

Seminar: Multimodal Deep Learning

Topic 7: Text supporting CV models

Author: Maximilian Schneider

Supervisor: Jann Goschenhofer

Department Of Statistics
Ludwig-Maximilians-Universität



21.07.2022

Introduction: Scale

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. . . . Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. . . .

Introduction: Scale

... One thing that should be learned from the bitter lesson is the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great. The two methods that seem to scale arbitrarily in this way are search and learning. (Sutton 2019)

Contents

1 Introduction

2 Concepts

2.1 Web-scale data

2.2 Contrastive objective

2.3 Zero shooting and foundation models

2.4 Connecting image representations to language

3 Architectures

3.1 CLIP

3.2 ALIGN

3.3 Florence

4 Performance comparison

5 Resources

6 Sources

Concepts: Web-scale data

- Internet full of naturally occurring image-text pairs
- No labor intensive manual labeling
- Large datasets
 - 400 million (CLIP; Radford et al. 2021)
 - 900 million (Florence; Yuan et al. 2021)
 - 1.8 billion (ALIGN; Jia et al. 2021)
- Pre-processing needed, resulting in arbitrary choices
- Social biases are reproduced

Concepts: Contrastive objective

$$\ell_1^{V1,V2} = - \mathbb{E}_{\{v_1^1, v_2^1, \dots, v_2^N\}} \left(\log \frac{h_\theta(\{v_1^1, v_2^1\})}{h_\theta(\{v_1^1, v_2^1\}) + \sum_{k=2}^N h_\theta(\{v_1^1, v_2^k\})} \right) \quad (1)$$

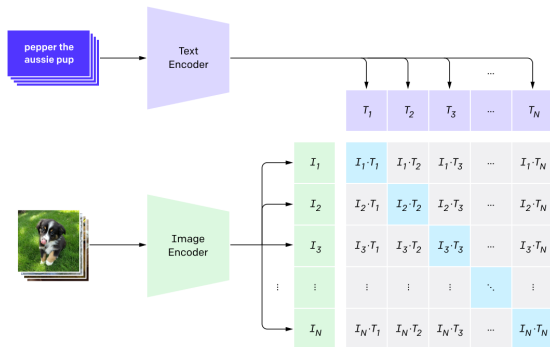


Figure 1: Visualization of contrastive objective (OpenAI 2021)

Concepts: Contrastive objective

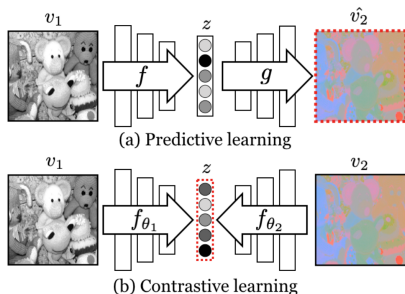


Figure 2: Predictive Learning vs Contrastive Learning: Contrastive Loss measured in representation space (Tian, Krishnan, and Isola 2020)

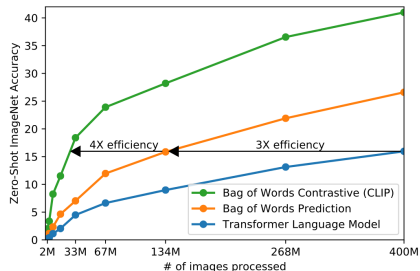


Figure 3: Data efficiency of contrastive objective (Radford et al. 2021)

Concepts: Zero shooting and foundation models

- Paradigms from NLP
- Zero shooting: Apply pre-trained model to new, unseen datasets; no deceivment through overfitting

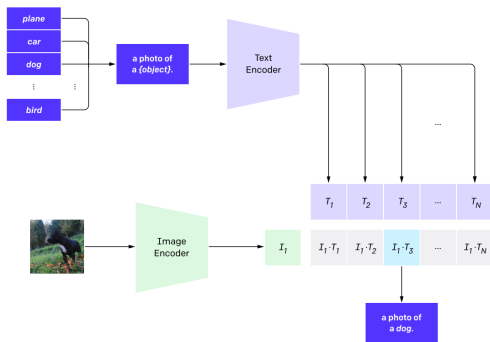


Figure 4: Visualization of zero-shot application (Radford et al. 2021)

- Foundation model: Reusing models, e.g., CLIP inside DALL·E 2 to embed images (Ramesh et al. 2022) or as a filter for creating LAION-400M (Schuhmann 2022)

Concepts: Connecting image representations to language

- Learned image representations are directly connected to natural language representations
- Direct specification of visual concepts possible through prompt engineering, e.g., “picture of ...”, “macro of ...” or “drawing of ...”

Contents

1 Introduction

2 Concepts

2.1 Web-scale data

2.2 Contrastive objective

2.3 Zero shooting and foundation models

2.4 Connecting image representations to language

3 Architectures

3.1 CLIP

3.2 ALIGN

3.3 Florence

4 Performance comparison

5 Resources

6 Sources

CLIP (Radford et al. 2021): Architecture

- Jointly trained image encoder and text encoder from scratch
- Image encoder: Some versions with modified ResNets, some versions with Vision Transformers (ViT; Dosovitskiy et al. 2020)
 - ResNet-50, ResNet-101, 3 which follow EfficientNet-style model scaling RN50x4, RN50x16, RN50x64
 - Vision Transformers: ViT-B/32, ViT-B/16, ViT-L/14, **ViT-L/14@336px**
- Text encoder: Transformer with modifications
- Maximization of cosine similarity of image embedding and text embedding

CLIP: Loss

Recap: general formulation of contrastive loss

$$\ell_1^{V1,V2} = - \mathbb{E}_{\{v_1^1, v_2^1, \dots, v_2^N\}} \left(\log \frac{h_\theta(\{v_1^1, v_2^1\})}{h_\theta(\{v_1^1, v_2^1\}) + \sum_{k=2}^N h_\theta(\{v_1^1, v_2^k\})} \right) \quad (2)$$

Application in CLIP:

$$\ell_i^{v_1 \rightarrow v_2} = - \log \frac{\exp(\langle v_1^i, v_2^i \rangle / \tau)}{\sum_{k=1}^N \exp(\langle v_1^i, v_2^k \rangle / \tau)}, \quad (3)$$

where $\langle v_1^i, v_2^i \rangle$ represents the cosine similarity, i.e., $v_1^{i\top} v_2^i / \|v_1^i\| \|v_2^i\|$; and $\tau \in \mathbb{R}^+$ represents a temperature parameter, which is directly learned during training (Zhang et al. 2020). This can be viewed as a symmetric cross entropy loss over the cosine similarity of the embeddings (Radford et al. 2021).

CLIP: Performance





































	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%

Figure 5: Robustness of zero-shot CLIP to distribution shifts (Radford et al. 2021)

ALIGN (Jia et al. 2021)

- Key difference to CLIP: Less (costly) data curation (400 million image-text pairs → 1.8 billion image-text pairs)
- Image encoder: EfficientNet (EfficientNet-L2)
- Text encoder: BERT (BERT-Large)
- Trained from scratch
- 800 million parameters (Alford 2021)
- Name: alignment of visual and language representations through contrastive loss or. Intuitively, "**A** Large-scale **I**ma**G**e and **N**oisy-text embedding"

ALIGN (Jia et al. 2021)

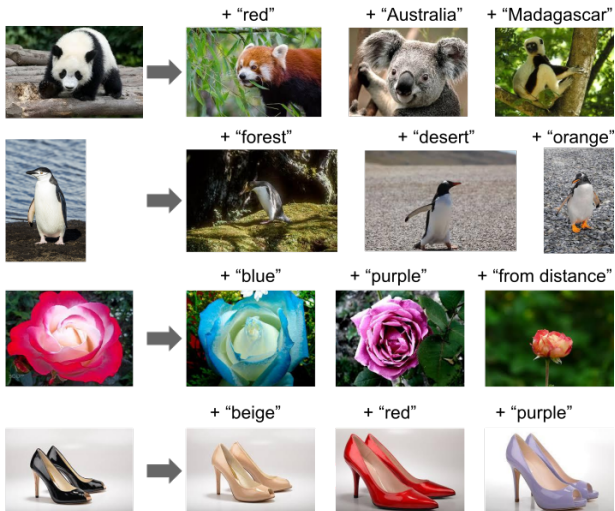


Figure 6: Addition of word and image embedding

Florence (Yuan et al. 2021)

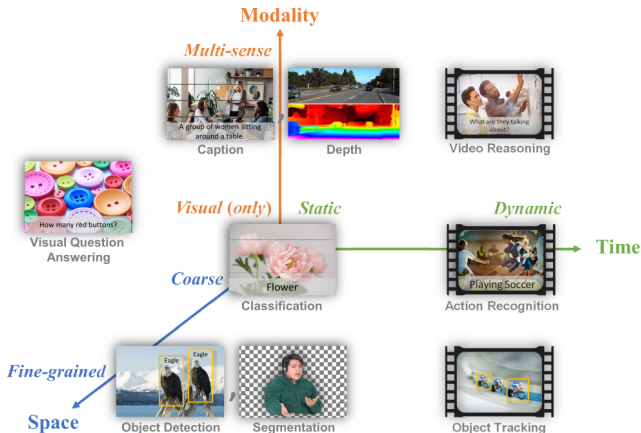


Figure 7: Florence' approach to foundation models: A general purpose vision system for all these tasks.

Florence (Yuan et al. 2021)

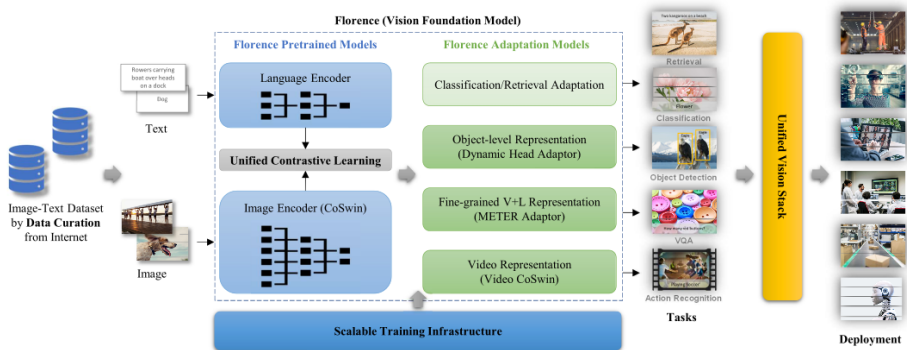


Figure 8: Modular architecture of florence

Florence (Yuan et al. 2021)

- Image encoder: hierarchical Vision Transformer (CoSwin Transformer)
- Text encoder: Transformer similar to CLIP
- Trained from scratch
- 893 million parameters (Alford 2021)
- 900 million image-text pairs
- Name: the origin of the trail for exploring vision foundation models, as well as the birthplace of Renaissance
- Loss: unified image-text contrastive learning in image-label-description space → All image-text pairs with the same label y are regarded as positive instances.

Performance comparison

	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
Florence	83.7			

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

Performance comparison

	Flickr30K (1K test set)				MSCOCO (5K test set)			
	Image→Text		Text→Image		Image→Text		Text→Image	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLIP	88.0	98.7	68.7	90.6	58.4	81.5	37.8	62.4
ALIGN	88.6	98.7	75.7	93.8	58.6	83.0	45.6	69.8
Florence	90.9	99.1	76.7	93.6	64.7	85.9	47.2	71.4

Table 2: Zero-shot image and text retrieval (Yuan et al. 2021)

Performance comparison

	Food101	CIFAR10	CIFAR100	SUN397	Stanford Cars	FGCV Aircraft	VOC2007	DTD	Oxford Pets	Caltech101	Flowers102	ImageNet
CLIP	93.8	95.7	77.5	68.4	78.8	37.2	84.3	55.7	93.5	92.8	78.3	76.2
Florence	95.1	94.6	77.6	77.0	93.2	55.5	85.5	66.4	95.9	94.7	86.2	83.7
CLIP (fine tuned)	95.9	97.9	87.4	82.2	91.5	71.6	89.9	83.0	95.1	96.0	99.2	85.4
ALIGN (fine tuned)	95.9				96.1				96.2			88.6
Florence (fine tuned)	96.2	97.6	87.1	84.2	95.7	83.9	90.5	86.0	96.4	96.6	99.7	90.1

Table 3: Top-1 Accuracy of CLIP, Florence and ALIGN on various datasets.

Resources

- Pre-trained CLIP models on Github:
<https://github.com/openai/CLIP>
- Command line image retrieval: rclip,
<https://github.com/yuriymikhalevich/rclip>
 - Wraps *Vit-B/32 CLIP*
- CLIP dataset not available → Open dataset LAION-400M (Schuhmann 2022)
- Florence code and model weights are not open source (Solawetz 2021)

Sources I

- Alford, A. (July 20, 2021). *Google Announces 800M Parameter Vision-Language AI Model ALIGN*. URL: <https://www.infoq.com/news/2021/07/google-vision-language-ai/>.
- Dosovitskiy, Alexey et al. (2020). “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *CoRR* abs/2010.11929. arXiv: 2010.11929. URL: <https://arxiv.org/abs/2010.11929>.
- Jia, Chao et al. (2021). “Scaling up visual and vision-language representation learning with noisy text supervision”. In: *International Conference on Machine Learning*. PMLR, pp. 4904–4916.
- OpenAI (Jan. 5, 2021). *CLIP: Connection Text and Images*. URL: <https://openai.com/blog/clip/>.
- Radford, Alec et al. (2021). “Learning transferable visual models from natural language supervision”. In: *International Conference on Machine Learning*. PMLR, pp. 8748–8763.

Sources II

- Ramesh, Aditya et al. (2022). “Hierarchical Text-Conditional Image Generation with CLIP Latents. 2022”. In: *arXiv preprint arXiv:2204.06125*.
- Schuhmann, C. (July 7, 2022). *Laion-400-Million Open Dataset*. URL: <https://laion.ai/blog/laion-400-open-dataset/>.
- Solawetz, J. (Dec. 9, 2021). *Florence: A New Foundation for Computer Vision*. URL: <https://blog.roboflow.com/florence-a-new-foundational-model-for-computer-vision/>.
- Sutton, R. S. (Mar. 13, 2019). *The Bitter Lesson*. URL: <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>.
- Tian, Yonglong, Dilip Krishnan, and Phillip Isola (2020). “Contrastive multiview coding”. In: *European conference on computer vision*. Springer, pp. 776–794.
- Yuan, Lu et al. (2021). “Florence: A New Foundation Model for Computer Vision”. In: *arXiv preprint arXiv:2111.11432*.

Sources III

Zhang, Yuhao et al. (2020). “Contrastive learning of medical visual representations from paired images and text”. In: *arXiv preprint arXiv:2010.00747*.