PRE-REPORT: Learning Transferable Visual Models From Natural Language Supervision

SongSeungWu

1 ABSTRACT

This paper introduces a groundbreaking method for computer vision by utilizing natural language supervision to learn transferable image representations. The approach involves pre-training on a dataset of 400 million (image, text) pairs to align images with corresponding captions. This enables the model to perform zero-shot transfer across various downstream tasks without additional training on specific datasets. Demonstrating competitive performance on over 30 benchmarks, the model matches ResNet-50's accuracy on ImageNet in a zero-shot setting, highlighting its potential as a scalable alternative to traditional supervised learning.

2 INTRODUCTION AND MOTIVATING WORK

The introduction emphasizes the transformative success of task-agnostic objectives in NLP, such as autoregressive and masked language modeling, which enable zero-shot transfer, as exemplified by GPT-3. In contrast, computer vision continues to rely on supervised datasets like ImageNet. Previous research into natural language supervision for image representation learning faced challenges in scalability and performance. To address these limitations, the authors propose CLIP, a model trained on 400 million (image, text) pairs. CLIP closes the gap between small-scale experiments and large-scale implementations, achieving robust zero-shot performance across 30+ datasets while maintaining computational efficiency. This advancement has significant implications for generalizing computer vision models and addressing policy considerations.

3 APPROACH

The paper outlines the methodology for learning image representations using natural language supervision, focusing on three key components: leveraging language as a training signal, constructing a large dataset, and selecting efficient pre-training and model scaling methods.

Natural Language Supervision

- Natural language serves as a scalable and flexible supervision source. Unlike traditional labeled datasets, it enables passive learning from internet text and supports zero-shot transfer by linking representations to language. The authors adopt a contrastive learning approach, addressing inefficiencies in previous predictive models by training on a proxy task: predicting matching (image, text) pairs.

Creating a Sufficiently Large Dataset

- Existing datasets like MS-COCO and Visual Genome are too small, while YFCC100M is sparse in quality metadata. To overcome these limitations, the authors construct a new dataset of 400 million (image, text) pairs from diverse internet sources, ensuring coverage of a broad set of visual concepts. The resulting dataset, WIT, enables large-scale natural language supervision.

Selecting an Efficient Pre-Training Method

- The authors use a contrastive objective for efficiency, training models to maximize cosine similarity between embeddings of matching pairs while minimizing mismatched pairs. This approach significantly reduces compute requirements compared to generative or predictive methods.

Choosing and Scaling a Model

- The image encoder employs modified ResNet-50 or Vision Transformer architectures, while the text encoder uses a Transformer-based architecture. The models are scaled across width, depth, and resolution to maximize performance. This design balances computational efficiency and model capacity, enabling large-scale pre-training and competitive performance in zero-shot transfer.

4 TRAINING

The training process involves five ResNet variants and three Vision Transformer models. ResNet models include ResNet-50, ResNet-101, and scaled-up versions using EfficientNet-style scaling. ViT models include ViT-B/32, ViT-B/16, and ViT-L/14. All models were trained for 32 epochs using the Adam optimizer with weight decay and a cosine learning rate schedule. Key techniques for optimizing training included mixed precision, gradient checkpointing, and sharded similarity calculations across GPUs to manage memory and efficiency. The largest ResNet, RN50x64, trained in 18 days on 592 GPUs, while the largest ViT model trained in 12 days on 256 GPUs. A higher resolution pre-training for ViT-L/14 improved performance, making it the best-performing CLIP model.

5 EXPERIMENTS

Zero-Shot Transfer

Motivation

- Zero-shot transfer evaluates a model's ability to generalize to unseen datasets and tasks, measuring task-learning capabilities. CLIP leverages natural language supervision to achieve robust generalization.

Using CLIP for Zero-Shot Transfer

- CLIP classifies by embedding text descriptions of classes and matching them with image embeddings using cosine similarity, enabling dynamic, task-specific classifiers.

Prompt Engineering and Ensembling

- Customizing prompts and ensembling multiple context templates improve accuracy, with ImageNet performance increasing by 5% using 80 prompts.

Analysis of Performance

- CLIP excels in diverse tasks and shows strong robustness to distribution shifts, reducing errors by up to 75%. However, it struggles with specialized domains like medical diagnostics and satellite imagery.

6 LIMITATIONS

Performance Gaps

- CLIP's zero-shot performance is competitive but falls short on fine-grained and abstract tasks, performing poorly on tasks outside its training data.

Generalization Issues

- Struggles with out-of-distribution data, as shown by poor MNIST results, where simpler models outperform CLIP.

Few-Shot Learning

- Performance drops when transitioning from zero-shot to few-shot learning, highlighting inefficiencies in adapting with minimal data.

7 CONCLUSION

This study explores the application of task-agnostic web-scale pre-training from NLP to computer vision. The results indicate that similar behaviors emerge, with CLIP models learning a wide range of tasks during pre-training. These capabilities enable zero-shot transfer to various datasets using natural language prompting.

PRE-REPORT: LINEARLY MAPPING FROM IMAGE TO TEXT SPACE

SongSeungWu

1 ABSTRACT

This study investigates how well text-only language models can learn features of the non-linguistic world. It tests the hypothesis that conceptual representations of frozen text-only and vision-only models can be linked via a linear transformation. By transferring image representations from vision models to frozen LMs through a single linear projection, the study achieves competitive performance in captioning and visual question answering tasks. Comparing three image encoders, the results show that CLIP, pretrained with linguistic supervision, encodes category information more effectively and outperforms others on vision-language benchmarks. These findings suggest that LMs encode conceptual information structurally similarly to vision-based models.

2 INTRODUCTION

This study tests the hypothesis that conceptual representations between text-only language models and image encoders can be connected through a linear transformation. The authors propose LiMBeR, which projects image representations into the embedding space of LMs as soft prompts. By using three image encoders, the study demonstrates that higher levels of linguistic supervision during pretraining lead to more effective information transfer to LMs.

The key findings

- Visual semantic information can be successfully mapped to LMs as soft prompts.
- This mapping enables generative models to describe images and answer questions at performance levels comparable to multimodal models.
- Linguistic supervision plays a critical role in concept formation and the transferability of visual features between vision and text spaces.

3 METHOD: LINEARLY MAPPING FROM IMAGE TO TEXT REPRESENTATIONS

Linearly Mapping Between Representation spaces is a method for linearly mapping the representation spaces of image encoders and language models. A single linear layer P is trained to transform hidden representations hi from an image encoder into the input space eL of a language model. These transformations, referred to as soft prompts, allow an analysis of conceptual similarities between the two models.

TRAINING PROCEDURE

- All models were trained on an image captioning task using the Conceptual Captions 3M dataset for 15,000 steps with consistent hyperparameters. The study used three image encoders with varying levels of linguistic supervision, to evaluate their performance based on linguistic alignment.

LIMITATIONS

- The main limitations of this study include a lack of control over the prompt length k, which may influence model performance. High values of k were used in experiments, but their exact impact was not systematically tested. Additionally, the runoff issue in the language model could cause it to generate plausible but unrelated outputs, creating the illusion of recognizing unseen image elements. Nonetheless, the overall performance on describing images and answering questions remained robust across datasets.

4 PERFORMANCE ON VISION-LANGUAGE TASKS

This section verifies that image representations linearly projected into the LM's input space can carry semantic information that the LM can process. Experiments were conducted on various datasets, and performance was evaluated using metrics such as CIDEr-D, CLIPScore, and RefCLIPScore. Results indicate that CLIP consistently outperformed NFRN50 and BEIT.

DATA

- Experiments utilized MSCOCO, NoCaps, and VQA2 datasets. Prompts like A picture of were appended to enhance the language model's performance.

METRICS

- For image captioning, CIDEr-D, CLIPScore, and RefCLIPScore were used. CIDEr-D rewards visually informative words, while CLIPScore assesses semantic similarity between images and captions. Visual question answering was evaluated based on accuracy.

RESULTS

- CLIP outperformed both NFRN50 and BEIT across tasks.
- BEIT, despite lacking linguistic supervision in pretraining, showed significantly better performance than randomly initialized NFRN50.
- BEIT struggled with transferring lexical category details, leading to poor performance in VQA.
- Linguistic supervision during pretraining proved critical for effective concept transfer, as demonstrated by BEIT's improved performance with additional finetuning, occasionally surpassing CLIP.

5 TRANSFER OF VISUAL CONCEPTS

Analyzing when image prompts succeed or fail helps understand the differences between text and image representation spaces. BEIT performed decently in captioning tasks but poorly in VQA, likely due to its inability to encode fine-grained lexical categories, despite its capability to encode visual properties effectively.

TRANSFER OF LEXICAL CATEGORICAL CONCEPTS

- Method: Using the COCO dataset, the top 50 nouns, adjectives, and relations from captions were analyzed. Precision, recall, and Wu-Palmer similarity were used to evaluate generated captions against ground truth.
- Results: BEIT had lower recall for nouns like people, vehicles, and objects compared to CLIP and NFRN50, but maintained similar Wup similarity. This suggests BEIT transfers coarse-grained visual information but struggles with distinguishing fine lexical categories.

6 CONCLUSION

This paper uses LiMBeR to analyze the similarity between image and text representation spaces. CLIP outperformed ResNet and BEIT, with linguistic supervision playing a key role in transfer success. LMs and vision models show similarity in coarse concepts, but fine-grained lexical concepts transfer only with linguistic supervision. The study suggests that linear transformations effectively transfer image information.