Learning Transferable Visual Models From Natural Language Supervision

Thành viên nhóm:

20521609 – Nguyễn Hoàng Minh

20521998 – Nguyễn Thiện Thuật

Link paper:

https://arxiv.org/pdf/2103.00020.pdf

Nội dung

01

Giới thiệu

02

(CLIP) Contrastive Language-Image Pre-Training

03

Kết quả thực nghiệm

04

Minh họa mô hình CLIP

05

Q&A

1. Giới thiệu

Động lực nghiên cứu:

- Những **nghiên cứu trước** đã chứng minh việc **giám sát từ ngôn ngữ tự nhiên đã cung cấp phương pháp hiệu quả để biểu diễn hình ảnh.**
- Tuy nhiên, hiệu suất trên các **benchmark vẫn thấp** và **bị giới hạn dữ liệu.**

1. Giới thiệu

Đóng góp của nghiên cứu:

- **Đề xuất và chứng minh** mô hình có tên là **CLIP** (mô hình đơn giản hơn của ConVIRT) có **hiệu suất tốt hơn** từ việc giám sát ngôn ngữ tự nhiên.
    - **Hiệu suất của mô hình zero-shot CLIP** trên **30 tập dữ liệu** có thể **cạnh tranh với những mô hình được train với nhiệm vụ cụ thể** của dataset đó.
    - **CLIP vượt trội các mô hình trên ImageNet** và có **hiệu quả tính toán** cao hơn.
    - **CLIP phù hợp** với **ResNet50** trên **ImageNet "zero-shot"** mà **không sử dụng bất kỳ tập huấn luyện nào được gắn nhãn ban đầu.**
- **CLIP có thể biểu diễn hình ảnh với ngôn ngữ** trên tập **dữ liệu quy mô lớn.**
- **CLIP (giống GPT) có thể thực hiện nhiều task thông qua pre-training** của nó:

OCR, geo-localization, action recognition,...

2. (CLIP) Contrastive Language-Image Pre-Training

Selecting an Efficient Pre-Training Method

- Các hệ thống **CV hiện đại** sử dụng **khối lượng tính toán rất lớn** và chỉ dự đoán 1000 class của ImageNet
  - 19 năm GPU để huấn luyện ResNeXt101-32x48d
  - 33 năm TPUv3 để huấn luyện Noisy Student EfficientNet-L2

⇒ **Pre-train là giải pháp tốt nhất để giải quyết vấn đề**

- Cách tiếp cận ban đầu, tương tự như VirTex, đã cùng train CNN images và text transformer từ đầu để dự đoán titles của ảnh.

⇒ **Rất khó để mở rộng mô hình.**

2. (CLIP) Contrastive Language-Image Pre-Training

Selecting an Efficient Pre-Training Method

- Cả hai cách tiếp cận này có chung một điểm chung là cố gắng dự đoán chính xác các từ trong văn bản với mỗi hình ảnh ***(predictive objective)***. (Rất khó vì rất nhiều mô tả, comment và văn bản đi kèm với hình ảnh).

⇒ Nghiên cứu gần đây đã phát hiện ra rằng các mục **contrastive objective** có thể **học cách biểu diễn tốt hơn** so với **predictive objective**

⇒ **CLIP sử dụng contrastive objective.**

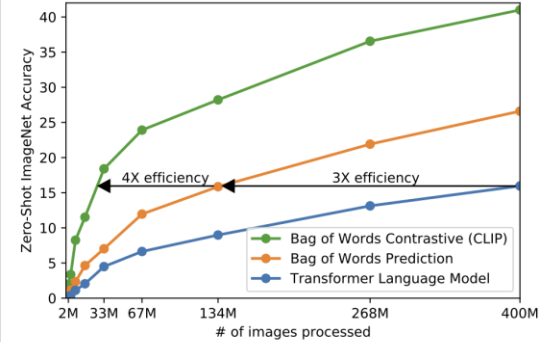


Figure 2. **CLIP is much more efficient at zero-shot transfer than our image caption baseline.** Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a bag-of-words (BoW) encoding of the text (Joulin et al., 2016). Swapping the prediction objective for the contrastive objective of CLIP further improves efficiency another 4x.

2. (CLIP) Contrastive Language-Image Pre-Training

Creating a Sufficiently Large Datasets

- **Tập dữ liệu WIT** (WebImageText) (Số từ tương đương WebText được sử dụng để huấn luyện GPT-2):
  - MS-COCO
  - Visual Genome
  - YFCC100M
  - 400 triệu cặp (image, text) được thu tập từ đa dạng nguồn trên intenet.
- Với **MS-COCO** and **Visual Genome** là **high quality crowd-labeled dataset**, nhưng dữ liệu **bị nhỏ so với tiêu chuẩn hiện tại** (khoảng 100,000 ảnh) trong khi các hệ thống CV thường train trên dữ liệu khoảng 3.5 tỷ ảnh.

2. (CLIP) Contrastive Language-Image Pre-Training

Creating a Sufficiently Large Datasets

- Với **YFCC100M**, khoảng **100 triệu ảnh**, nhưng ảnh mỗi **nhãn thưa thớt và có chất lượng khác nhau**. **Sau khi lọc các ảnh** chỉ có titles hoặc descriptions bằng Tiếng Anh, **dataset còn lại 15 triệu ảnh** *(gần bằng kích thước ImageNet)*.

- **Dữ liệu vẫn chưa đủ và chúng sẽ đánh giá thấp tiềm năng của hướng nghiên cứu này**. Họ **thu thập trên internet thêm** nhằm bao quát các trường hợp càng rộng càng tốt. Với **500.000 queries (20.000 cặp/queries)**, **điều kiện là các từ Tiếng Anh** xuất hiện **ít nhất 100 lần của Wikipedia**.

2. (CLIP) Contrastive Language-Image Pre-Training

CLIP Overview

- CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on a variety of (image, text) pairs.

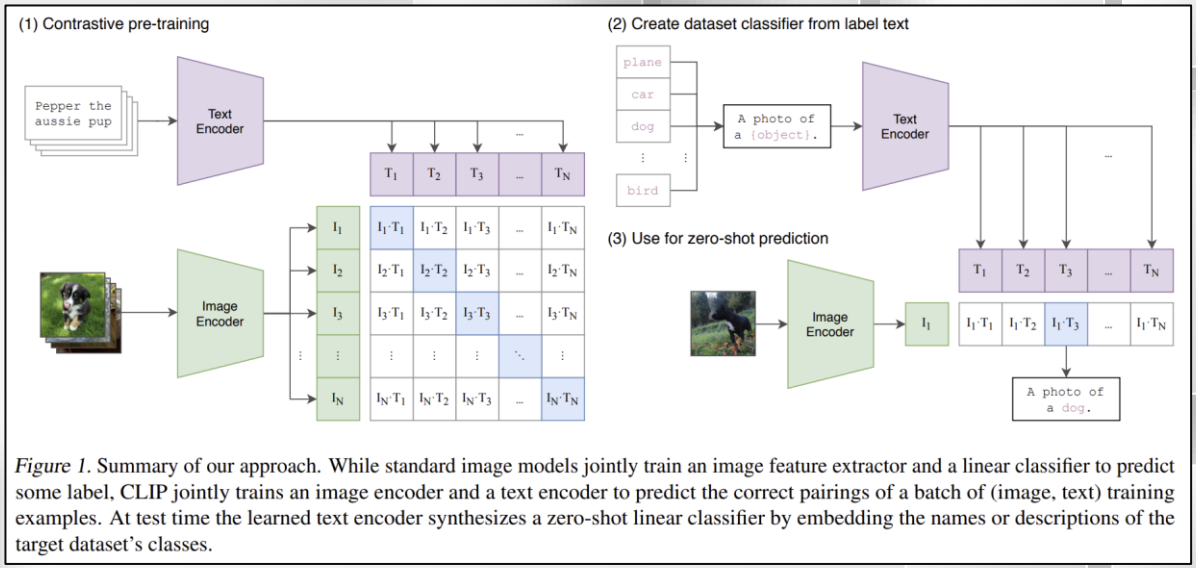


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the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

- It can be instructed in natural language to predict the most relevant text snippet, given an image without directly optimizing for the task, similarly to the zero-shot capabilities of GPT-2 and 3.

2. (CLIP) Contrastive Language-Image Pre-Training

Background

The loss function is an image-to-text:

$$\ell_i^{(v\to u)} = -\log \frac{\exp(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau)},$$

The loss function is an text-to-image:

$$\ell_i^{(u\to v)} = -\log \frac{\exp(\langle \mathbf{u}_i, \mathbf{v}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{u}_i, \mathbf{v}_k \rangle / \tau)}.$$

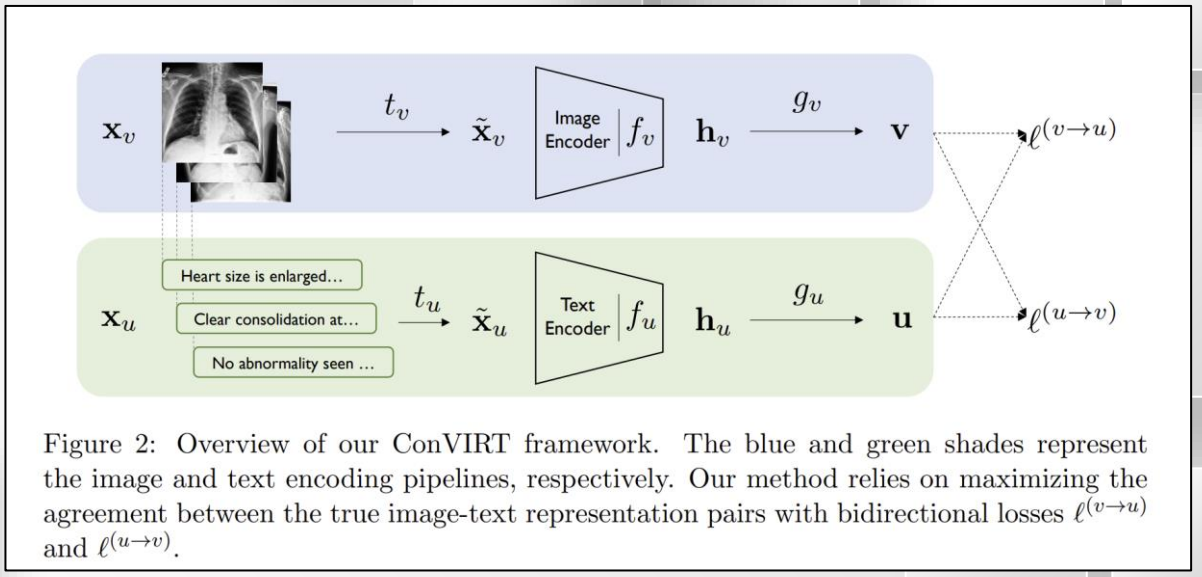


Figure 2: Overview of our ConVIRT framework. The blue and green shades represent the image and text encoding pipelines, respectively. Our method relies on maximizing the agreement between the true image-text representation pairs with bidirectional losses $\ell^{(v\to u)}$ and $\ell^{(u\to v)}$.

- <vi, ui> biểu diễn cosine similarity
- N input pairs.
- Temperature parameter (τ ∈ R+) là tham số học, để kiểm soát phạm vi log trong softmax.

2. (CLIP) Contrastive Language-Image Pre-Training

Background

Final training loss over all positive image-text pairs in each minibatch:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda \ell_i^{(v\to u)} + (1-\lambda)\ell_i^{(u\to v)}\right),$$

- $\lambda \in [0, 1]$ là siêu tham số.

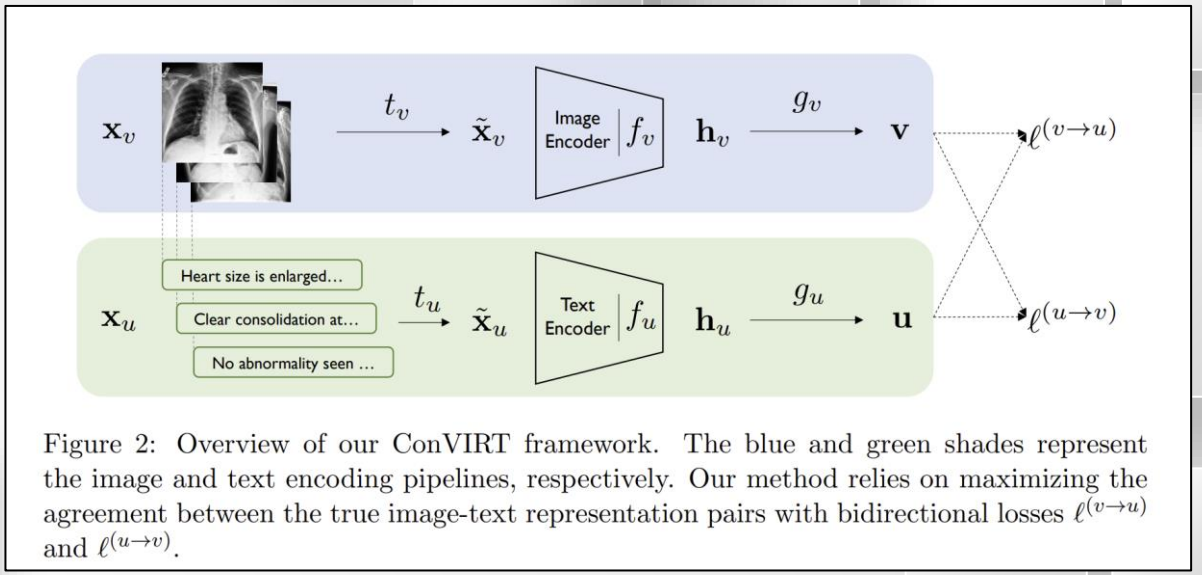


Figure 2: Overview of our ConVIRT framework. The blue and green shades represent the image and text encoding pipelines, respectively. Our method relies on maximizing the agreement between the true image-text representation pairs with bidirectional losses $\ell^{(v\to u)}$ and $\ell^{(u\to v)}$.

2. (CLIP) Contrastive Language-Image Pre-Training

CLIP model

- Cho **N** cặp (image, text), **CLIP được huấn luyện** để **dự đoán các cặp đúng** trong số N × N cặp (image, text)
- CLIP học **multi-modal embedding** bằng cách **cùng huấn luyện image encoder và text encoder** để **tối đa cosine similarity** của **N cặp đúng**.
- Tối ưu hóa bằng **a symmetric cross entropy loss (InfoNCE loss)** từ **similarity scores**.

2. (CLIP) Contrastive Language-Image Pre-Training

CLIP model

- Lấy ý tưởng từ **ConVIRT** nhưng CLIP được thiết kế **đơn giản hơn**. Vì **kích thước dữ liệu lớn** nên **không** lo ngại về **overfiting:**
  - Huấn luyện CLIP từ đầu mà **không cần weights của image encoder** trên ImageNet và **weights pre-trained** của text encoder
  - **Không sử dụng the non-linear projection g_v** giữa **encoder's representation** và **contrastive embedding space**.
  - **Chỉ** sử dụng **linear projection** để ánh xạ từ **encoder's representation** sang **contrastive embedding space**.

2. (CLIP) Contrastive Language-Image Pre-Training

CLIP model

- Lấy ý tưởng từ **ConVIRT** nhưng CLIP được thiết kế đơn giản hơn. Vì **kích thước dữ liệu lớn** nên **không** lo ngại về **overfiting:**
  - **Loại bỏ text transformation t_u** dùng để lấy mẫu một câu thống nhất từ văn bản cho mỗi hình *(vì trong tập dữ liệu huấn luyện của CLIP chỉ là một câu duy nhất).*
  - **Đơn giản** hóa **image transformation t_v** bằng cách **chỉ** sử dụng **random square crop** cho **ảnh** để **tăng cường dữ liệu**

2. (CLIP) Contrastive Language-Image Pre-Training

The image encoder

- **ResNet-50**, có biến đổi
  - ResNetD
  - Antialiased rect-2 blur pooling
  - Thay thế global average pooling layer thành attention pooling
- **Vision Transformer (ViT)** chỉ sửa đổi nhỏ là thêm **layer normalization** vào **trước combined patch** và **position embeddings** của transformer.

2. (CLIP) Contrastive Language-Image Pre-Training

The text encoder

- **Transformer**
  - 63M-parameter
  - 12- layer 512-wide model
  - 8 attention heads
  - 63M-parameter
  - Max sequence length là 76
  - The text sequence is bracketed with [SOS] and [EOS] tokens
  - Sử dụng [EOS] token như là text feature được normalized và sử dụng để đưa vào the multi-modal embedding space

2. (CLIP) Contrastive Language-Image Pre-Training

Training

- **5 ResNet:** ResNet-50, ResNet-101, RN50x4, RN50x16, RN50x64 *(3 EfficientNet-style models of ResNet-50).*
- **3 Vision Transformers**: ViT-B/32, ViT-B/16, ViT-L/14.
- Adam optimizer with decoupled weight decay regularization với 32 epochs.
- Initial hyperparameters: grid searches, random search, and manual tuning.
- The learnable temperature parameter: 0.07
- Minibatch size: 32,768.
- Mixed-precision được sử dụng để tăng tốc độ học.
- Gradient checkpointing, half-precision Adam, and half-precision stochastically rounded text encoder weights để tiết kiệm bộ nhớ.
- The RN50x64 (largest ResNet model) mất 18 ngày trên 592 V100 GPUs.
- Transformer (largest Vision model) mất 12 ngày trên 256 V100 GPUs.

3. Kết quả thực nghiệm

Initial comparison to visual n-grams

	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. This improvement reflects many differences in the 4 years since the development of Visual N-Grams (Li et al., 2017).

3. Kết quả thực nghiệm

Analysis of zero-shot clip performance

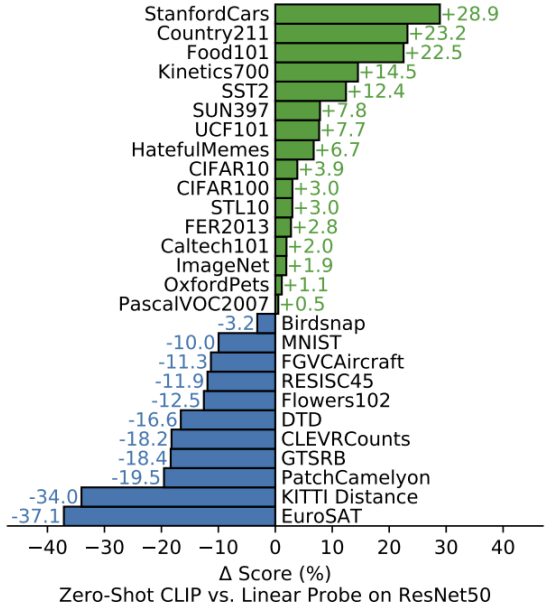


Figure 5. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet-50 features on 16 datasets, including ImageNet.

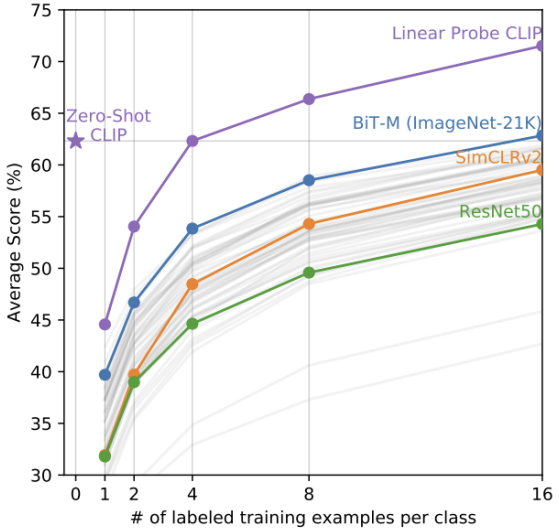


Figure 6. **Zero-shot CLIP outperforms few-shot linear probes.** Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, 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.

3. Kết quả thực nghiệm

Representation Learning

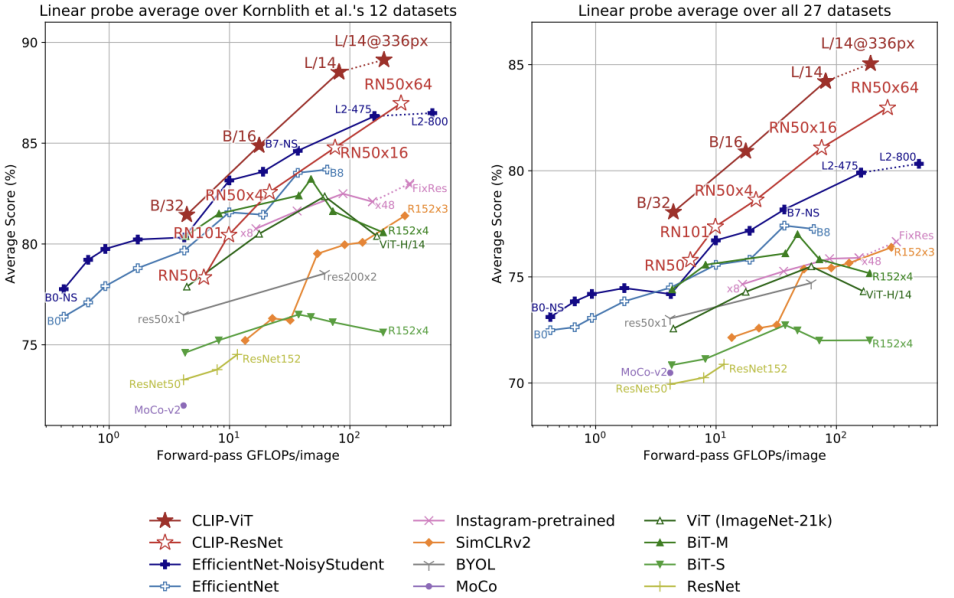


Figure 10. **Linear probe performance of CLIP models in comparison with state-of-the-art computer vision models**, including EfficientNet (Tan & Le, 2019; Xie et al., 2020), MoCo (Chen et al., 2020d), Instagram-pretrained ResNeXt models (Mahajan et al., 2018; Touvron et al., 2019), BiT (Kolesnikov et al., 2019), ViT (Dosovitskiy et al., 2020), SimCLRv2 (Chen et al., 2020c), BYOL (Grill et al., 2020), and the original ResNet models (He et al., 2016b). (Left) Scores are averaged over 12 datasets studied by Kornblith et al. (2019). (Right) Scores are averaged over 27 datasets that contain a wider variety of distributions. Dotted lines indicate models fine-tuned or evaluated on images at a higher-resolution than pre-training. See Table 10 for individual scores and Figure 20 for plots for each dataset.

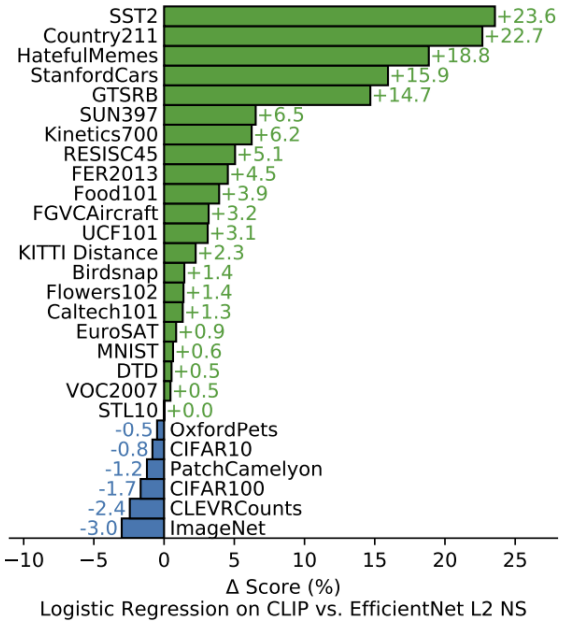


Figure 11. **CLIP's features outperform the features of the best ImageNet model on a wide variety of datasets.** Fitting a linear classifier on CLIP's features outperforms using the Noisy Student EfficientNet-L2 on 21 out of 27 datasets.

3. Kết quả thực nghiệm

Representation Learning

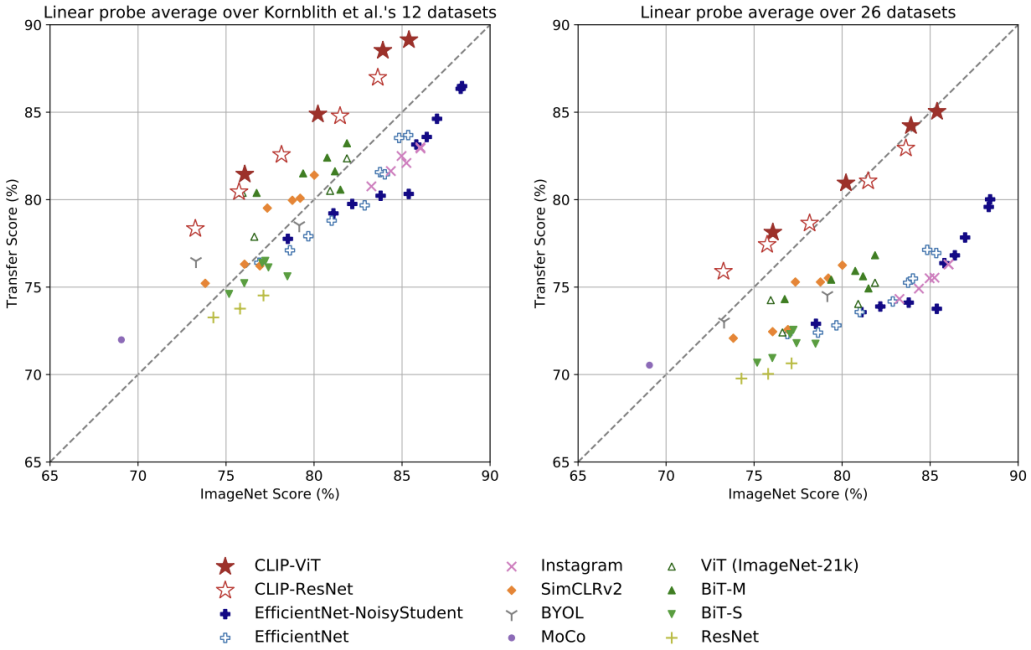
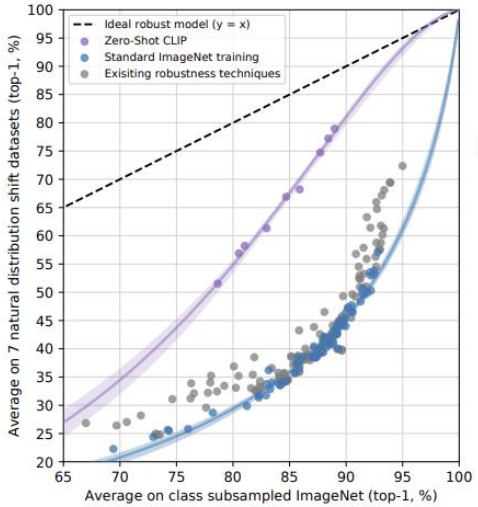


Figure 12. **CLIP's features are more robust to task shift when compared to models pre-trained on ImageNet.** For both dataset splits, the transfer scores of linear probes trained on the representations of CLIP models are higher than other models with similar ImageNet performance. This suggests that the representations of models trained on ImageNet are somewhat overfit to their task.

3. Kết quả thực nghiệm

Representation Learning



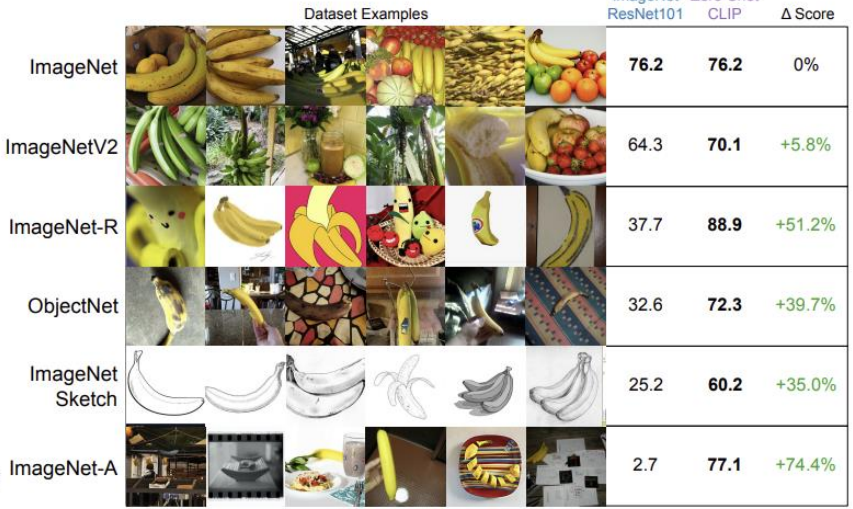


Figure 13. **Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models.** (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this "robustness gap" by up to 75%. Linear fits on logit transformed values are shown with bootstrap estimated 95% confidence intervals. (Right) Visualizing distribution shift for bananas, a class shared across 5 of the 7 natural distribution shift datasets. The performance of the best zero-shot CLIP model, ViT-L/14@336px, is compared with a model that has the same performance on the ImageNet validation set, ResNet-101.

4. Minh họa mô hình CLIP

Q&A

Cảm Ơn!