leverage Masked Language Modeling(MLM) within a text sentence. ② represents cross-modal **Multi-View Supervision(MVS)**. We first have two augmented views of both image and text, then contrast the 2×2 image-text pairs. ③ indicates **Nearest-Neighbor Supervision(NNS)**. We sample text NN in the embedding space to serve as additional supervision. The combination of the three supervision leads to efficient multi-modal learning.

自监督学习 (SS)



(a) SimSiam Framework



(b) Masked Language Modeling

最近邻监督 (NNS)

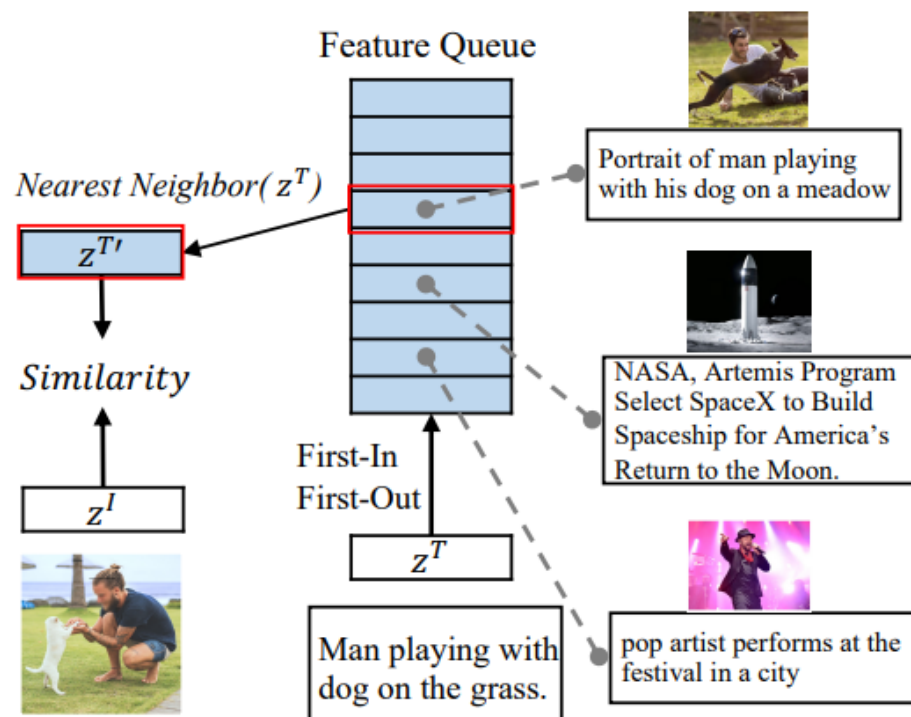


Figure 6: Nearest-Neighbor Supervision. $z^{T'}$ is the NN of feature z^T in the embedding space. $z^{T'}$ will serve as an additional objective for z^I . We use the feature-level nearest neighbor for the text descriptions as the supervision.

分类结果

Table 2: Zero-shot top1 accuracy on ImageNet. Our DeCLIP shows great data-efficiency.

| METHOD | IMAGE ENCODER | # PARAMS | TRAINING SIZE | ZERO-SHOT TOP1 ACC. |
|-------------------|---------------|-------------|--------------------|---------------------|
| CLIP [†] | RESNET50 | 24M | 88M | 56.9 |
| CLIP | RESNET50 | 24M | 400M | 59.6 |
| DECLIP | RESNET50 | 24M | 88M(↓ 4.5×) | 62.5(↑ +2.9) |
| CLIP | RESNET101 | 42M | 400M | 62.2 |
| CLIP [†] | ViT-B/32 | 88M | 88M | 57.4 |
| CLIP | ViT-B/32 | 88M | 400M | 63.2 |
| DECLIP | ViT-B/32 | 88M | 88M(↓ 4.5×) | 66.2(↑ +3.0) |
| CLIP | RESNET50×64 | 291M | 400M | 73.6 |
| DECLIP | REGNETY-64GF | 276M | 88M(↓ 4.5×) | 73.7(↑ +0.1) |

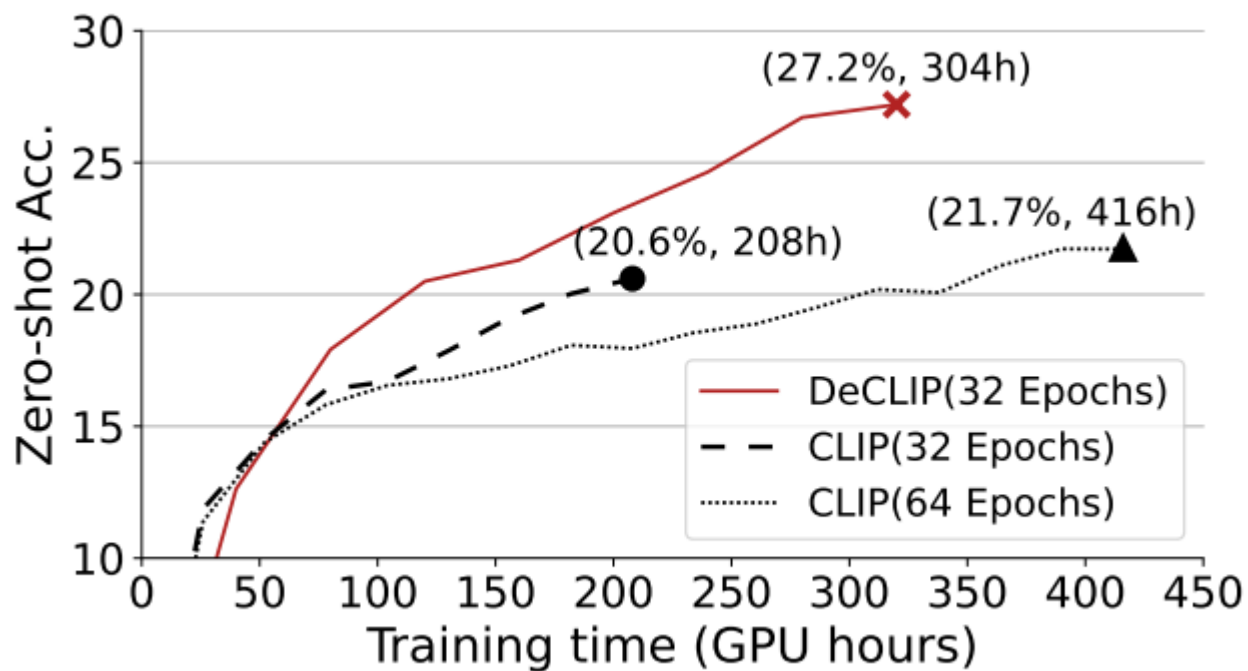
[†] OUR REIMPLEMENTATION.

消融实验结果

Table 4: Ablation on additional supervision. SS/MVS/NNS denotes Self-Supervision, Multi-View Supervision and Nearest-Neighbor Supervision, respectively.

| CLIP | MVS | SS | NNS | ZERO-SHOT |
|------|-----|----|-----|---------------------|
| ✓ | × | × | × | 20.6 |
| ✓ | ✓ | × | × | 24.8(↑ +4.2) |
| ✓ | ✓ | ✓ | × | 25.4(↑ +4.8) |
| ✓ | ✓ | ✓ | ✓ | 27.2(↑ +6.6) |

训练时间结果



数据NN可视化

Original pair



fans show their support during olympic games



lilies and roses on a blue background



Find the Saltine Slim Fit T-Shirt



The person - cookies and cream ice blended



Reading chair



Woolworth Building The Woolworth Building will ki...



Surfing in Sea Kayaks | SKILS



Black Solid Icon For Wander, Rove And Peregrinate Stock Vector...

NN pair



fans show their support during country



pattern with flowers on a blue background



the ting tings slim fit t - shirt



cookies & cream and nutella waffle cream in a ice cone



the special chair



cathedral of learning exterior floors a wonderful art deco neo got...



stand up paddle boards california | pau hana surf supply



knockdown in action icon . element of fight for mobile concept...

CC3M

CC12M

YFCC

Web-crawled