

A Practitioner's Guide to CLIP

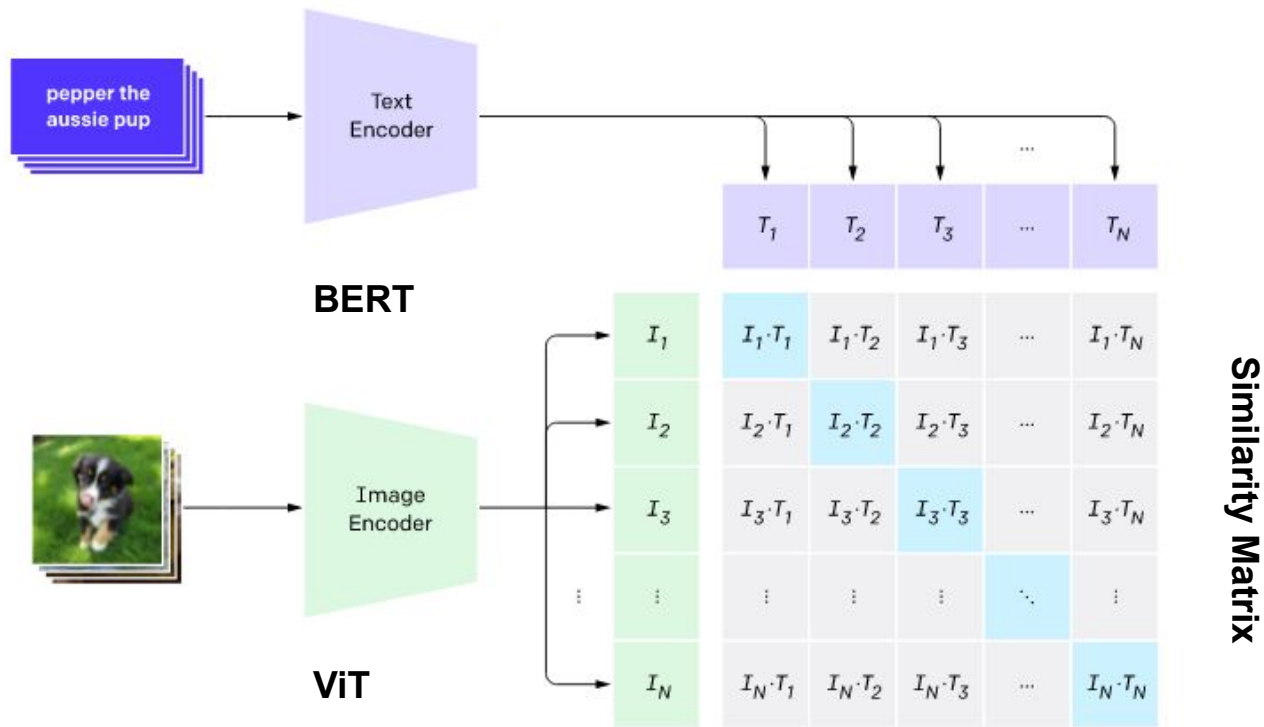
Presenter: Mert Kilickaya

Research Question

How can we induce useful visual & textual embeddings by using freely available Web image captions?

CLIP

1. Contrastive pre-training

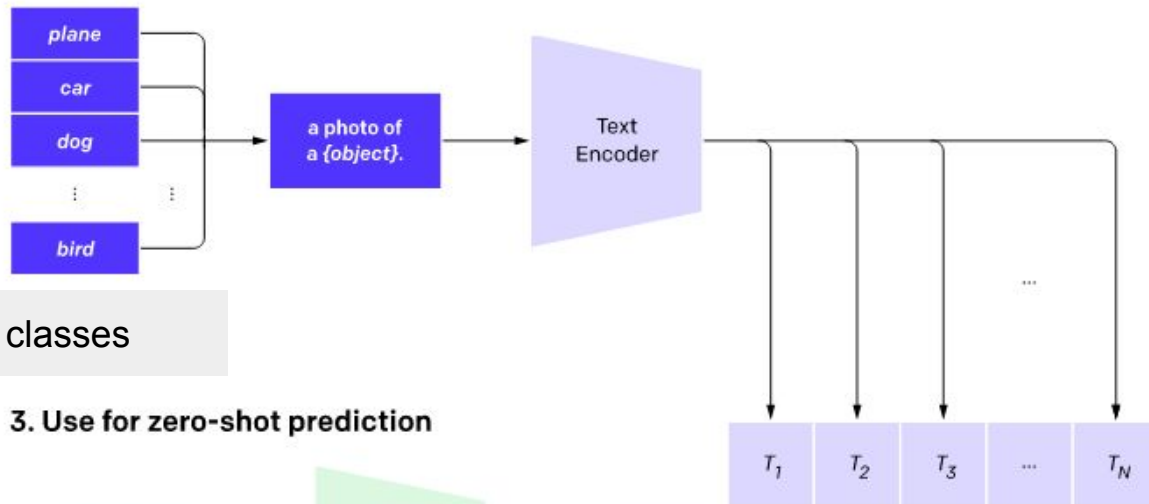


Idea: Given N image-caption pairs, predict image's own caption

How: Force on-diagonal values to be **1s**, and all off-diagonal to be **0s**

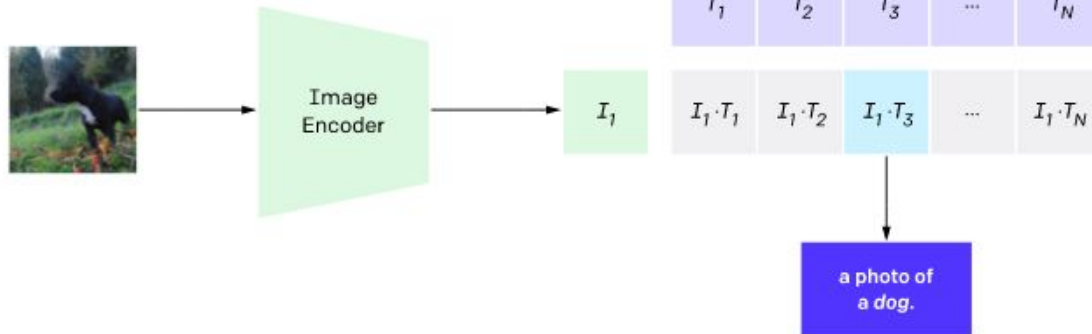
CLIP: Inference

2. Create dataset classifier from label text



Prompt the text encoder with your classes

3. Use for zero-shot prediction



Classify by computing the dot-product between image encoding and all prompts

CLIP outperforms ResNet on 16 benchmarks (of 27)

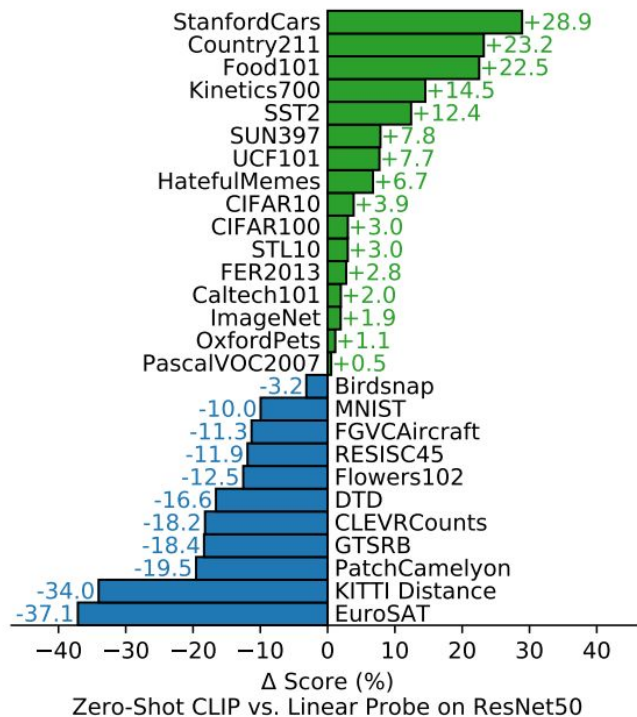








Figure 5. Zero-shot CLIP is competitive with a fully supervised baseline. Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

No-fine-tuning is applied

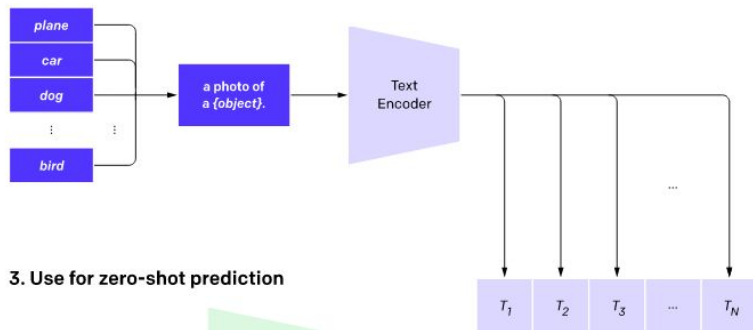
CLIP is highly robust against distributional shifts (ImageNet)

	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%

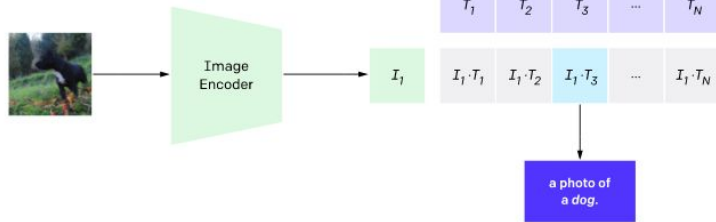
Zero-shot CLIP exhibits more robust behaviour on ImageNet vs. a fully supervised ResNet on ImageNet

CLIP: Fine-tuning for your task

2. Create dataset classifier from label text



3. Use for zero-shot prediction



Prompt Engineering: How to best *captionify* your list of categories? Experiment

Then: Typical classifier training with cross entropy (target 1 for the correct prompt, 0 otherwise)

Details: 1) Half-precision training (fp16), 2) ADAM optimizer parameters matter! (beta, epsilon, weight decay)

CLIP: Fine-tuning for your task

Official training code not available!

Simple script: <https://github.com/openai/CLIP/issues/83>

Comprehensive: https://github.com/mlfoundations/open_clip

PyTorch lightning library: <https://github.com/Zasder3/train-CLIP>

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t            - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

CLIP: Very recent advances

State-of-the-Art in Vision and Language tasks via fine-tuning

How Much Can CLIP Benefit Vision-and-Language Tasks?

**Sheng Shen^{*†}, Liunian Harold Li^{*†}, Hao Tan[°], Mohit Bansal[°],
Anna Rohrbach[†], Kai-Wei Chang[†], Zhewei Yao[†] and Kurt Keutzer[†]**
[†]University of California, Berkeley, [‡]University of California, Los Angeles
[°]University of North Carolina at Chapel Hill

{sheng.s, anna.rohrbach, zhewei, keutzer}@berkeley.edu,
{liunian.harold.li, kwchang}@cs.ucla.edu, {haotan, mbansal}@cs.unc.edu

Off-the-shelf metric to evaluate image captioning models

CLIPScore: A Reference-free Evaluation Metric for Image Captioning

Jack Hessel[†] Ari Holtzman[†] Maxwell Forbes^{†‡} Ronan Le Bras[†] Yejin Choi^{†‡}
[†]Allen Institute for AI

[‡]Paul G. Allen School of Computer Science & Engineering, University of Washington
{jackh, ronanlb}@allenai.org {ahai, mbforbes, yejin}@cs.washington.edu

CLIP: Very recent advances

Does language help generalization in vision models?

Benjamin Devillers^{*1}, Bhavin Choksi^{*2}, Romain Bielawski¹, Rufin VanRullen^{1,2}

¹ Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France
{firstname.lastname}@univ-tlse3.fr

² CerCO, CNRS UMR5549, Toulouse
{firstname.lastname}@cnrs.fr

Abstract

Vision models trained on multimodal datasets can benefit from the wide availability of large image-caption datasets. A recent model (CLIP) was found to generalize well in zero-shot and transfer learning settings. This could imply that linguistic or “semantic grounding” confers additional generalization abilities to the visual feature space. Here, we systematically evaluate various multimodal architectures and vision-only models in terms of unsupervised clustering, few-shot learning, transfer learning and adversarial robustness. In each setting, multimodal training produced no additional generalization capability compared to standard supervised visual training. We conclude that work is still required for semantic grounding to help improve vision models.

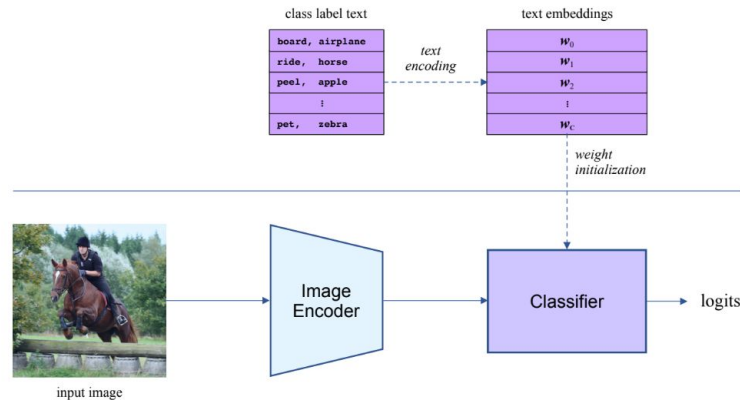


Figure 2: **DEtection FRee (DEFR) HOI recognition pipeline.** It has an image encoder and a linear classifier. The weight of the linear classifier is initialized by w , the text embeddings of class labels encoded by BERT or CLIP’s text encoder. We call this *embedding initialization*, detailed in [subsection 3.2](#). Compared to the detection-supervised approaches, DEFR significantly simplifies the pipeline.

Helps with vision-only tasks as well.

Discussion

Concern: How likely CLIP has already observed images of specific datasets (trained on 400 Million images)?

Dynamic classifier vs. Static classifier: Is generating classifiers on-the-go the future of recognition?

Generic vs. Domain-specific: What is more promising? (Data, Architecture, Labels)