Lab 8: Connecting language and vision pre-report

CHI YEONG HEO[1],
[1]School of Computer Science and Engineering, Pusan National University, Busan 46241 Republic of Korea

1. Introduction

This report provides a summary of papers in the field of connecting language and vision: Learning Transferable Visual Models From Natural Language Supervision [2] and Linearly Mapping from Image to Text Space [1].

2. Learning Transferable Visual Models From Natural Language Supervision

2.1. Overview

Contrastive Language-Image Pre-training (CLIP) is a method for training computer vision models using natural language supervision. By leveraging raw text associated with images, CLIP facilitates: (1) the construction of training datasets from diverse sources, and (2) the development of models capable of generalizing effectively to a wide range of visual tasks without the need for task-specific fine-tuning. Proposed by OpenAI, CLIP represents a foundational advancement in the field of vision-language models.

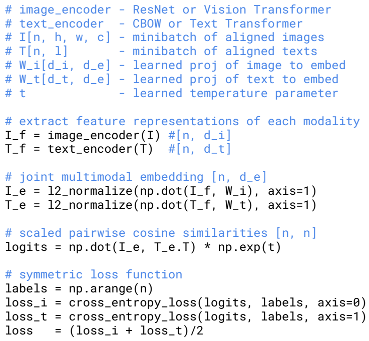
2.2. Natural Language Supervision

Learning from natural language supervision has several strengths: (1) It's much easier to scale since it doesn't require annotations from crowd-sourced labeling. Instead, it leverages the tremendous text on the Internet. (2) It doesn't "just" learn a representation but also connects that representation to language which enables flexible zero-shot transfer.

2.3. Contrastive Objective

Given a batch of N (image, text) pairs, CLIP is trained to identify which of the $N \times N$ possible (image, text) pairings within the batch correspond to the true matches. To achieve this, CLIP learns a multi-modal embedding space by jointly training an image encoder and a text encoder. The objective is to maximize the cosine similarity of the embeddings for the N true (image, text) pairings in the batch while minimizing the cosine similarity for the remaining $N^2 - N$ incorrect pairings. A symmetric cross-entropy loss over these similarities is employed during optimization.

This contrastive learning approach offers significant advantages over methods that predict exact words. The diversity of descriptions, comments, and related text accompanying images makes exact word prediction a highly challenging task. In contrast, training with a contrastive objective simplifies the problem by focusing on embedding alignment rather than precise text generation. The authors of the paper report that this method improves efficiency by a factor of four in zero-shot transfer performance on the ImageNet dataset.



```
# 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 2: Numpy-like pseudocode for the core of an implementation of CLIP.

2.4. Using CLIP for zero-shot Transfer

CLIP is pre-trained to determine whether an image and a text snippet are correctly paired. This capability is adapted for zero-shot classification by using the class names of a dataset as candidate text pairings. For a given image, CLIP computes embeddings for the image and the set of text descriptions using its respective encoders. The cosine similarity between the image and text embeddings is scaled by a temperature parameter τ and converted into a probability distribution using a softmax function. This process effectively implements a multinomial logistic regression classifier with normalized inputs, weights, and temperature scaling. In this framework, the image encoder serves as the vision backbone generating feature representations, while the text encoder acts as a hypernetwork that generates the classifier weights based on the textual descriptions of the visual concepts represented by the classes.

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

Table 1: Comparing CLIP to prior zero-shot transfer image classification results. CLIP improves performance on all three datasets by a large amount.

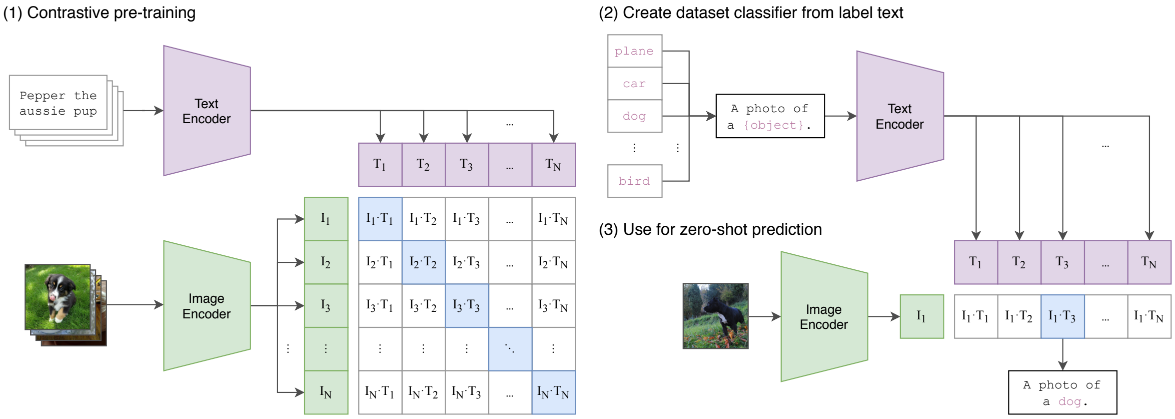

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Figure 1: Summary of CLIP. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

3. Linearly Mapping from Image to Text Space

3.1. Overview

This study examines the relationship between representations in language models and image encoders. It demonstrates three key findings: (1) visual semantic information can be effectively mapped to language models as soft prompts via a linear transformation, without modifying model parameters; (2) this mapping enables generative models to describe and interpret images at a performance level comparable to multimodal models that jointly train vision and language components; (3) linguistic supervision during pretraining significantly enhances concept formation in models, improving the transferability of visual features between vision and language spaces.

3.2. Method: Linearly Mapping from Image to Text Representations

The authors propose a simplified framework, termed `LiMBeR` (Linearly Mapping Between Representation Spaces), to bridge image and text representations. This approach uses a single linear projection layer P to map image features from a pretrained encoder E into the input space of a generative language model LM, without altering the parameters of either E or LM. The projected features, referred to as soft prompts, enable the LM to interpret and describe visual inputs.

The method builds on prior multimodal systems such as MAGMA and Frozen but significantly reduces complexity by relying solely on the linear mapping layer for modality alignment. This approach allows the authors to investigate the structural similarities between vision and language representations.

From E an image encoding of dimensionality h_I is extracted. Then that encoding is projected to a $e_L * k$ sequence of soft prompts, an image prompts.

The language model, GPT-J, is paired with various image encoders:

- **CLIP RN50x16**: Trained on multimodal image-text embeddings, providing strong linguistic supervision.
- **NF-ResNet50**: Pretrained on WordNet-labeled data for indirect linguistic supervision, with variations including random initialization and fine-tuning.
- **BEIT-Large**: Self-supervised with no linguistic supervision, including pretrained and fine-tuned variants.

3.3. Performance on Vision-Language Tasks

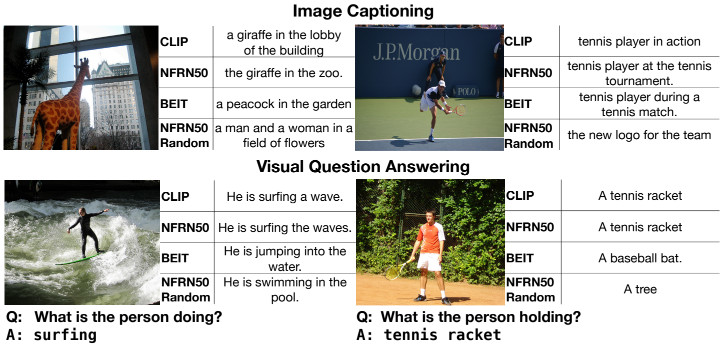


Image Captioning

CLIP	a giraffe in the lobby of the building	CLIP	tennis player in action
NFRN50	the giraffe in the zoo.	NFRN50	tennis player at the tennis tournament.
BEIT	a peacock in the garden	BEIT	tennis player during a tennis match.
NFRN50 Random	a man and a woman in a field of flowers	NFRN50 Random	the new logo for the team

Visual Question Answering

CLIP	He is surfing a wave.	CLIP	A tennis racket
NFRN50	He is surfing the waves.	NFRN50	A tennis racket
BEIT	He is jumping into the water.	BEIT	A baseball bat.
NFRN50 Random	He is swimming in the pool.	NFRN50 Random	A tree

Q: What is the person doing?
A: `surfing`

Q: What is the person holding?
A: `tennis racket`

Figure 3: Selected examples of captioning and zero-shot visual question answering (VQA), demonstrating each model's capability to transfer information to the language model without fine-tuning either component.

Table 2 summarizes the primary results on vision-language tasks. The findings indicate that fine-tuning parameters within either the encoder or decoder provides limited performance gains. In many cases, jointly tuned models underperform compared to those where only the projection layer was trained with frozen encoder and decoder models.

Moreover, the results suggest a significant correlation between the linguistic supervision employed during pretraining and the efficacy of transferring information to the language model. This highlights the importance of pretraining tasks in establishing alignment between vision and language modalities.

Image Captioning	NoCaps - CIDEr-D				NoCaps (All)		**CoCo**	CoCo	
	In	Out	Near	All	CLIP-S	Ref-S	CIDEr-D	CLIP-S	Ref-S
🔥NFRN50 Tuned	20.9	30.8	25.3	27.3	66.5	72.5	35.3	69.7	74.8
🔥MAGMA (released)	18.0	12.7	18.4	16.9	63.2	68.8	**52.1**	76.7	79.4
🔥MAGMA (ours)	**30.4**	**43.4**	**36.7**	**38.7**	74.3	78.7	47.5	75.3	**79.6**
❄BEIT Random	5.5	3.6	4.1	4.4	46.8	55.1	5.2	48.8	56.2
❄NFRN50 Random	5.4	4.0	4.9	5.0	47.5	55.7	4.8	49.5	57.1
❄BEIT	20.3	16.3	18.9	18.9	62.0	69.1	22.3	63.6	70.0
❄NFRN50	21.3	31.2	26.9	28.5	65.6	71.8	36.2	68.9	74.1
❄BEIT FT.	**38.5**	**48.8**	**43.1**	**45.3**	73.0	78.1	51.0	74.2	78.9
❄CLIP	34.3	48.4	41.6	43.9	**74.7**	**79.4**	**54.9**	**76.2**	**80.4**

VQA n-shots	0	1	**2**	4
Blind	20.60	35.11	36.17	36.99
🔥NFRN50 Tuned	27.15	37.47	38.48	39.18
🔥MAGMA (ours)	24.62	39.27	40.58	41.51
🔥MAGMA (reported)	32.7	40.2	**42.5**	43.8
❄NFRN50 Random	25.34	36.15	36.79	37.43
❄BEIT	24.92	34.35	34.70	31.72
❄NFRN50	27.63	37.51	38.58	39.17
❄CLIP	33.33	39.93	**40.82**	40.34

Table 2: Captioning Performance and Visual Question Answering (VQA) accuracy for all variations on model architecture and image encoders used. There's a consistent increasing trend in performance that correlates with an increase in linguistic supervision.

3.4. Transfer of Visual Concepts

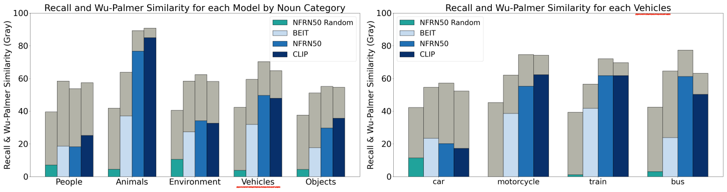


Figure 4: Average noun recall in generated captions follows the expected pattern (CLIP>NFRN50>BEIT). However, based on Wu-Palmer similarity, BEIT achieves comparable or superior performance to NFRN50 and CLIP across 4 out of 5 noun categories. This indicates that BEIT, while struggling to transfer exact concepts, effectively transfers related concepts based on visual similarity. The right-hand side illustrates this phenomenon with vehicle-related terms, where BEIT, despite not recognizing 'bus,' enables the language model to infer a semantically related concept such as another type of vehicle. Average random Wu-Palmer similarity remains approximately 0.4 across datasets.

The results, presented in Figure 4, demonstrate that BEIT exhibits lower noun recall in categories such as 'people,' 'environment,' 'vehicles,' and 'objects' compared to NFRN50 and CLIP. However, it achieves comparable performance in terms of Wu-Palmer similarity across many categories. Unlike the pretraining paradigms of NFRN50 and CLIP, BEIT's pretraining does not explicitly enforce the learning of fine-grained conceptual distinctions between visually similar objects described by different lexical terms.

Despite these limitations, prompts from BEIT enable the language model to infer broader conceptual meanings. For instance, BEIT often transfers coarse-grained visual concepts, as evidenced by high Wu-Palmer similarity scores and corroborated by CLIPScore results in Table 2. This suggests that while BEIT may struggle with lexical precision, it successfully bridges visual and textual domains at a broader conceptual level.

[1] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.