

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford^{*1} Jong Wook Kim^{*1} Chris Hallacy¹ Aditya Ramesh¹ Gabriel Goh¹ Sandhini Agarwal¹
Girish Sastry¹ Amanda Askell¹ Pamela Mishkin¹ Jack Clark¹ Gretchen Krueger¹ Ilya Sutskever¹

Abstract

SOTA computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer of the model to downstream tasks. We study performance on over 30 different computer vision datasets, spanning tasks such as OCR, action recognition in videos, geo-localization, and many types of fine-grained object classification. The model transfers non-trivially to most tasks and is often competitive with a fully supervised baseline without the need for any dataset specific training. For instance, we match the accuracy of the original ResNet50 on ImageNet zero-shot without needing to use any of the 1.28 million training examples it was trained on. We release our code and pre-trained model weights at <https://github.com/OpenAI/CLIP>.

1. Introduction and Motivating Work

Pre-training methods which learn directly from raw text have revolutionized NLP over the last few years (Dai & Le, 2015; Peters et al., 2018; Howard & Ruder, 2018; Radford et al., 2018; Devlin et al., 2018; Raffel et al., 2019). The development of “text-to-text” as a standardized input-output

^{*}Equal contribution ¹OpenAI, San Francisco, CA 94110, USA.
Correspondence to: <{alec, jongwook}@openai.com>.

interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on crowd-labeled datasets such as ImageNet (Deng et al., 2009). Could scalable pre-training methods which learn directly from web text result in a similar breakthrough in computer vision? Prior work is encouraging.

Joulin et al. (2016) demonstrated that CNNs trained to predict words in image captions can learn representations competitive with ImageNet training. Li et al. (2017) then extended this approach to predicting phrase n-grams in addition to individual words and demonstrated the ability of their system to zero-shot transfer to other image classification datasets. Adopting more recent architectures and pre-training approaches, VirTex (Desai & Johnson, 2020), ICMLM (Bulent Sariyildiz et al., 2020), and ConVIRT (Zhang et al., 2020) have recently demonstrated the potential of transformer-based language modeling, masked language modeling, and contrastive objectives to learn image representations from text.

However, the aforementioned models still under-perform current SOTA computer vision models such as Big Transfer (Kolesnikov et al., 2019) and the weakly supervised ResNeXt (Mahajan et al., 2018). A crucial difference is scale. While Mahajan et al. (2018) and Kolesnikov et al. (2019) trained for accelerator years on millions to billions of images, VirTex, ICMLM, and ConVIRT trained for accelerator days on one to two hundred thousand images. We close this gap and study the behaviors of image models trained from natural language supervision at large scale. We demonstrate that a simplified version of ConVIRT trained from scratch, which we call CLIP, for Contrastive Language-Image Pre-training, is an efficient and scalable method of learning from natural language supervision. We find that

Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên

Alec Radford^{*1} Kim Jong Wook^{*1} Chris Hallacy¹ Aditya Ramesh¹ Gabriel Goh¹ Sandhini Agarwal¹
Girish Sastry¹ Amanda Askell¹ Pamela Mishkin¹ Jack Clark¹ Gretchen Krueger¹ Ilya Sutskever¹

Tóm tắt

Hệ thống thị giác máy tính SOTA được đào tạo để dự đoán một tập hợp cố định các danh mục đối tượng được xác định trước. Hình thức giám sát hạn chế này giới hạn họ tính tổng quát và khả năng sử dụng vì có thêm nhãn dữ liệu cần thiết để chỉ định bất kỳ khái niệm trực quan nào khác. Học trực tiếp từ văn bản thô về hình ảnh là một giải pháp thay thế đầy hứa hẹn tận dụng nguồn giám sát rộng hơn nhiều. Chúng tôi chứng minh rằng nhiệm vụ đào tạo trước đơn giản là dự đoán chủ thích nào đi kèm với hình ảnh nào là một cách hiệu quả và có thể mở rộng để học hình ảnh SOTA biểu diễn từ đầu trên một tập dữ liệu gồm 400 triệu cặp (hình ảnh, văn bản) được thu thập từ internet. Sau khi đào tạo trước, ngôn ngữ tự nhiên được sử dụng để tham khảo các khái niệm trực quan đã học (hoặc mô tả khái niệm mới) những cái) cho phép chuyển giao mô hình không cần bắn sang nhiệm vụ hạ lưu. Chúng tôi nghiên cứu hiệu suất trên 30 bộ dữ liệu thị giác máy tính khác nhau, trải dài các tác vụ như OCR, nhận dạng hành động trong video, định vị địa lý và nhiều loại hạt mìn phân loại đối tượng. Mô hình chuyển giao không tầm thường cho hầu hết các nhiệm vụ và thường mang tính cạnh tranh với đường cơ sở được giám sát đầy đủ mà không cần cho bất kỳ tập dữ liệu đào tạo cụ thể nào. Ví dụ, chúng tôi phù hợp với độ chính xác của ResNet50 gốc trên ImageNet zero-shot mà không cần sử dụng bất kỳ 1,28 triệu ví dụ đào tạo đã được đào tạo trên. Chúng tôi phát hành mã và mô hình được đào tạo trước của mình trọng số tại <https://github.com/OpenAI/CLIP>.

1. Giới thiệu và động viên công việc

Phương pháp đào tạo trước học trực tiếp từ văn bản thô đã cách mạng hóa NLP trong vài năm qua (Dai & Le, 2015; Peters và cộng sự, 2018; Howard & Ruder, 2018; Radford và cộng sự, 2018; Devlin và cộng sự, 2018; Raffel và cộng sự, 2019). Các sự phát triển của “text-to-text” như một đầu vào-dầu ra chuẩn hóa

*Đóng góp ngang nhau 1OpenAI, San Francisco, CA 94110, Hoa Kỳ.
Liên hệ: <{alec, jongwook}@openai.com>.

giao diện (McCann và cộng sự, 2018; Radford và cộng sự, 2019; Raffel et al., 2019) đã cho phép các kiến trúc không phụ thuộc vào tác vụ chuyên giao không cần thực hiện sang các tập dữ liệu hạ lưu. Các hệ thống chủ chốt như GPT-3 (Brown và cộng sự, 2020) hiện đang cạnh tranh trên nhiều nhiệm vụ với các mô hình tùy chỉnh trong khi hầu như không cần dữ liệu đào tạo cụ thể của tập dữ liệu.

Những kết quả này cho thấy rằng giám sát tổng hợp có thể tiếp cận được với các phương pháp đào tạo trước hiện đại trong các bộ sưu tập văn bản quy mô web vượt trội hơn so với giám sát đám đông chất lượng cao được gắn nhãn Bộ dữ liệu NLP. Tuy nhiên, trong các lĩnh vực khác như máy tính tầm nhìn vẫn là thông lệ chuẩn để đào tạo trước các mô hình trên các tập dữ liệu được gắn nhãn đám đông như ImageNet (Deng và cộng sự, 2009). Có thể mở rộng các phương pháp đào tạo trước có thể học trực tiếp từ văn bản web dẫn đến một bước đột phá tương tự trong máy tính tầm nhìn? Công việc trước đây rất đáng khích lệ.

Joulin và cộng sự (2016) đã chứng minh rằng CNN được đào tạo để dự đoán các từ trong chủ đề hình ảnh có thể học các biểu diễn cạnh tranh với đào tạo ImageNet. Sau đó, Li và cộng sự (2017) đã mở rộng cách tiếp cận này để dự đoán các n-gram cụm từ ngoài các từ riêng lẻ và chứng minh khả năng của hệ thống của họ để chuyển giao không-shot sang các tập dữ liệu phân loại hình ảnh khác. Áp dụng các kiến trúc mới hơn và phương pháp tiếp cận trước khi đào tạo, VirTex (Desai & Johnson, 2020), ICMLM (Bulent Sariyildiz và cộng sự, 2020) và ConVIRT (Zhang và cộng sự, 2020) gần đây đã chứng minh tiềm năng của mô hình ngôn ngữ dựa trên bộ chuyển đổi, được che giấu mô hình hóa ngôn ngữ và mục tiêu tương phản để tìm hiểu biểu diễn hình ảnh từ văn bản.

Tuy nhiên, các mô hình nói trên vẫn hoạt động kém các mô hình thị giác máy tính SOTA hiện tại như Big Transfer (Kolesnikov và cộng sự, 2019) và mô hình giám sát yếu

ResNeXt (Mahajan và cộng sự, 2018). Một sự khác biệt quan trọng là thang đo. Trong khi Mahajan et al. (2018) và Kolesnikov et al. (2019) được đào tạo cho những năm tăng tốc trên hàng triệu đến hàng tỷ của hình ảnh, VirTex, ICMLM và ConVIRT được đào tạo trong những ngày tăng tốc trên một đến hai trăm nghìn hình ảnh. Chúng tôi thu hẹp khoảng cách này và nghiên cứu hành vi của các mô hình hình ảnh được đào tạo từ sự giám sát ngôn ngữ tự nhiên ở quy mô lớn. Chúng tôi chứng minh rằng một phiên bản đơn giản hóa của ConVIRT đã được đào tạo từ đầu, mà chúng tôi gọi là CLIP, viết tắt của Ngôn ngữ tương phản - Đào tạo trước hình ảnh, là một phương pháp hiệu quả và có thể mở rộng học từ sự giám sát ngôn ngữ tự nhiên. Chúng tôi thấy rằng

Learning Transferable Visual Models From Natural Language Supervision

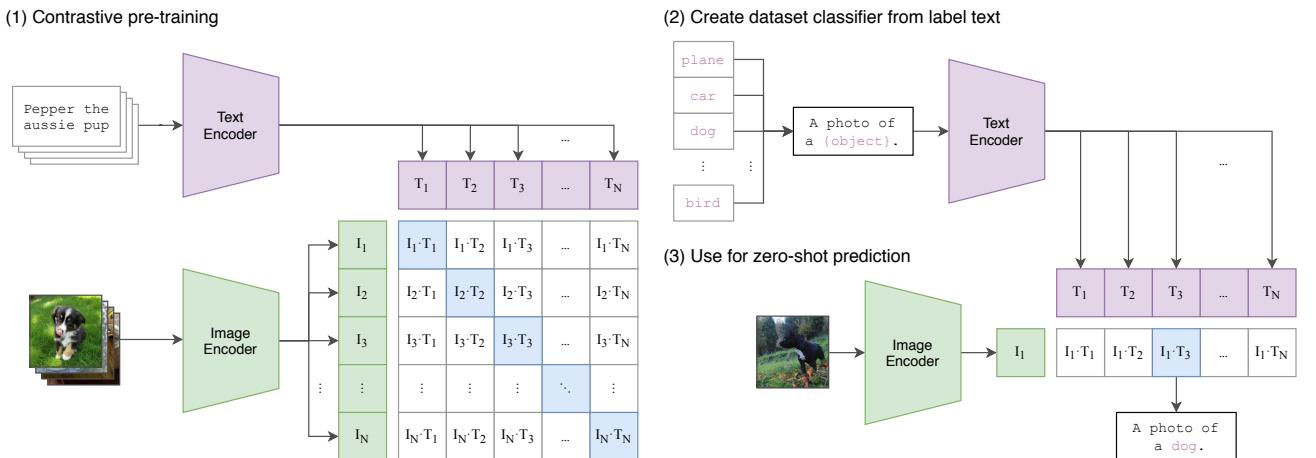


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the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

CLIP learns to perform a wide set of tasks during pre-training including OCR, geo-localization, action recognition, and outperforms the best publicly available ImageNet model while being more computationally efficient. We also find that zero-shot CLIP models are much more robust than equivalent accuracy supervised ImageNet models.

2. Approach

At the core of our work is the idea of learning perception from the supervision contained in natural language paired with images. In the following subsections we detail our specific approach.

2.1. Creating a Sufficiently Large Dataset

Existing work has mainly used three datasets, MS-COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), and YFCC100M (Thomee et al., 2016). While MS-COCO and Visual Genome are high quality crowd-labeled datasets, they are small by modern standards with approximately 100,000 training photos each. By comparison, other computer vision systems are trained on up to 3.5 billion Instagram photos (Mahajan et al., 2018). YFCC100M, at 100 million photos, is a possible alternative, but the metadata for each image is sparse and of varying quality. Many images use automatically generated filenames like 20160716_113957.JPG as “titles” or contain “descriptions” of camera exposure settings. After filtering to keep only images with natural language titles and/or descriptions in English, the dataset shrunk by a factor of 6 to only 15 million photos. This is approximately the same size as ImageNet.

A major motivation for natural language supervision is the

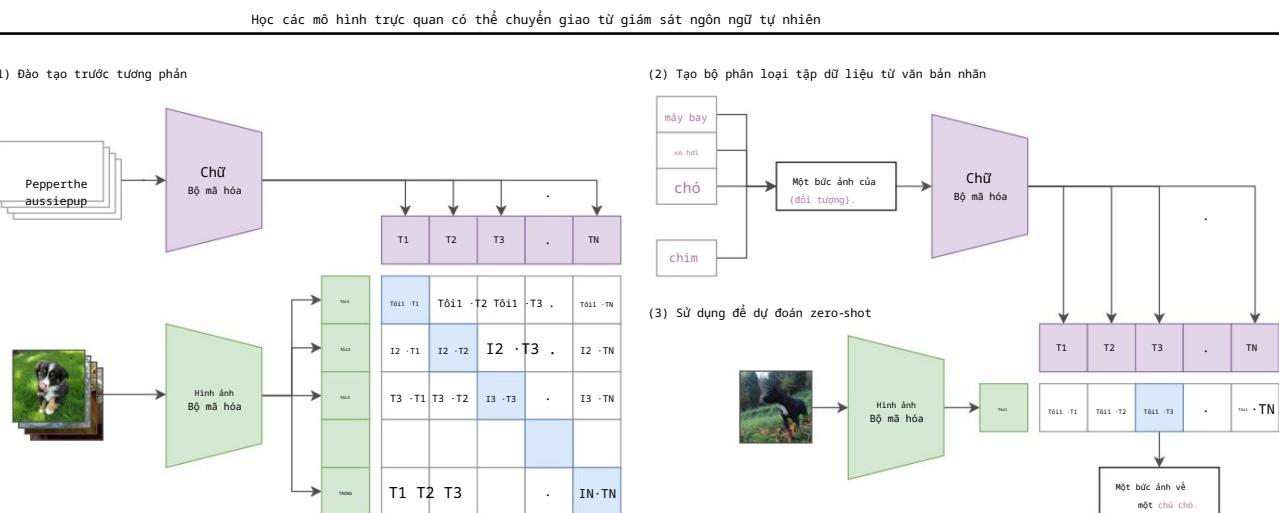
large quantities of data of this form available publicly on the internet. To test this we constructed a new dataset of 400 million (image, text) pairs collected from a variety of publicly available sources on the Internet. To attempt to cover as broad a set of visual concepts as possible, we search for (image, text) pairs as part of the construction process whose text includes one of a set of 500,000 queries. We approximately class balance the results by including up to 20,000 (image, text) pairs per query. The resulting dataset has a similar total word count as the WebText dataset used to train GPT-2. We refer to this dataset as WIT for WebImageText.¹

2.2. Selecting an Efficient Pre-Training Method

Our initial approach, similar to VirTex, jointly trained an image CNN and text transformer from scratch to predict the caption of an image. However, we encountered difficulties efficiently scaling this method. In Figure 2 we show that a 63 million parameter transformer language model, which already uses twice the compute of its ResNet50 image encoder, learns to recognize ImageNet classes three times slower than an approach similar to Joulin et al. (2016) that predicts a bag-of-words encoding of the same text.

Recent work in contrastive representation learning has found that contrastive objectives can outperform the equivalent predictive objective (Tian et al., 2019). Noting this finding,

¹The base query list is all words occurring at least 100 times in the English version of Wikipedia. This is augmented with bi-grams with high pointwise mutual information for the pair (Church & Hanks, 1990) as well as the names of all Wikipedia articles above a certain search volume. Finally all WordNet (Miller, 1995) synsets not already in the query list are added.



Hình 1. Tóm tắt phương pháp tiếp cận của chúng tôi. Trong khi các mô hình hình ảnh chuẩn cùng nhau đào tạo một trình trích xuất đặc điểm hình ảnh và một trình phân loại tuyến tính để dự đoán một số nhãn, CLIP cùng nhau đào tạo một trình mã hóa hình ảnh và một trình mã hóa văn bản để dự đoán các cặp chính xác của một loạt các ví dụ đào tạo (hình ảnh, văn bản). Vào thời điểm kiểm tra, trình mã hóa văn bản đã học tổng hợp một trình phân loại tuyến tính zero-shot bằng cách nhúng tên hoặc mô tả các lớp trong tập dữ liệu mục tiêu.

CLIP học cách thực hiện một loạt các tác vụ trong quá trình đào tạo trước bao gồm OCR, định vị địa lý, nhận dạng hành động và vượt trội hơn mô hình ImageNet công khai tốt nhất trong khi vẫn hiệu quả hơn về mặt tính toán. Chúng tôi cũng thấy rằng các mô hình CLIP zero-shot mạnh mẽ hơn nhiều so với các mô hình ImageNet có giám sát độ chính xác tương đương.

2. Cách tiếp cận

Cốt lõi công việc của chúng tôi là ý tưởng học nhận thức từ sự giám sát chưa trong ngôn ngữ tự nhiên kết hợp với hình ảnh. Trong các tiêu mục sau, chúng tôi trình bày chi tiết cách tiếp cận cụ thể của mình.

2.1. Tạo một tập dữ liệu đủ lớn

Công việc hiện tại chủ yếu sử dụng ba tập dữ liệu, MS-COCO (Lin và cộng sự, 2014), Visual Genome (Krishna và cộng sự, 2017) và YFCC100M (Thomee và cộng sự, 2016). Mặc dù MS-COCO và Visual Genome là các tập dữ liệu được gắn nhãn cộng đồng chất lượng cao, nhưng chúng có quy mô nhỏ theo tiêu chuẩn hiện đại với khoảng 100.000 ảnh đào tạo cho mỗi tập. Để so sánh, các hệ thống thị giác máy tính khác được đào tạo trên 3,5 tỷ ảnh Instagram (Mahajan và cộng sự, 2018). YFCC100M, với 100 triệu ảnh, là một giải pháp thay thế khả thi, nhưng siêu dữ liệu cho mỗi hình ảnh rất ít và chất lượng khác nhau. Nhiều hình ảnh sử dụng tên tệp được tạo tự động như 20160716_113957.JPG làm “tiêu đề” hoặc chứa “mô tả” về cài đặt phơi sáng của máy ảnh. Sau khi lọc để chỉ giữ lại những hình ảnh có tiêu đề và/hoặc mô tả bằng ngôn ngữ tự nhiên bằng tiếng Anh, tập dữ liệu đã giảm đi 6 lần, chỉ còn 15 triệu ảnh. Kích thước này gần bằng ImageNet.

Một động lực chính cho việc giám sát ngôn ngữ tự nhiên là

lượng lớn dữ liệu dạng này có sẵn công khai trên internet. Để kiểm tra điều này, chúng tôi đã xây dựng một tập dữ liệu mới 400 triệu cặp (hình ảnh, văn bản) được thu thập từ nhiều nguồn công khai trên Internet. Để cố gắng bao quát một tập hợp các khái niệm trực quan rộng nhất có thể, chúng tôi tìm kiếm các cặp (hình ảnh, văn bản) như một phần của quá trình xây dựng có văn bản bao gồm một trong 500.000 truy vấn.

Chúng tôi cân bằng kết quả bằng cách bao gồm tối đa 20.000 cặp (hình ảnh, văn bản) cho mỗi truy vấn. Bộ dữ liệu kết quả có tổng số từ tương tự như bộ dữ liệu WebText được sử dụng để đào tạo GPT-2. Chúng tôi gọi bộ dữ liệu này là WIT cho WebImageText.

¹

2.2. Lựa chọn phương pháp đào tạo trước hiệu quả

Cách tiếp cận ban đầu của chúng tôi, tương tự như VirTex, đã cùng nhau đào tạo một CNN hình ảnh và bộ chuyển đổi văn bản từ đầu để dự đoán chủ đề của một hình ảnh. Tuy nhiên, chúng tôi gặp phải những khó khăn khi mở rộng hiệu quả phương pháp này. Trong Hình 2, chúng tôi chỉ ra rằng một mô hình ngôn ngữ bộ chuyển đổi 63 triệu tham số, vốn đã sử dụng gấp đôi khả năng tính toán của bộ mã hóa hình ảnh ResNet50, học cách nhận dạng các lớp ImageNet chậm hơn bាន so với một cách tiếp cận tương tự như Joulin et al. (2016) dự đoán mã hóa một túi từ của cùng một văn bản.

Nghiên cứu gần đây về học tập biểu diễn tương phản đã phát hiện ra rằng các mục tiêu tương phản có thể vượt trội hơn mục tiêu dự đoán tương đương (Tian và cộng sự, 2019). Lưu ý phát hiện này,

1Danh sách truy vấn cơ sở là tất cả các từ xuất hiện ít nhất 100 lần trong phiên bản tiếng Anh của Wikipedia. Danh sách này được bổ sung thêm các bi-gram có thông tin tương hỗ tăng cao cho cặp (Church & Hanks, 1990) cũng như tên của tất cả các bài viết Wikipedia trên một khối lượng tìm kiếm nhất định. Cuối cùng, tất cả các synset WordNet (Miller, 1995) chưa có trong danh sách truy vấn đều được thêm vào.

Learning Transferable Visual Models From Natural Language Supervision

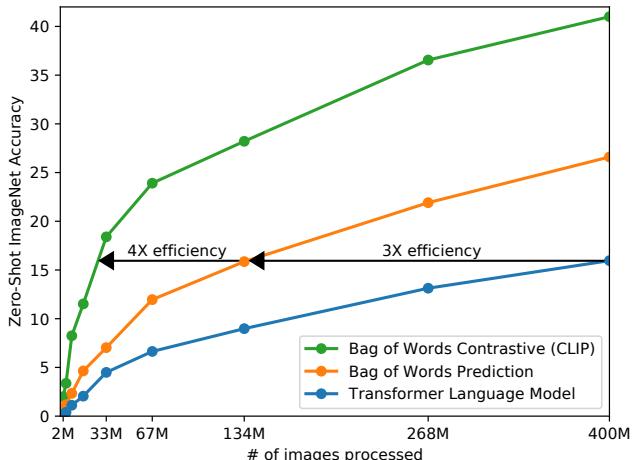


Figure 2. CLIP is much more efficient at zero-shot transfer than our image caption baseline. Although highly expressive, we found that transformer-based language models are relatively weak at zero-shot ImageNet classification. Here, we see that it learns 3x slower than a baseline which predicts a bag-of-words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

We explored training a system to solve the potentially easier proxy task of predicting only which text *as a whole* is paired with which image and not the exact words of that text. Starting with the same bag-of-words encoding baseline, we swapped the predictive objective for a contrastive objective in Figure 2, observed a further 4x efficiency improvement in the rate of zero-shot transfer to ImageNet.

Given a batch of N (image, text) pairs, CLIP is trained to predict which of the $N \times N$ possible (image, text) pairings across a batch actually occurred. To do this, CLIP learns a multi-modal embedding space by jointly training an image encoder and text encoder to maximize the cosine similarity of the image and text embeddings of the N real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings. We optimize a symmetric cross entropy loss over these similarity scores. In Figure 3 we include pseudocode for the core of an implementation of CLIP. This batch construction technique and objective was first introduced as the *multi-class N-pair loss* Sohn (2016) and was recently adapted for contrastive (text, image) representation learning in the domain of medical imaging by Zhang et al. (2020).

Since over-fitting is not a major concern, the details of training CLIP are simplified compared to Zhang et al. (2020). We train CLIP from scratch instead of initializing with pre-trained weights. We remove the non-linear projection between the representation and the contrastive embedding space. We use only a linear projection to map from each encoder’s representation to the multi-modal embedding space.

```

# 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 = 12_normalize(np.dot(I_f, W_i), axis=1)
T_e = 12_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

```

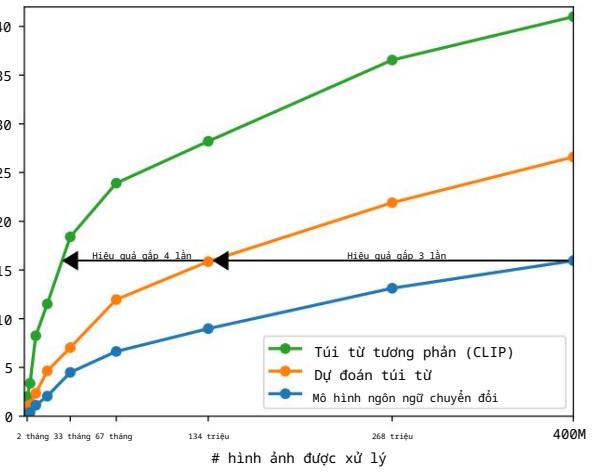
Figure 3. Numpy-like pseudocode for the core of an implementation of CLIP.

We also remove the text transformation function t_u which samples a single sentence at uniform from the text since many of the (image, text) pairs in CLIP’s pre-training dataset are only a single sentence. We also simplify the image transformation function t_v . A random square crop from resized images is the only data augmentation used during training. Finally, the temperature parameter which controls the range of the logits in the softmax, τ , is directly optimized during training as a log-parameterized multiplicative scalar to avoid turning as a hyper-parameter.

2.3. Choosing and Scaling a Model

We consider two different architectures for the image encoder. For the first, we use ResNet50 (He et al., 2016a) as the base architecture for the image encoder due to its widespread adoption and proven performance. We make several modifications to the original version using the ResNetD improvements from He et al. (2019) and the antialiased rect-2 blur pooling from Zhang (2019). We also replace the global average pooling layer with an attention pooling mechanism. The attention pooling is implemented as a single layer of “transformer-style” multi-head QKV attention where the query is conditioned on the global average-pooled representation of the image. For the second architecture, we experiment with the recently introduced Vision Transformer (ViT) (Dosovitskiy et al., 2020). We closely follow their implementation with only the minor modification of adding an additional layer normalization to the combined patch and position embeddings before the transformer and use a slightly different initialization scheme.

Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên



Hình 2. CLIP hiệu quả hơn nhiều trong việc chuyển zero-shot so với đường cơ sở chúng tôi. Mặc dù có khả năng biểu đạt cao, chúng tôi thấy rằng các mô hình ngôn ngữ dựa trên bộ chuyển đổi tương đối yếu trong phân loại ImageNet zero-shot. Ở đây, chúng tôi thấy rằng nó học chậm hơn 3 lần so với đường cơ sở dự đoán mã hóa túi từ (BoW) của văn bản (Joulin và cộng sự, 2016). Việc hoàn đổi mục tiêu dự đoán cho mục tiêu tương phản của CLIP cải thiện hiệu quả thêm 4 lần nữa.

Chúng tôi đã khám phá việc đào tạo một hệ thống để giải quyết nhiệm vụ ủy nhiệm có khả năng dễ hơn là chỉ dự đoán văn bản nào được ghép nối với hình ảnh nào chứ không phải từng từ chính xác trong văn bản đó. Bắt đầu với cùng một nhóm từ mã hóa cơ sở, chúng tôi đã hoán đổi mục tiêu dự đoán cho mục tiêu tương phản trong Hình 2, quan sát thấy hiệu quả cải thiện thêm 4 lần về tốc độ truyền dữ liệu không cần chỉnh sửa đến ImageNet.

Với một lô gồm N cặp (hình ảnh, văn bản), CLIP được đào tạo để dự đoán cặp nào trong số $N \times N$ cặp (hình ảnh, văn bản) có thể có trong một lô thực sự xảy ra. Để thực hiện điều này, CLIP học một không gian nhúng đa phương thức bằng cách đào tạo chung một bộ mã hóa hình ảnh và bộ mã hóa văn bản để tối ưu hóa cosin tương tự

tính chất của hình ảnh và nhúng văn bản của N cặp thực trong lô trong khi giảm thiểu độ tương đồng cosin của nhúng của $N^2 - N$ cặp không chính xác. Chúng tôi tối ưu hóa mất mát entropy chéo đối xứng trên các điểm tương đồng này. Trong Hình 3, chúng tôi bao gồm mã giả cho lối của một triển khai CLIP. Kỹ thuật xây dựng hàng loạt và mục tiêu này lần đầu tiên được giới thiệu là mất mát N-pair đa lớp Sohn (2016) và gần đây đã được điều chỉnh cho việc học biểu diễn tương phản (văn bản, hình ảnh) trong lĩnh vực hình ảnh y tế bởi Zhang và cộng sự (2020).

Vì việc quá phù hợp không phải là mối quan tâm lớn nên các chi tiết về đào tạo CLIP được đơn giản hóa so với Zhang et al. (2020). Chúng tôi đào tạo CLIP từ đầu thay vì khởi tạo với các trọng số được đào tạo trước. Chúng tôi loại bỏ phép chiếu phi tuyến tính giữa biểu diễn và không gian nhúng tương phản. Chúng tôi chỉ sử dụng phép chiếu tuyến tính để ánh xạ từ biểu diễn của mỗi bộ mã hóa đến không gian nhúng đa phương thức.

```

# image_encoder - ResNet hoặc Vision Transformer
# text_encoder - CBOW hoặc Text Transformer
# I[n, h, w, c] - minibatch của hình ảnh được căn chỉnh
# T[n, l] - minibatch của các văn bản được căn chỉnh
# W_i[d_i, d_e] - học được proj của hình ảnh để nhúng
# W_t[d_t, d_e] - học được proj của văn bản để nhúng # t - học được tham số nhiệt độ

# trích xuất các biểu diễn tính năng của từng phương thức
I_f = bộ mã hóa hình ảnh(I) #[n, d_i]
T_f = bộ mã hóa văn bản(T) #[n, d_t]

# nhúng đa phương thức chung [n, d_e]
I_e = 12_normalize(np.dot(I_f, W_i), trục=1)
T_e = 12_normalize(np.dot(T_f, W_t), trục=1)

# sự tương đồng cosin theo cặp được chia tỷ lệ [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# hàm mất mát đối xứng nhẫn =
np.arange(n) loss_i =
cross_entropy_loss(logits, nhẫn, trục=0) loss_t =
cross_entropy_loss(logits, nhẫn, trục=1) loss = (loss_i + loss_t)/2

```

Hình 3. Mã giả giống Numpy cho lối của một triển khai CLIP.

Chúng tôi cũng loại bỏ hàm chuyển đổi văn bản tu láy mẫu một câu duy nhất ở dạng dòng nhất từ văn bản vì nhiều cặp (hình ảnh, văn bản) trong tập dữ liệu tiền đào tạo của CLIP chỉ là một câu duy nhất. Chúng tôi cũng đơn giản hóa hàm chuyển đổi hình ảnh tv. Một hình vuông cắt ngẫu nhiên từ hình ảnh đã thay đổi kích thước là phép tăng cường dữ liệu duy nhất được sử dụng trong quá trình đào tạo. Cuối cùng, tham số nhiệt độ kiểm soát phạm vi của logit trong softmax, τ được tối ưu hóa trực tiếp trong quá trình đào tạo như một số vô hướng nhân tham số logarit để tránh chuyển đổi như một siêu tham số.

2.3. Lựa chọn và mở rộng mô hình

Chúng tôi xem xét hai kiến trúc khác nhau cho bộ mã hóa hình ảnh. Đối với kiến trúc đầu tiên, chúng tôi sử dụng ResNet50 (He et al., 2016a) làm kiến trúc cơ sở cho bộ mã hóa hình ảnh do được áp dụng rộng rãi và hiệu suất đã được chứng minh. Chúng tôi thực hiện một số sửa đổi đối với phiên bản gốc bằng cách sử dụng các cải tiến ResNetD từ He et al. (2019) và nhóm làm mờ rect-2 chống răng cửa từ Zhang (2019). Chúng tôi cũng thay thế lớp nhóm trung bình toàn cục bằng cơ chế nhóm chú ý. Nhóm chú ý được triển khai dưới dạng một lớp duy nhất của sự chú ý QKV đa đầu “kiểu máy biến áp” trong đó truy vấn được điều kiện hóa dựa trên biểu diễn nhóm trung bình toàn cục của hình ảnh. Đối với kiến trúc thứ hai, chúng tôi thử nghiệm với Vision Transformer (ViT) mới được giới thiệu gần đây (Dosovitskiy et al., 2020). Chúng tôi theo dõi chặt chẽ quá trình triển khai của họ với chỉ một sửa đổi nhỏ là thêm một lớp chuẩn hóa bổ sung vào các nhúng vị trí và bản vá kết hợp trước máy biến áp và sử dụng một lược đồ khởi tạo hơi khác một chút.

Learning Transferable Visual Models From Natural Language Supervision

The text encoder is a Transformer (Vaswani et al., 2017) with the architecture modifications described in Radford et al. (2019). As a base size we use a 12-layer 512-wide model with 8 attention heads. The transformer operates on a lower-cased byte pair encoding (BPE) representation of the text (Sennrich et al., 2015). The text sequence is bracketed with [SOS] and [EOS] tokens and the activations of the highest layer of the transformer at the [EOS] token are used as the feature representation of the text which is layer normalized and then linearly projected into the multi-modal embedding space. Masked self-attention was used in the text encoder to preserve the ability to add language modeling as an auxiliary objective, though exploration of this is left as future work.

While previous computer vision research has often scaled models by increasing the width (Mahajan et al., 2018) or depth (He et al., 2016a) in isolation, for the ResNet image encoders we adapt the approach of Tan & Le (2019) which found that allocating additional compute across all of width, depth, and resolution outperforms allocating it to only one dimension. We use a simple variant which allocates additional compute equally to increasing the width, depth, and resolution of the model. For the text encoder, we only scale the width of the model to be proportional to the calculated increase in width of the ResNet and do not scale the depth at all, as we found CLIP’s performance to be less sensitive to the text encoder.

2.4. Pre-training

We train a series of 5 ResNets and 3 Vision Transformers. For the ResNets we train a ResNet50, a ResNet101, and then 3 more which follow EfficientNet-style model scaling and use approximately 4x, 16x, and 64x the compute of a ResNet50. They are denoted as RN50x4, RN50x16, and RN50x64 respectively. For the Vision Transformers we train a ViT-B/32, a ViT-B/16, and a ViT-L/14. The largest ResNet model, RN50x64, took 18 days to train on 592 V100 GPUs while the largest Vision Transformer took 12 days on 256 V100 GPUs. For the ViT-L/14 we also pre-train at a higher 336 pixel resolution for one additional epoch to boost performance similar to FixRes (Touvron et al., 2019). We denote this model as ViT-L/14@336px. Unless otherwise specified, all results reported in this paper as “CLIP” use this model which we found to perform best. Full model hyperparameters and details are in supplementary material.

2.5. Using CLIP

CLIP is pre-trained to predict if an image and a text snippet are paired together in WIT. To apply CLIP to downstream tasks, we reuse this capability and study the zero-shot transfer performance of CLIP on standard computer vision datasets. Similar to Radford et al. (2019) we motivate

	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 work. CLIP improves performance on all three datasets by a large amount. This improvement reflects many differences since the development of Visual N-Grams (Li et al., 2017).

this as a way of measuring the *task learning* capability of a system (as opposed to its *representation learning* capability). For each dataset, we use the names of all the classes in the dataset as the set of potential text pairings and predict the most probable (image, text) pair according to CLIP. We additionally experiment with providing CLIP with text prompts to help specify the task as well as ensembling multiple of these templates in order to boost performance. However, since the vast majority of unsupervised and self-supervised computer vision research focuses on representation learning, we also investigate this for CLIP using the common linear probe protocol.

3. Analysis

3.1. Initial Comparison to Visual N-Grams

To our knowledge, Visual N-Grams (Li et al., 2017) first studied zero-shot transfer to existing image classification datasets in the manner described above. It is also the only other work we are aware of that has studied zero-shot transfer to standard image classification datasets using a task agnostic pre-trained model. In Table 1 we compare Visual N-Grams to CLIP. The best CLIP model improves accuracy on ImageNet from a proof of concept 11.5% to 76.2% and matches the performance of the original ResNet50 despite using none of the 1.28 million crowd-labeled training examples. Additionally, the top-5 accuracy of CLIP models are noticeably higher and this model has a 95% top-5 accuracy, matching Inception-V4 (Szegedy et al., 2016). The ability to match the performance of a strong, fully supervised baseline in a zero-shot setting suggests CLIP is a significant step towards flexible and practical zero-shot computer vision classifiers. This comparison is not direct because many differences between CLIP and Visual N-Grams were not controlled for. As a closer comparison, we trained a CLIP ResNet50 on the same YFCC100M dataset that Visual N-Grams was trained on and found it matched their reported ImageNet performance within a V100 GPU day. This baseline was also trained from scratch instead of being initialized from pre-trained ImageNet weights as in Visual N-Grams.

Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên

Bộ mã hóa văn bản là một Transformer (Vaswani và cộng sự, 2017) với các sửa đổi về kiến trúc được mô tả trong Radford và cộng sự (2019). Với kích thước cơ sở, chúng tôi sử dụng mô hình 12 lớp rộng 512 với 8 đầu chú ý. Bộ biến đổi hoạt động trên biểu diễn mã hóa cặp byte viết thường (BPE) của văn bản (Sennrich và cộng sự, 2015). Chuỗi văn bản được đóng ngoặc với các mã thông báo [SOS] và [EOS] và các kích hoạt của lớp cao nhất của bộ biến đổi tại mã thông báo [EOS] được sử dụng làm biểu diễn tinh nang của văn bản được chuẩn hóa theo lớp và sau đó được chiếu tuyển tính vào không gian nhung đa phương thức. Sự tự chú ý được che giấu đã được sử dụng trong bộ mã hóa văn bản để duy trì khả năng thêm mô hình hóa ngôn ngữ làm mục tiêu phụ trợ, mặc dù việc khám phá điều này được để lại cho công việc trong tương lai.

Trong khi nghiên cứu thị giác máy tính trước đây thường mở rộng mô hình bằng cách tăng chiều rộng (Mahajan và cộng sự, 2018) hoặc chiều sâu (He và cộng sự, 2016a) một cách riêng biệt, đối với bộ mã hóa hình ảnh ResNet, chúng tôi áp dụng phương pháp của Tan & Le (2019), phát hiện ra rằng việc phân bổ thêm tính toán trên toàn bộ chiều rộng, chiều sâu và độ phân giải sẽ hiệu quả hơn việc chỉ phân bổ cho một chiều. Chúng tôi sử dụng một biến thể đơn giản phân bổ thêm tính toán như nhau khi tăng chiều rộng, chiều sâu và độ phân giải của mô hình. Đối với bộ mã hóa văn bản, chúng tôi chỉ mở rộng chiều rộng của mô hình theo tỷ lệ thuận với mức tăng chiều rộng được tính toán của ResNet và không mở rộng chiều sâu chút nào vì chúng tôi thấy hiệu suất của CLIP ít nhạy cảm hơn với bộ mã hóa văn bản.

2.4. Đào tạo trước

Chúng tôi đào tạo một loạt 5 ResNet và 3 Vision Transformer. Đối với ResNet, chúng tôi đào tạo một ResNet50, một ResNet101 và sau đó là 3 mô hình nữa theo mô hình EfficientNet và sử dụng khoảng 4x, 16x và 64x khả năng tính toán của ResNet50. Chúng được ký hiệu lần lượt là RN50x4, RN50x16 và RN50x64. Đối với Vision Transformer, chúng tôi đào tạo một ViT-B/32, một ViT-B/16 và một ViT-L/14. Mô hình ResNet lớn nhất, RN50x64, mất 18 ngày để đào tạo trên 592 GPU V100 trong khi Vision Transformer lớn nhất mất 12 ngày trên 256 GPU V100. Đối với ViT-L/14, chúng tôi cũng đào tạo trước ở độ phân giải 336 pixel cao hơn trong một ký nguyên bổ sung để tăng hiệu suất tương tự như FixRes (Touvron và cộng sự, 2019). Chúng tôi ký hiệu mô hình này là ViT-L/14@336px. Trừ khi được chỉ định khác, tất cả các kết quả được báo cáo trong bài báo này là “CLIP” đều sử dụng mô hình này mà chúng tôi thấy là hoạt động tốt nhất. Siêu tham số và chi tiết mô hình đầy đủ có trong tài liệu bổ sung.

2.5. Sử dụng CLIP

CLIP được đào tạo trước để dự đoán xem một hình ảnh và một đoạn văn bản có được ghép nối với nhau trong WIT hay không. Để áp dụng CLIP cho các tác vụ hạ nguồn, chúng tôi sử dụng lại khả năng này và nghiên cứu hiệu suất truyền zero-shot của CLIP trên các tập dữ liệu thị giác máy tính tiêu chuẩn. Tương tự như Radford et al. (2019), chúng tôi thúc đẩy

	aYahoo	ImageNet	MẶT TRỜI
N-Gram trực quan	72,4	CLIP	11,5
98,4			23,0
		76,2	58,5

Bảng 1. So sánh CLIP với công việc phân loại hình ảnh chuyển zero-shot trước đây. CLIP cải thiện hiệu suất trên cả ba tập dữ liệu với số lượng lớn. Sự cải thiện này phản ánh nhiều điểm khác biệt kể từ khi phát triển Visual N-Grams (Li et al., 2017).

đây là một cách đo lường khả năng học tác vụ của một hệ thống (khác với khả năng học biểu diễn của nó).

Đối với mỗi tập dữ liệu, chúng tôi sử dụng tên của tất cả các lớp trong tập dữ liệu làm tập hợp các cặp văn bản tiềm năng và dự đoán cặp có khả năng xảy ra nhất (hình ảnh, văn bản) theo CLIP. Chúng tôi cũng thử nghiệm bằng cách cung cấp cho CLIP các lời nhắc văn bản để giúp chỉ định nhiệm vụ cũng như tập hợp nhiều mẫu này để tăng hiệu suất. Tuy nhiên, vì phần lớn các nghiên cứu về thị giác máy tính không giám sát và tự giám sát tập trung vào việc học biểu diễn, chúng tôi cũng nghiên cứu điều này cho CLIP bằng cách sử dụng giao thức thăm dò tuyển tính chung.

3. Phân tích

3.1. So sánh ban đầu với Visual N-Gram

Theo hiểu biết của chúng tôi, Visual N-Grams (Li và cộng sự, 2017) đã nghiên cứu đầu tiên về việc chuyển zero-shot sang các tập dữ liệu phân loại hình ảnh hiện có theo cách được mô tả trên. Đây cũng là công trình duy nhất khác mà chúng tôi biết đã nghiên cứu việc chuyển zero-shot sang các tập dữ liệu phân loại hình ảnh tiêu chuẩn bằng cách sử dụng một mô hình được đào tạo trước không phụ thuộc vào tác vụ. Trong Bảng 1, chúng tôi so sánh Visual N-Grams với CLIP. Mô hình CLIP tốt nhất cải thiện độ chính xác trên ImageNet từ bằng chứng khái niệm 11,5% lên 76,2% và phù hợp với hiệu suất của ResNet50 ban đầu mặc dù không sử dụng bất kỳ ví dụ đào tạo nào trong số 1,28 triệu ví dụ được gắn nhãn đám đông. Ngoài ra, độ chính xác trong top-5 của các mô hình CLIP cao hơn đáng kể và mô hình này có độ chính xác trong top-5 là 95%, phù hợp với Inception-V4 (Szegedy và cộng sự, 2016). Khả năng khớp với hiệu suất của một đường cơ sở mạnh, được giám sát đầy đủ trong bài đặt zero-shot cho thấy CLIP là một bước tiến đáng kể hướng tới các bộ phân loại thị giác máy tính zero-shot linh hoạt và thiết thực. So sánh này không trực tiếp vì nhiều điểm khác biệt giữa CLIP và Visual N-Gram không được kiểm soát. Để so sánh chặt chẽ hơn, chúng tôi đã đào tạo một CLIP ResNet50 trên cùng một tập dữ liệu YFCC100M mà Vi-sual N-Gram đã được đào tạo và thấy rõ ràng nó khớp với hiệu suất ImageNet được báo cáo của họ trong một ngày GPU V100.

Đường cơ sở này cũng được đào tạo từ đầu thay vì được khởi tạo từ trọng số ImageNet được đào tạo trước như trong Visual N-Gram.

Learning Transferable Visual Models From Natural Language Supervision

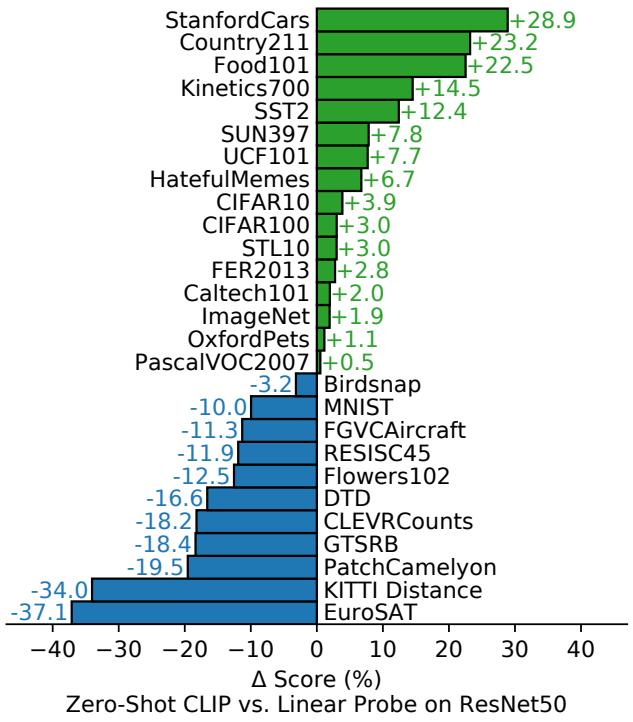


Figure 4. Zero-shot CLIP is competitive with a fully supervised baseline. Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet50 features on 16 datasets, including ImageNet.

3.2. Zero-Shot Performance

In computer vision, zero-shot learning usually refers to the study of generalizing to unseen object categories in image classification (Lampert et al., 2009). We instead use the term in a broader sense and study generalization to unseen datasets. We motivate this as a proxy for performing unseen tasks, as aspired to in the zero-data learning paper of Larochelle et al. (2008). While much research in the field of unsupervised learning focuses on the *representation learning* capabilities of machine learning systems, we motivate studying zero-shot transfer as a way of measuring the *task-learning* capabilities of machine learning systems. In this view, a dataset evaluates performance on a task on a specific distribution. However, many popular computer vision datasets were created by the research community primarily as benchmarks to guide the development of generic image classification methods rather than measuring performance on a specific task. To our knowledge, Visual N-Grams (Li et al., 2017) first studied zero-shot transfer to existing image classification datasets in the manner described above.

To conduct a more comprehensive analysis, we implement models contextualizes the task-learning capabilities of CLIP, comparing to few-shot methods is a more direct comparison, since zero-shot is its limit. In Figure 5, we visualize how zero-shot CLIP compares to few-shot logistic regression on

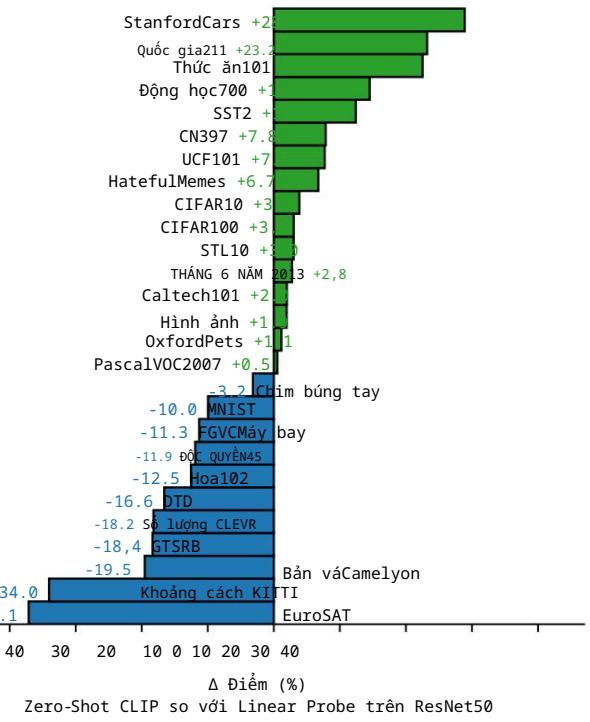
results. To start, we look at how well CLIP’s zero-shot classifiers perform when compared to the a simple off-the-shelf baseline: fitting a fully supervised, regularized, logistic regression classifier on the features of the canonical ResNet50. In Figure 4 we show this comparison across 27 datasets.

Zero-shot CLIP outperforms this baseline slightly and wins on 16 of the 27 datasets. The dataset zero-shot CLIP improves by the most is STL10, a dataset designed to encourage unsupervised learning by containing only a limited number of labeled examples. Zero-shot CLIP, without using any training examples, achieves 99.3% on this dataset which appears to be a new SOTA. On fine-grained classification tasks, we observe a wide spread in performance. On two of these datasets, Stanford Cars and Food101, zero-shot CLIP outperforms logistic regression on ResNet50 features by over 20% while on Flowers102 and FGVC Aircraft, zero-shot CLIP underperforms by over 10%. We suspect these differences are primarily due to varying amounts of per-task supervision between WIT and ImageNet. On “general” object classification datasets such as ImageNet, CIFAR10, and PascalVOC2007 performance is relatively similar with a slight advantage for zero-shot CLIP. Zero-shot CLIP significantly outperforms a ResNet50 on two datasets measuring action recognition in videos. On Kinetics700, CLIP outperforms a ResNet50 by 14.5%. Zero-shot CLIP also outperforms a ResNet50’s features by 7.7% on UCF101. We speculate this is due to natural language providing wider supervision for visual concepts involving verbs, compared to the noun-centric object supervision in ImageNet.

Looking at where zero-shot CLIP notably underperforms, we see that zero-shot CLIP is quite weak on several specialized, complex, or abstract tasks such as satellite image classification (EuroSAT and RESISC45), lymph node tumor detection (PatchCamelyon), counting objects in synthetic scenes (CLEVRCounts), self-driving related tasks such as German traffic sign recognition (GTSRB), recognizing distance to the nearest car (KITTI Distance). These results highlight the poor capability of zero-shot CLIP on more complex tasks. By contrast, non-expert humans can robustly perform several of these tasks, such as counting, satellite image classification, and traffic sign recognition, suggesting significant room for improvement. However, we caution that it is unclear whether measuring zero-shot transfer, as opposed to few-shot transfer, is a meaningful evaluation for difficult tasks that a learner has no prior experience with, such as lymph node tumor classification for almost all humans (and possibly CLIP).

While comparing zero-shot performance to fully supervised models contextualizes the task-learning capabilities of CLIP, comparing to few-shot methods is a more direct comparison, since zero-shot is its limit. In Figure 5, we visualize how zero-shot CLIP compares to few-shot logistic regression on

Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên



Hình 4. Zero-shot CLIP có tính cạnh tranh với đường cơ sở được giám sát hoàn toàn. Trong bộ đánh giá 27 tập dữ liệu, một zero-shot CLIP bộ phân loại hoạt động tốt hơn bộ phân loại tuyến tính được giám sát đầy đủ được lắp trên ResNet50 có trong 16 tập dữ liệu, bao gồm ImageNet.

3.2. Hiệu suất Zero-Shot

Trong tầm nhìn máy tính, học không bắn thường để cập đến nghiên cứu về việc khai quật hóa các loại đối tượng không nhìn thấy trong hình ảnh phân loại (Lampert và cộng sự, 2009). Thay vào đó, chúng tôi sử dụng thuật ngữ theo nghĩa rộng hơn và nghiên cứu khai quật hóa đến những điều chưa thấy bộ dữ liệu. Chúng tôi thúc đẩy điều này như một proxy để thực hiện các nhiệm vụ chưa thấy, như mong muốn trong bài báo học tập dữ liệu bằng không của Larochelle et al. (2008). Trong khi nhiều nghiên cứu trong lĩnh vực học tập không giám sát tập trung vào khả năng học tập biểu diễn của các hệ thống học máy, chúng tôi thúc đẩy nghiên cứu chuyển giao không bắn như một cách để đo lường khả năng học tác vụ của các hệ thống máy học. Trong này xem, một tập dữ liệu đánh giá hiệu suất của một nhiệm vụ trên một phân phối cụ thể. Tuy nhiên, nhiều thí giác máy tính phổ biến các tập dữ liệu được tạo ra chủ yếu bởi cộng đồng nghiên cứu như là chuẩn mực để hướng dẫn sự phát triển của hình ảnh chung phương pháp phân loại thay vì đo lường hiệu suất trên một nhiệm vụ cụ thể. Theo hiểu biết của chúng tôi, Visual N-Grams (Li et al., 2017) lần đầu tiên nghiên cứu chuyển giao không ảnh sang hình ảnh hiện có bộ dữ liệu phân loại theo cách mô tả ở trên

Để tiến hành phân tích toàn diện hơn, chúng tôi thực hiện một bộ đánh giá lớn hơn nhiều được trình bày chi tiết trong phần bổ sung vật liệu. Tổng cộng chúng tôi mở rộng từ 3 tập dữ liệu được báo cáo trong Visual N-Grams bao gồm hơn 30 tập dữ liệu và so sánh với hơn 50 hệ thống thí giác máy tính hiện có để ngửi cảnh hòa

kết quả. Đầu tiên, chúng ta hãy xem các bộ phân loại zero-shot của CLIP hoạt động tốt như thế nào khi so sánh với một bộ phân loại đơn giản có sẵn đường cơ sở: lắp một bộ phân loại hồi quy logistic được giám sát đầy đủ, chính quy trên các đặc điểm của ResNet50 chuẩn. Trong Hình 4, chúng tôi trình bày sự so sánh này trên 27 tập dữ liệu.

Zero-shot CLIP vượt trội hơn chút so với đường cơ sở này và giành chiến thắng trên 16 trong số 27 tập dữ liệu. Tập dữ liệu mà CLIP zero-shot cải thiện nhiều nhất là STL10, một tập dữ liệu được thiết kế để khuyến khích học không giám sát bằng cách chỉ chứa một số lượng ví dụ được dán nhãn. Clip Zero-shot, không sử dụng bất kỳ ví dụ đào tạo nào, đạt được 99,3% trên tập dữ liệu này có vẻ như là một SOTA mới. Về phân loại chi tiết

nhiệm vụ, chúng tôi quan sát thấy sự lan rộng trong hiệu suất. Trên hai các tập dữ liệu này, Stanford Cars và Food101, zero-shot CLIP vượt trội hơn hồi quy logistic trên các tính năng ResNet50 bằng 20% trong khi trên Flowers102 và FGVC Aircraft, CLIP không bắn phát nào hoạt động kém hơn 10%. Chúng tôi nghĩ rằng những điều này sự khác biệt chủ yếu là do số lượng khác nhau của mỗi nhiệm vụ giám sát giữa WIT và ImageNet. Trên các tập dữ liệu phân loại đối tượng “chung” như ImageNet, CIFAR10 và

Hiệu suất của PascalVOC2007 tương đối giống với lợi thế nhỏ cho CLIP zero-shot. CLIP zero-shot vượt trội hơn đáng kể so với ResNet50 trên hai tập dữ liệu do lưỡng nhận dạng hành động trong video. Trên Kinetics700, CLIP vượt trội hơn ResNet50 14,5%. Zero-shot CLIP cũng vượt trội hơn tính năng của ResNet50 7,7% trên UCF101.

Chúng tôi suy đoán điều này là do ngôn ngữ tự nhiên cung cấp phạm vi rộng hơn giám sát các khái niệm trực quan liên quan đến động từ, so sánh để giám sát đối tượng lấy danh từ làm trung tâm trong ImageNet.

Nhìn vào nơi mà CLIP không bắn phát nào có hiệu suất kém đáng kể, chúng ta thấy rằng CLIP zero-shot khá yếu trong một số nhiệm vụ chuyên biệt, phức tạp hoặc trừu tượng như hình ảnh vệ tinh phân loại (EuroSAT và RESISC45), khối u hạch bạch huyết phát hiện (PatchCamelyon), đếm các đối tượng trong tổng hợp cảnh (CLEVRCounts), các nhiệm vụ liên quan đến lái xe tự động như Nhận dạng biển báo giao thông của Đức (GTSRB), nhận dạng khoảng cách đến xe gần nhất (Khoảng cách KITTI). Những kết quả này làm nổi bật khả năng kém của CLIP zero-shot trên nhiều hơn nhiệm vụ phức tạp. Ngược lại, con người không chuyên nghiệp có thể mạnh mẽ thực hiện một số nhiệm vụ này, chẳng hạn như đếm, vẽ tinh phân loại hình ảnh và nhận dạng biển báo giao thông, ghi ý cùn nhiều chỗ để cải thiện. Tuy nhiên, chúng tôi cảnh báo rằng không rõ liệu việc đo lường chuyển giao không có phát bắn, như trái ngược với việc chuyển giao ít lần, là một đánh giá có ý nghĩa đối với những nhiệm vụ khó mà người học chưa có kinh nghiệm trước đó, chẳng hạn như phân loại khối u hạch bạch huyết cho hầu hết con người (và có thể là CLIP).

Trong khi so sánh hiệu suất không bắn với hiệu suất được giám sát hoàn toàn các mô hình ngôn ngữ cảnh hóa khả năng học tập nhiệm vụ của CLIP, so sánh với phương pháp ít phát bắn là một sự so sánh trực tiếp hơn, vi zero-shot là giới hạn của nó. Trong Hình 5, chúng ta hình dung cách so sánh hồi quy logistic zero-shot với few-shot

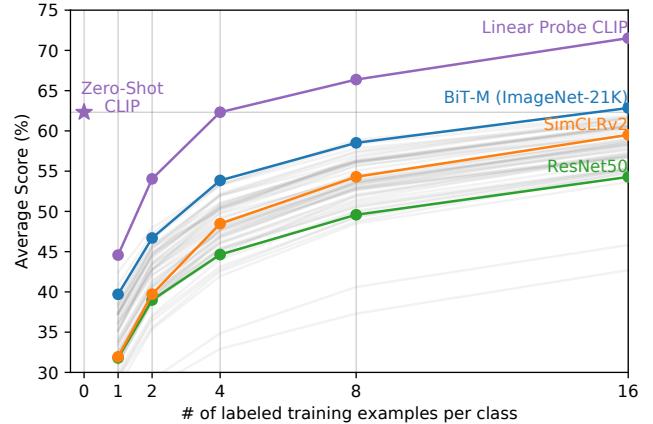


Figure 5. Zero-shot CLIP outperforms few-shot linear probes. Zero-shot CLIP matches the average performance of a 4-shot linear classifier trained on the same feature space and nearly matches the best results of a 16-shot linear classifier across publicly available models. For both BiT-M and SimCLRv2, 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.

the features of many image models including the best publicly available ImageNet models, self-supervised learning methods, and CLIP itself. While one might expect zero-shot to underperform one-shot, we instead find that zero-shot CLIP matches the performance of 4-shot logistic regression on the same feature space. This is likely due to a key difference between the zero-shot and few-shot approach. First, CLIP’s zero-shot classifier is generated via natural language which allows for visual concepts to be directly specified (“communicated”). By contrast, “normal” supervised learning must infer concepts indirectly from training examples. Context-less example-based learning has the drawback that many different hypotheses can be consistent with the data, especially in the one-shot case. A single image often contains many different visual concepts. Although a capable learner is able to exploit visual cues and heuristics, such as assuming that the concept being demonstrated is the primary object in an image, there is no guarantee.

When comparing zero-shot CLIP to few-shot logistic regression on the features of other models, zero-shot CLIP roughly matches the performance of the best performing 16-shot classifier in our evaluation suite, which uses the features of a BiT-M ResNet152x2 trained on ImageNet-21K.

We are certain that a BiT-L model trained on JFT-300M would perform even better but these models have not been publicly released. That a BiT-M ResNet152x2 performs best in a 16-shot setting is somewhat surprising since, as analyzed in Section 3.3, the Noisy Student EfficientNet-L2 outperforms it in a fully supervised setting by almost 5% on average across 27 datasets.

3.3. Representation Learning

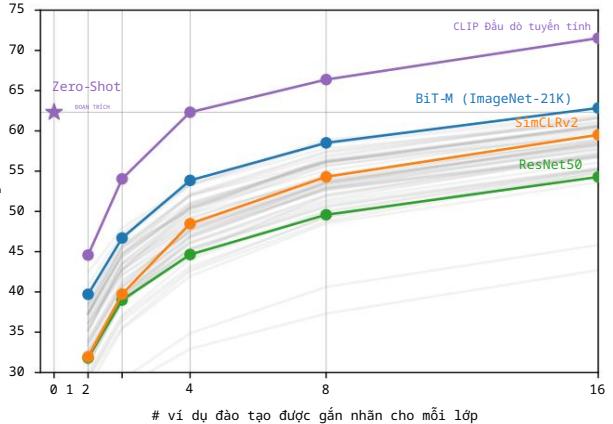
While we have focused on studying the task-learning capabilities of CLIP through zero-shot transfer, it is more common to study the representation learning capabilities of a model. We use a linear probe evaluation protocol because it requires minimal hyper-parameter tuning and has standardized evaluation procedures. Please see the supplementary material for further details on evaluation.

Figure 6 summarizes our findings. To minimize selection effects that could raise concerns of confirmation or reporting bias, we first study performance on the 12 dataset evaluation suite from Kornblith et al. (2019). Models trained with CLIP scale very well with compute and our largest model slightly outperforms the best existing model (a Noisy Student EfficientNet-L2) on both overall score and compute efficiency. We also find that CLIP vision transformers are about 3x more compute efficient than CLIP ResNets, which allows higher overall performance within our compute budget. These results replicate the findings of Dosovitskiy et al. (2020) which reported that vision transformers are more compute efficient than convnets when trained on sufficiently large datasets. Our best overall model ViT-L/14@336px outperforms the best existing model across this evaluation suite by an average of 2.6%.

CLIP models learn a wider set of tasks than has previously been demonstrated in a single computer vision model trained end-to-end from random initialization. These tasks include geo-localization, optical character recognition, facial emotion recognition, and action recognition. None of these tasks are measured in the evaluation suite of Kornblith et al. (2019). This could be argued to be a form of selection bias in Kornblith et al. (2019)’s study towards tasks that overlap with ImageNet. To address this, we also measure performance on a broader 27 dataset evaluation suite. This evaluation suite, detailed in Appendix A includes datasets representing the aforementioned tasks, German Traffic Signs Recognition Benchmark (Stallkamp et al., 2011), as well as several other datasets adapted from VTAB (Zhai et al., 2019). On this broader evaluation suite, the benefits of CLIP are more clear. All CLIP models, regardless of scale, outperform all evaluated systems in terms of compute efficiency. The improvement in average score of the best model over previous systems increases from 2.6% to 5%.

3.4. Robustness to Natural Distribution Shift

In 2015, it was announced that a deep learning model exceeded human performance on the ImageNet test set (He et al., 2015). However, research in the subsequent years has repeatedly found that these models still make many simple mistakes (Dodge & Karam, 2017; Geirhos et al., 2018; Alcorn et al., 2019), and new benchmarks testing these systems has often found their performance to be much lower than



Hình 5. Đầu dò CLIP không phát có hiệu suất tốt hơn đầu dò tuyến tính ít phát. Zero-shot CLIP khớp với hiệu suất trung bình của bộ phân loại tuyến tính 4-shot được đào tạo trên cùng một không gian đặc điểm và gần khớp với kết quả tốt nhất của bộ phân loại tuyến tính 16-shot trên các mô hình có sẵn công khai. Đối với cả BiT-M và SimCLRv2, mô hình có hiệu suất tốt nhất được đánh dấu. Các đường màu xám nhạt là các mô hình khác trong bộ đánh giá. 20 tập dữ liệu với ít nhất 16 ví dụ trên mỗi lớp đã được sử dụng trong phân tích này.

các tính năng của nhiều mô hình hình ảnh bao gồm các mô hình ImageNet tốt nhất có sẵn công khai, các phương pháp học tự giám sát và bản thân CLIP. Trong khi người ta có thể mong đợi zero-shot hoạt động kém hơn one-shot, thay vào đó chúng tôi thấy rằng zero-shot CLIP phù hợp với hiệu suất của hồi quy logistic 4-shot trên cùng một không gian tính năng. Điều này có thể là do sự khác biệt chính giữa phương pháp zero-shot và few-shot. Đầu tiên, bộ phân loại zero-shot của CLIP được tạo thông qua ngôn ngữ tự nhiên cho phép các khái niệm trực quan được chỉ định trực tiếp (“giao tiếp”). Ngược lại, học có giám sát “bình thường” phải suy ra các khái niệm gián tiếp từ các ví dụ đào tạo.

Học tập dựa trên ví dụ không có ngữ cảnh có nhược điểm là nhiều giả thuyết khác nhau có thể phù hợp với dữ liệu, đặc biệt là trong trường hợp một lần. Một hình ảnh duy nhất thường chứa nhiều khái niệm trực quan khác nhau. Mặc dù người học có năng lực có thể khai thác các tín hiệu trực quan và phương pháp tìm kiếm, chẳng hạn như giả định rằng khái niệm đang được chứng minh là đối tượng chính trong hình ảnh, nhưng không có gì đảm bảo.

Khi so sánh CLIP không có cú đánh nào với hồi quy logistic ít cú đánh trên các tính năng của các mô hình khác, CLIP không có cú đánh nào gần giống với hiệu suất của bộ phân loại 16 cú đánh tốt nhất trong bộ đánh giá của chúng tôi, sử dụng các tính năng của BiT-M ResNet152x2 được đào tạo trên ImageNet-21K. Chúng tôi chắc chắn rằng mô hình BiT-L được đào tạo trên JFT-300M sẽ hoạt động tốt hơn nữa nhưng những mô hình này chưa được công bố rộng rãi. Việc BiT-M ResNet152x2 hoạt động tốt nhất trong bài đặt 16 lần chụp là điều khá đáng ngạc nhiên vì, như đã phân tích trong Phần 3.3, Noisy Student EfficientNet-L2 hoạt động tốt hơn trong bài đặt được giám sát hoàn toàn trung bình gần 5% trên 27 tập dữ liệu.

3.3. Học biểu diễn

Trong khi chúng tôi tập trung vào việc nghiên cứu khả năng học nhiệm vụ của CLIP thông qua chuyển giao zero-shot, thì việc nghiên cứu khả năng học biểu diễn của một mô hình lại phổ biến hơn. Chúng tôi sử dụng giao thức đánh giá thăm dò tuyến tính vì nó yêu cầu điều chỉnh siêu tham số tối thiểu và có các quy trình đánh giá chuẩn hóa. Vui lòng xem tài liệu bổ sung để biết thêm chi tiết về đánh giá.

Hình 6 tóm tắt những phát hiện của chúng tôi. Để giảm thiểu các hiệu ứng lựa chọn có thể gây ra lỗi lo ngại về sai lệch xác nhận hoặc báo cáo, trước tiên chúng tôi nghiên cứu hiệu suất trên bộ đánh giá 12 tập dữ liệu từ Kornblith et al. (2019). Các mô hình được đào tạo bằng CLIP có khả năng mở rộng rất tốt với tính toán và mô hình lớn nhất của chúng tôi vượt trội hơn một chút so với mô hình hiện có tốt nhất (Noisy Student EfficientNet-L2) về cả điểm tổng thể và hiệu quả tính toán. Chúng tôi cũng thấy rằng bộ chuyển đổi thị giác CLIP có hiệu quả tính toán cao hơn khoảng 3 lần so với CLIP ResNet, cho phép hiệu suất tổng thể cao hơn trong phạm vi ngân sách tính toán của chúng tôi. Những kết quả này lập lại những phát hiện của Dosovitskiy et al. (2020), người đã báo cáo rằng bộ chuyển đổi thị giác có hiệu quả tính toán cao hơn convnet khi được đào tạo trên các tập dữ liệu đủ lớn. Mô hình tổng thể tốt nhất của chúng tôi ViT-L/14@336px vượt trội hơn mô hình hiện có tốt nhất trên toàn bộ bộ đánh giá này trung bình là 2.6%.

Các mô hình CLIP học một tập hợp rộng hơn các tác vụ đã được chứng minh trước đây trong một mô hình thị giác máy tính duy nhất được đào tạo từ đầu đến cuối từ khai tạo ngẫu nhiên. Các tác vụ này bao gồm định vị địa lý, nhận dạng ký tự quang học, nhận dạng cảm xúc khuôn mặt và nhận dạng hành động. Không có tác vụ nào trong số này được đo lường trong bộ đánh giá của Kornblith et al.

(2019). Có thể lập luận rằng đây là một dạng sai lệch lựa chọn trong nghiên cứu của Kornblith và cộng sự (2019) đối với các nhiệm vụ chồng chéo với ImageNet. Để giải quyết vấn đề này, chúng tôi cũng đo hiệu suất trên bộ đánh giá 27 tập dữ liệu rộng hơn. Bộ đánh giá này, được trình bày chi tiết trong Phụ lục A, bao gồm các tập dữ liệu đại diện cho các nhiệm vụ đã đề cập ở trên, German Traffic Signs Recognition Benchmark (Stallkamp và cộng sự, 2011), cũng như một số tập dữ liệu khác được điều chỉnh từ VTAB (Zhai và cộng sự, 2019). Trên bộ đánh giá rộng hơn này, lợi ích của CLIP rõ ràng hơn. Tất cả các mô hình CLIP, bắt kể quy mô, đều vượt trội hơn tất cả các hệ thống được đánh giá về hiệu quả tính toán.

Sự cải thiện về điểm trung bình của mô hình tốt nhất so với các hệ thống trước đó tăng từ 2,6% lên 5%.

3.4. Sự vững chắc đối với sự thay đổi phân phối tự nhiên

Vào năm 2015, người ta đã công bố rằng một mô hình học sâu đã vượt qua hiệu suất của con người trên bộ thử nghiệm ImageNet (He et al., 2015). Tuy nhiên, nghiên cứu trong những năm tiếp theo đã nhiều lần phát hiện ra rằng các mô hình này vẫn mắc nhiều lỗi đơn giản (Dodge & Karam, 2017; Geirhos et al., 2018; Alcorn et al., 2019) và các điểm chuẩn mới kiểm tra các hệ thống này thường thấy hiệu suất của chúng thấp hơn nhiều so với

Learning Transferable Visual Models From Natural Language Supervision

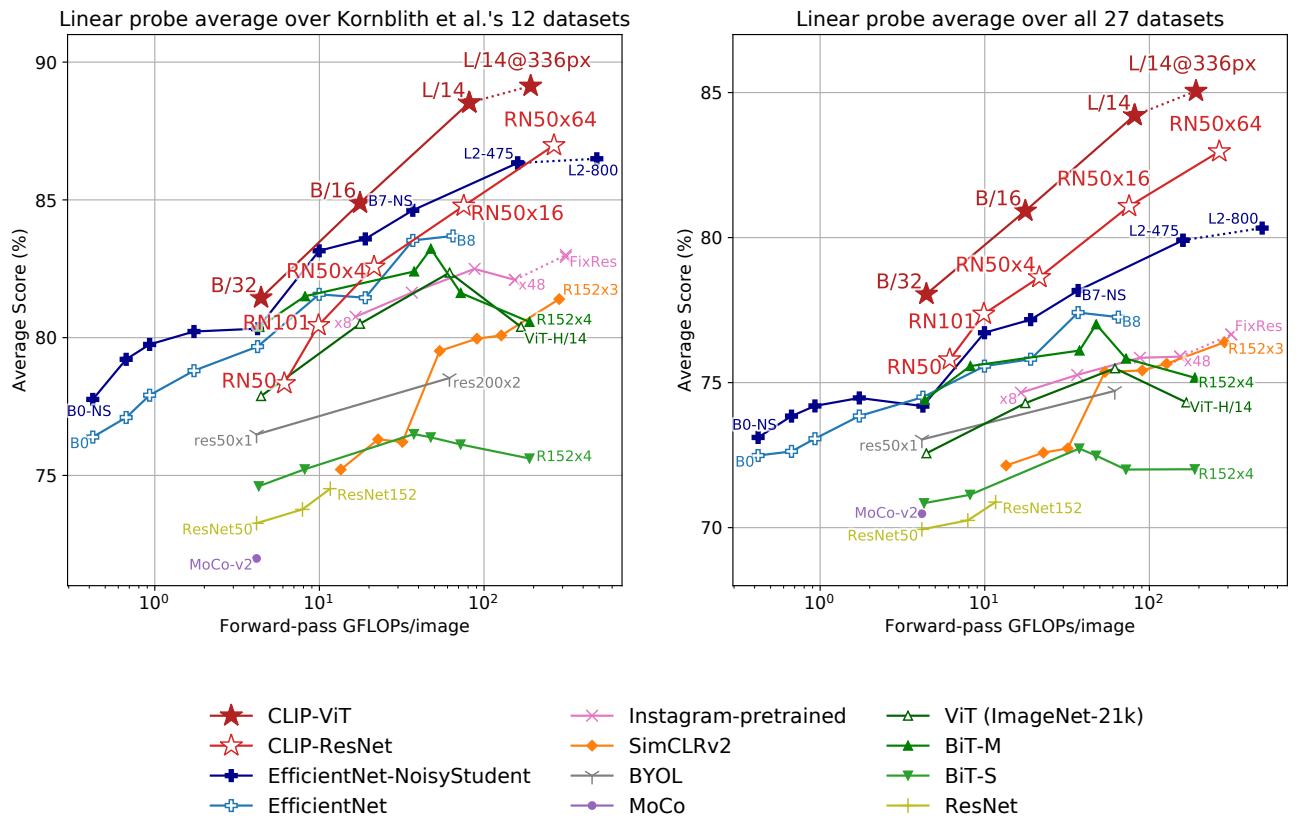


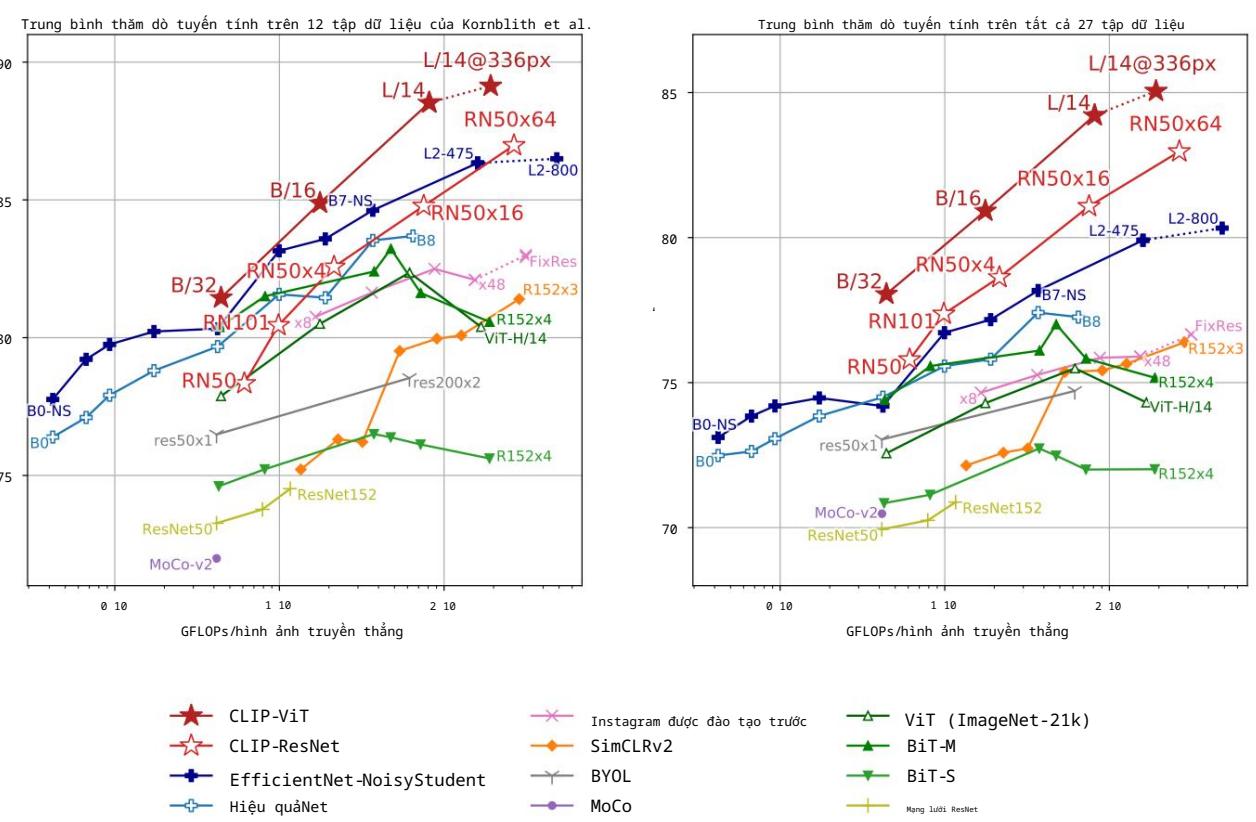
Figure 6. Linear probe performance of CLIP models in comparison with SOTA computer vision models, including EfficientNet (Tan & Le, 2019; Xie et al., 2020), MoCo (Chen et al., 2020b), Instagram-pretrained ResNeXt models (Mahajan et al., 2018; Touvron et al., 2019), BiT (Kolesnikov et al., 2019), ViT (Dosovitskiy et al., 2020), SimCLRv2 (Chen et al., 2020a), 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. Please see supplementary material for individual model scores for each dataset.

both human accuracy and ImageNet performance (Recht et al., 2019; Barbu et al., 2019). Taori et al. (2020) is a recent comprehensive study moving towards quantifying and understanding this for ImageNet models. Taori et al. (2020) study how the performance of ImageNet models change when evaluated on *natural distribution shifts*. They measure performance on a set of 7 distribution shifts. Taori et al. (2020) find that accuracy under distribution shift increases predictably with ImageNet accuracy and is well modeled as a linear function of logit-transformed accuracy. Taori et al. (2020) use this finding to propose that robustness analysis should distinguish between *effective* and *relative* robustness. Effective robustness measures improvements in accuracy under distribution shift above what is predicted by the documented relationship between in-distribution and out-of-distribution accuracy. Taori et al. (2020) argue that robustness techniques should aim to improve both effective robustness and relative robustness.

However, almost all models studied in Taori et al. (2020) are

trained or fine-tuned on the ImageNet dataset. Is training or adapting to the ImageNet dataset distribution the cause of the observed robustness gap? Intuitively, a zero-shot model should not be able to exploit spurious correlations or patterns that hold only on a specific distribution, since it is not trained on that distribution. Thus it is possible that zero-shot models exhibit higher effective robustness. In Figure 7, we compare the performance of zero-shot CLIP with existing ImageNet models on natural distribution shifts. All zero-shot CLIP models improve effective robustness by a large amount and reduce the gap between ImageNet accuracy and accuracy under distribution shift by up to 75%. Zero-shot CLIP models trace a completely distinct robustness frontier from all 204 prior models studied in Taori et al. (2020). These results suggest that the recent shift towards large-scale task and dataset agnostic pre-training combined with a reorientation towards zero-shot transfer evaluation (as advocated by Yogatama et al. (2019) and Linzen (2020)) promotes the development of more robust systems and provides a more accurate assessment of true model performance.

Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên



Hình 6. Hiệu suất thăm dò tuyến tính của các mô hình CLIP khi so sánh với các mô hình thị giác máy tính SOTA, bao gồm EfficientNet (Tan & Le, 2019; Xie và cộng sự, 2020), MoCo (Chen và cộng sự, 2020b), các mô hình ResNeXt được đào tạo trước trên Instagram (Mahajan và cộng sự, 2018; Touvron và cộng sự, 2019), BiT (Kolesnikov và cộng sự, 2019), ViT (Dosovitskiy và cộng sự, 2020), SimCLRv2 (Chen và cộng sự, 2020a), BYOL (Grill và cộng sự, 2020) và các mô hình ResNet ban đầu (He và cộng sự, 2016b). (Trái) Điểm được tính trung bình trên 12 tập dữ liệu do Kornblith và cộng sự nghiên cứu (2019). (Phải) Điểm được tính trung bình trên 27 tập dữ liệu có chứa nhiều loại phân phối hơn. Các đường chấm chấm biểu thị các mô hình được tinh chỉnh hoặc đánh giá trên hình ảnh có độ phân giải cao hơn so với đào tạo trước. Vui lòng xem tài liệu bổ sung để biết điểm mô hình riêng lẻ cho từng tập dữ liệu.

cả độ chính xác của con người và hiệu suất ImageNet (Recht và cộng sự, 2019; Barbu và cộng sự, 2019). Taori và cộng sự (2020) là một nghiên cứu toàn diện gần đây hướng tới việc định lượng và hiểu điều này cho các mô hình ImageNet. Taori và cộng sự (2020) nghiên cứu cách hiệu suất của các mô hình ImageNet thay đổi khi được đánh giá trên các ca chuyển dịch phân phối tự nhiên. Họ đo lường hiệu suất trên một tập hợp gồm 7 ca chuyển dịch phân phối. Taori và cộng sự (2020) nhận thấy rằng độ chính xác theo ca chuyển dịch phân phối tăng theo độ chính xác của ImageNet và được mô hình hóa tốt dưới dạng hàm tuyến tính của độ chính xác được chuyển đổi logit. Taori và cộng sự (2020) sử dụng phát hiện này để đề xuất rằng phân tích độ mạnh nén phân biệt giữa độ mạnh hiệu quả và độ mạnh tương đối. Độ mạnh hiệu quả đo lường sự cải thiện về độ chính xác theo ca chuyển dịch phân phối trên mức được dự đoán bởi mối quan hệ được ghi nhận giữa độ chính xác trong phân phối và ngoài phân phối. Độ mạnh tương đối nắm bắt mọi cải thiện về độ chính xác ngoài phân phối. Taori và cộng sự (2020) cho rằng các kỹ thuật tăng cường độ bền nên hướng tới mục tiêu cải thiện cả độ bền hiệu quả và độ bền tương đối.

Tuy nhiên, hầu hết tất cả các mô hình được nghiên cứu trong Taori et al. (2020) đều

được đào tạo hoặc tinh chỉnh trên tập dữ liệu ImageNet. Việc đào tạo hoặc thích ứng với phân phối tập dữ liệu ImageNet có phải là nguyên nhân gây ra khoảng cách độ mạnh quan sát được không? Theo trực giác, một mô hình zero-shot không thể khai thác các mối tương quan hoặc mẫu sai lệch chỉ tồn tại trên một phân phối cụ thể, vì nó không được đào tạo trên phân phối đó. Do đó, có thể các mô hình zero-shot thể hiện độ mạnh hiệu quả cao hơn. Trong Hình 7, chúng tôi so sánh hiệu suất của zero-shot CLIP với các mô hình ImageNet hiện có trên các dịch chuyển phân phối tự nhiên.

Tất cả các mô hình CLIP zero-shot đều cải thiện độ mạnh hiệu quả một cách đáng kể và thu hẹp khoảng cách giữa độ chính xác của ImageNet và độ chính xác khi dịch chuyển phân phối tới 75%. Các mô hình CLIP zero-shot theo dõi một ranh giới độ mạnh hoàn toàn khác biệt so với tất cả 204 mô hình trước đó được nghiên cứu trong Taori et al. (2020). Những kết quả này cho thấy sự thay đổi gần đây hướng tới tiền đào tạo không phụ thuộc vào nhiệm vụ và tập dữ liệu quy mô lớn kết hợp với định hướng lại theo hướng đánh giá chính xác hơn về hiệu suất mô hình thực sự.

Learning Transferable Visual Models From Natural Language Supervision

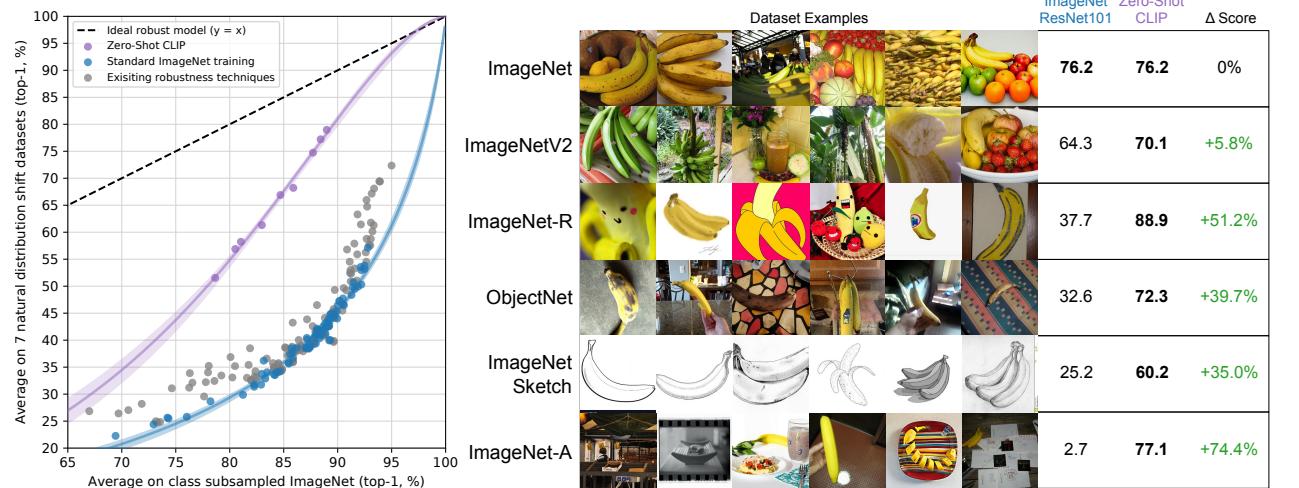


Figure 7. Zero-shot CLIP is much more robust to distribution shift than standard ImageNet models. (Left) An ideal robust model (dashed line) performs equally well on the ImageNet distribution and on other natural image distributions. Zero-shot CLIP models shrink this “robustness gap” by up to 75%. Linear 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 is compared with a model that has the same performance on the ImageNet validation set, ResNet101.

4. Data Overlap Analysis

A concern with pre-training on a very large internet dataset is unintentional overlap with downstream evals. We conducted de-duplication analysis to investigate this with full details in the supplementary material. Out of 35 datasets studied, 9 datasets have no detected overlap at all. There is a median overlap of 2.2% and an average overlap of 3.2%. Due to this small amount of overlap, overall accuracy is rarely shifted by more than 0.1% with only 7 datasets above this threshold. Of these, only 2 are statistically significant after Bonferroni correction. The max detected improvement is only 0.6% on Birdsnap. This echos the findings of similar duplicate analysis in previous work on large scale pre-training. Mahajan et al. (2018) and Kolesnikov et al. (2019) detected similar overlap rates for their models and also observed minimal changes in overall performance.

5. Broader Impacts

CLIP allows people to design their own classifiers and removes the need for task-specific training data. How these classes are designed heavily influences both model performance and model biases. For example, we find that when given a set of labels including Fairface race labels (Käkkäinen & Joo, 2019) and a handful of egregious terms such as “criminal” and “animal” the model tends to classify images of people aged 0–20 in the egregious category at a rate of 32.3%. However, when we add the class “child” to the list of possible classes, this behaviour drops to 8.7%. We also found discrepancies across gender and race for people categorized into the ‘crime’ and ‘non-human’ categories,

highlighting the potential for disparate impact even when extreme care is taken for thoughtful class design.

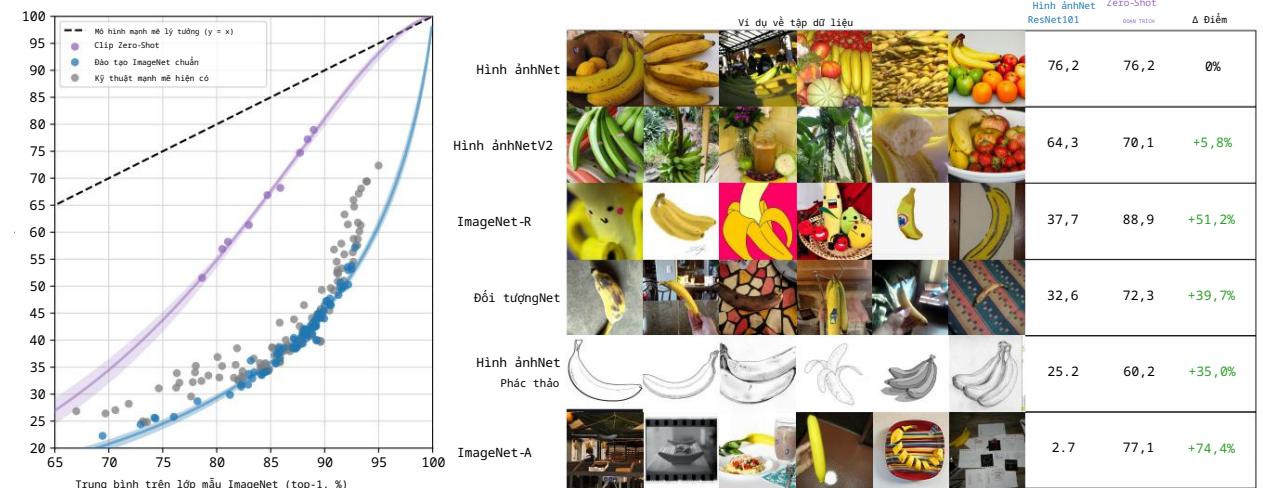
Additionally, given that CLIP does not need task-specific training data, it can unlock certain niche tasks with greater ease. Some of these tasks may raise privacy or surveillance related risks, which we explore by testing CLIP’s performance on celebrity identification using the CelebA dataset (Liu et al., 2018). CLIP has a top-1 accuracy of 59.2% for “in the wild” celebrity image classification when choosing from 100 candidates and of 43.3% when choosing from 1000 possible choices. Although it’s noteworthy to achieve these results with task agnostic pre-training, this performance is not competitive with widely available production level models. We explore challenges that CLIP poses in our supplemental materials and hope that this work motivates future research on the characterization of the capabilities, shortcomings, and biases of such models.

6. Limitations

The performance of zero-shot CLIP is often just competitive with the supervised baseline of a linear classifier on ResNet-50 features. This baseline is now well below the overall SOTA. Significant work is still needed to improve the task learning and transfer capabilities of CLIP. We estimate around a 1000x increase in compute is required for zero-shot CLIP to reach overall SOTA performance across our evaluation suite. This is infeasible to train with current hardware. Further research into improving upon the computational and data efficiency of CLIP will be necessary.

Despite our emphasis on zero-shot transfer, we repeatedly

Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên



Hình 7. Zero-shot CLIP mạnh mẽ hơn nhiều đối với sự thay đổi phân phối so với các mô hình mạnh mẽ lý tưởng (dường đứt nét). (Bên trái) Một mô hình mạnh mẽ lý tưởng (dường đứt nét) hoạt động tốt như nhau trên phân phối ImageNet và trên các phân phối hình ảnh tự nhiên khác. Các mô hình CLIP không có cảnh quay nào thu nhỏ “khoảng cách mạnh mẽ” này lên đến 75%. Các giá trị chuyển đổi logit phù hợp tuyến tính được hiển thị với khoảng tin cậy bootstrap ưới tinh 95%. (Bên phải) Hình dung sự thay đổi phân phối cho chuối, một lớp được chia sẻ trên 5 trong số 7 tập dữ liệu thay đổi phân phối tự nhiên. Hiệu suất của mô hình CLIP zero-shot tốt nhất được so sánh với mô hình có cùng hiệu suất trên bộ xác thực ImageNet, ResNet101.

4. Phân tích chồng chéo dữ liệu

Mỗi quan tâm với việc đào tạo trước trên một tập dữ liệu internet rất lớn là sự chồng chéo không chủ ý với các đánh giá hạ lưu. Chúng tôi đã tiến hành phân tích loại bỏ trùng lặp để điều tra điều này với đầy đủ chi tiết trong tài liệu bổ sung. Trong số 35 tập dữ liệu đã nghiên cứu, 9 tập dữ liệu không phát hiện thấy sự chồng chéo nào cả. Có sự chồng chéo trung bình là 2,2% và sự chồng chéo trung bình là 3,2%. Do lượng chồng chéo nhỏ này, độ chính xác tổng thể là hiếm khi thay đổi hơn 0,1% với chỉ 7 tập dữ liệu ở trên ngưỡng này. Trong số này, chỉ có 2 là có ý nghĩa thống kê sau khi hiệu chỉnh Bonferroni. Sự cải thiện tối đa được phát hiện chỉ là 0,6% trên Birdsnap. Điều này phản ánh những phát hiện của phân tích trùng lặp tương tự trong công trình trước đây trên quy mô lớn đào tạo trước. Mahajan và cộng sự. (2018) và Kolesnikov và cộng sự. (2019) đã phát hiện ra tỷ lệ chồng chéo tương tự cho các mô hình của họ và cũng quan sát thấy những thay đổi tối thiểu trong hiệu suất tổng thể.

5. Tác động rộng hơn

CLIP cho phép mọi người thiết kế bộ phân loại của riêng họ và loại bỏ nhu cầu về dữ liệu đào tạo cụ thể cho từng nhiệm vụ. Làm thế nào những các lớp được thiết kế ảnh hưởng rất lớn đến cả hiệu suất mô hình và độ lệch mô hình. Ví dụ, chúng tôi thấy rằng khi được cung cấp một bộ nhãn bao gồm nhãn chủng tộc Fairface (Käkkäinen & Joo, 2019) và một số thuật ngữ nghiêm trọng chẳng hạn như “tội phạm” và “động vật”, mô hình có xu hướng phân loại hình ảnh của những người trong độ tuổi từ 0-20 trong danh mục nghiêm trọng tại một tỷ lệ 32,3%. Tuy nhiên, khi chúng ta thêm lớp “child” vào danh sách các lớp có thể, hành vi này giảm xuống còn 8,7%. Chúng tôi cũng tìm thấy sự khác biệt giữa giới tính và chủng tộc đối với những người được phân loại vào nhóm “tội phạm” và “không phải con người”,

làm nổi bật tiềm năng tác động khác biệt ngay cả khi

thiết kế lớp học được thực hiện hết sức cẩn thận.

Ngoài ra, vì CLIP không cần nhiệm vụ cụ thể dữ liệu đào tạo, nó có thể mở khóa một số nhiệm vụ thích hợp với dễ dàng. Một số nhiệm vụ này có thể làm tăng sự riêng tư hoặc giám sát rủi ro liên quan, mà chúng tôi khám phá bằng cách kiểm tra hiệu suất của CLIP về nhận dạng người nổi tiếng bằng cách sử dụng tập dữ liệu CelebA (Liu và cộng sự, 2018). CLIP có độ chính xác hàng đầu là 59,2% đối với phân loại hình ảnh người nổi tiếng “trong tự nhiên” khi lựa chọn từ 100 ứng viên và 43,3% khi lựa chọn từ 1000 ứng viên có thể. Mặc dù đáng chú ý để đạt được những kết quả này với quá trình đào tạo trước không phụ thuộc vào nhiệm vụ, hiệu suất này không cạnh tranh được với sản xuất có sẵn rộng rãi mô hình cấp độ. Chúng tôi khám phá những thách thức mà CLIP đặt ra trong tài liệu bổ sung và hy vọng rằng công việc này thúc đẩy nghiên cứu trong tương lai về đặc điểm của các khả năng, những thiếu sót và thành kiến của các mô hình như vậy.

6. Hạn chế

Hiệu suất của CLIP zero-shot thường chỉ cạnh tranh với đường cơ sở có giám sát của bộ phân loại tuyến tính trên. Các tính năng của ResNet-50. Đường cơ sở này hiện đang ở mức thấp hơn nhiều SOTA nói chung. Vẫn cần phải làm việc đáng kể để cải thiện khả năng học và chuyển giao nhiệm vụ của CLIP. Chúng tôi ước tính cần tăng khoảng 1000 lần khả năng tính toán cho CLIP zero-shot để đạt được hiệu suất SOTA tổng thể trên bộ đánh giá của chúng tôi. Điều này không khả thi để đào tạo với phản ứng hiện tại. Nghiên cứu sâu hơn về việc cải thiện hiệu quả tính toán và dữ liệu của CLIP sẽ là cần thiết. Mặc dù chúng tôi nhấn mạnh vào việc chuyển giao không cần bán, chúng tôi liên tục

Learning Transferable Visual Models From Natural Language Supervision

queried performance on validation sets to guide development. This is unrealistic for true zero-shot scenarios. Similar concerns have been raised in the field of semi-supervised learning (Oliver et al., 2018). Another potential issue is our selection of evaluation datasets. While we report results on Kornblith et al. (2019)’s 12 dataset evaluation suite as a standardized collection, our main analysis uses a somewhat haphazard collection of 27 datasets that is undeniably co-adapted with the capabilities of CLIP. A new benchmark of tasks designed to evaluate broad zero-shot transfer capabilities would help address this issue.

We emphasize that specifying image classifiers through natural language is a flexible interface but this has its own limitations. Many complex tasks can be difficult to specify just through text. Actual training examples are undeniably useful but CLIP does not optimize for few-shot performance directly. We fall back to fitting linear classifiers on top of CLIP’s features. This results in a counter-intuitive drop in performance when transitioning from a zero-shot to a few-shot setting.

7. Related Work

The idea of learning to perform computer vision tasks from natural language supervision is by no means new. Rather, our main contribution is studying its behavior at large scale. Over 20 years ago Mori et al. (1999) explored improving content based image retrieval by training a model to predict the nouns and adjectives in text paired with images. Quattoni et al. (2007) demonstrated it was possible to learn more data efficient image representations via manifold learning in the weight space of classifiers trained to predict words in image captions. Srivastava & Salakhutdinov (2012) explored deep representation learning by training multimodal Deep Boltzmann Machines on top of low-level image and text tag features. More recent work inspiring CLIP is described in the Introduction.

Learning from collections of internet images is commonly investigated in webly supervised learning with Fergus et al. (2005) demonstrating the ability to train competitive computer vision classifiers by treating image search engine results as supervision. Of this line of work, *Learning Everything about Anything: Webly-Supervised Visual Concept Learning* (Divvala et al., 2014) has a notably similar ambition and goal as CLIP.

Developments in zero-shot computer vision (Larochelle et al., 2008; Lampert et al., 2009) were essential for CLIP. Socher et al. (2013a) demonstrated that connecting image and language representations enabled zero-shot transfer to unseen classes on CIFAR10 and Frome et al. (2013) improved and scaled this finding to ImageNet. The idea of generating a classifier from natural language dates back to

at least Elhoseiny et al. (2013) and a form similar to CLIP’s zero-shot classifier was explored in Lei Ba et al. (2015).

Natural language supervision has also been explored for tasks beyond image classification including video understanding (Ramanathan et al., 2013; Miech et al., 2019), Reinforcement Learning (Hermann et al., 2017), and a burst of recent work on learning joint models of vision and language (Lu et al., 2019; Tan & Bansal, 2019; Chen et al., 2019; Li et al., 2020b; Yu et al., 2020) for complex joint tasks beyond those studied here including visual question answering.

8. Conclusion

We have investigated whether it is possible to transfer the success of task-agnostic web-scale pre-training in NLP to another domain. We find that adopting this formula results in similar behaviors emerging in the field of computer vision and discuss the social implications of this line of research. In order to optimize their training objective, CLIP models learn to perform a wide variety of tasks during pre-training. This task learning can then be leveraged via natural language prompting to enable zero-shot transfer to many existing datasets. At sufficient scale, the performance of this approach can be competitive with task-specific supervised models although there is still room for much improvement.

ACKNOWLEDGMENTS

We’d like to thank the millions of people involved in creating the data CLIP is trained on. We’d also like to thank Susan Zhang for her work on image conditional language models while at OpenAI, Ishaaan Gulrajani for catching an error in the pseudocode, and Irene Solaiman, Miles Brundage, and Gillian Hadfield for their thoughtful feedback on the broader impacts section of the paper. We are also grateful to the Acceleration and Supercomputing teams at OpenAI for their critical work on software and hardware infrastructure this project used. Finally, we’d also like to thank the developers of the many software packages used throughout this project including, but not limited, to Numpy (Harris et al., 2020), SciPy (Virtanen et al., 2020), ftfy (Speer, 2019), TensorFlow (Abadi et al., 2016), PyTorch (Paszke et al., 2019), pandas (pandas development team, 2020), and scikit-learn (Pedregosa et al., 2011).

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283, 2016.

Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović,

Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên

hiệu suất được truy vấn trên các tập xác thực để hướng dẫn phát triển . Điều này là không thực tế đối với các kịch bản zero-shot thực sự. Những lo ngại tương tự đã được nêu ra trong lĩnh vực học bán giám sát (Oliver và cộng sự, 2018). Một vấn đề tiềm ẩn khác là lựa chọn các tập dữ liệu đánh giá của chúng tôi. Trong khi chúng tôi báo cáo kết quả về bộ đánh giá 12 tập dữ liệu của Kornblith và cộng sự (2019) dưới dạng một bộ sưu tập chuẩn hóa, thì phân tích chính của chúng tôi sử dụng một bộ sưu tập 27 tập dữ liệu có phần ngẫu nhiên, không thể phủ nhận là được điều chỉnh đồng thời với các khả năng của CLIP. Một chuẩn mực mới về các tác vụ được thiết kế để đánh giá các khả năng truyền zero-shot rộng sẽ giúp giải quyết vấn đề này.

Chúng tôi nhấn mạnh rằng việc chỉ định bộ phân loại hình ảnh thông qua ngôn ngữ tự nhiên là một giao diện linh hoạt nhưng điều này có những hạn chế riêng . Nhiều tác vụ phức tạp có thể khó chỉ định chỉ thông qua văn bản. Các ví dụ đào tạo thực tế chắc chắn hữu ích nhưng CLIP không tối ưu hóa trực tiếp cho hiệu suất chụp ít. Chúng tôi quay lại việc lắp bộ phân loại tuyến tính trên các tính năng của CLIP. Điều này dẫn đến hiệu suất giảm đi một cách phản trực giác khi chuyển từ cài đặt chụp không ảnh sang cài đặt chụp ít ảnh.

7. Công trình liên quan

Ý tưởng học cách thực hiện các tác vụ thị giác máy tính từ giám sát ngôn ngữ tự nhiên không phải là mới. Thay vào đó, đóng góp chính của chúng tôi là nghiên cứu hành vi của nó ở quy mô lớn.

Hơn 20 năm trước, Mori et al. (1999) đã khám phá cách cải thiện việc truy xuất hình ảnh dựa trên nội dung bằng cách đào tạo một mô hình để dự đoán danh từ và tính từ trong văn bản ghép nối với hình ảnh. Quattoni et al. (2007) đã chứng minh rằng có thể học được nhiều biểu diễn hình ảnh hiệu quả hơn về dữ liệu thông qua việc học đa tạp trong không gian trọng số của các bộ phân loại được đào tạo để dự đoán các từ trong chủ thích hình ảnh. Srivastava & Salakhutdinov (2012) đã khám phá cách học biểu diễn sâu bằng cách đào tạo Máy Boltzmann sâu đa phương thức trên các tính năng thé hình ảnh và văn bản cấp thấp. Các công trình gần đây hơn truyền cảm hứng cho CLIP được mô tả trong Phần giới thiệu.

Học từ các bộ sưu tập hình ảnh internet thường được nghiên cứu trong học tập có giám sát trên web với Fergus et al. (2005) chứng minh khả năng đào tạo các bộ phân loại thị giác máy tính cạnh tranh bằng cách xử lý kết quả của công cụ tìm kiếm hình ảnh như giám sát. Trong lĩnh vực này, Học mọi thứ về bất cứ thứ gì: Học khái niệm trực quan có giám sát trên web (Divvala et al., 2014) có tham vọng và mục tiêu tương tự đáng kể như CLIP.

Sự phát triển trong công nghệ thị giác máy tính không cần chụp (Larochelle và cộng sự, 2008; Lampert và cộng sự, 2009) là rất cần thiết cho CLIP. Socher et al. (2013a) đã chứng minh rằng việc kết nối các biểu diễn hình ảnh và ngôn ngữ cho phép chuyển giao không-shot sang các lớp chưa thấy trên CIFAR10 và Frome et al. (2013) đã cải thiện và mở rộng phát hiện này thành ImageNet. Ý tưởng tạo ra một bộ phân loại từ ngôn ngữ tự nhiên có từ

ít nhất là Elhoseiny et al. (2013) và một hình thức tương tự như bộ phân loại zero-shot của CLIP đã được khám phá trong Lei Ba et al. (2015).

Giám sát ngôn ngữ tự nhiên cũng đã được khám phá cho các nhiệm vụ vượt ra ngoài phân loại hình ảnh bao gồm hiểu video (Ramanathan và cộng sự, 2013; Miech và cộng sự, 2019), Học tăng cường (Hermann và cộng sự, 2017) và một loạt công trình gần đây về học các mô hình chung về thị giác và ngôn ngữ (Lu và cộng sự, 2019; Tan & Bansal, 2019; Chen và cộng sự, 2019; Li và cộng sự, 2020b; Yu và cộng sự, 2020) cho các nhiệm vụ chung phức tạp vượt ra ngoài những nhiệm vụ được nghiên cứu ở đây bao gồm trả lời câu hỏi trực quan.

8. Kết luận

Chúng tôi đã nghiên cứu xem liệu có thể chuyển giao thành công của tiền đào tạo web không phụ thuộc vào tác vụ trong NLP sang một miền khác hay không. Chúng tôi thấy rằng việc áp dụng công thức này dẫn đến các hành vi tương tự xuất hiện trong lĩnh vực thị giác máy tính và thảo luận về các tác động xã hội của hướng nghiên cứu này. Để tối ưu hóa mục tiêu đào tạo của mình, các mô hình CLIP học cách thực hiện nhiều loại tác vụ khác nhau trong quá trình đào tạo trước. Sau đó, việc học tác vụ này có thể được tận dụng thông qua lối nhắc ngôn ngữ tự nhiên để cho phép chuyển giao zero-shot sang nhiều tập dữ liệu hiện có. Ở quy mô lớn, hiệu suất của phương pháp tiếp cận này có thể cạnh tranh với các mô hình có giám sát theo tác vụ cụ thể mặc dù vẫn còn nhiều chỗ để cải thiện.

LỜI CẢM ƠN

Chúng tôi muốn cảm ơn hàng triệu người tham gia vào việc tạo dữ liệu mà CLIP được đào tạo. Chúng tôi cũng muốn cảm ơn Susan Zhang vì công trình của cô ấy về các mô hình ngôn ngữ có điều kiện hình ảnh khi còn làm việc tại OpenAI, Ishaaan Gulrajani vì đã phát hiện ra lỗi trong mã giả và Irene Solaiman, Miles Brundage và Gillian Hadfield vì đã phản hồi chi đáo về phản tác động rõ ràng của bài báo. Chúng tôi cũng biết ơn các nhóm Acceleration và Supercomputing tại OpenAI vì công trình quan trọng của họ về cơ sở hạ tầng phần mềm và phần cứng mà dự án này đã sử dụng. Cuối cùng, chúng tôi cũng muốn cảm ơn các nhà phát triển của nhiều gói phần mềm được sử dụng trong suốt dự án này bao gồm nhưng không giới hạn ở Numpy (Harris và cộng sự, 2020), SciPy (Virtanen và cộng sự, 2020), ftfy (Speer, 2019), TensorFlow (Abadi và cộng sự, 2016), PyTorch (Paszke và cộng sự, 2019), pandas (nhóm phát triển pandas, 2020) và scikit-learn (Pedregosa và cộng sự, 2011).

Tài liệu tham khảo

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., và cộng sự. TensorFlow: Một hệ thống cho máy học quy mô lớn. Trong hội thảo {USENIX} lần thứ 12 về thiết kế và triển khai hệ điều hành ({OSDI} 16), trang 265-283, 2016.

Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelovic,

Learning Transferable Visual Models From Natural Language Supervision

- R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., and Zisserman, A. Self-supervised multimodal versatile networks. *arXiv preprint arXiv:2006.16228*, 2020.
- Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4845–4854, 2019.
- Assiri, Y. Stochastic optimization of plain convolutional neural networks with simple methods. *arXiv preprint arXiv:2001.08856*, 2020.
- Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, pp. 9453–9463, 2019.
- Bechmann, A. and Bowker, G. C. Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, 6(1):205395171881956, January 2019. doi: 10.1177/2053951718819569. URL <https://doi.org/10.1177/2053951718819569>.
- Blaise Aguera y Arcas, M. M. and Todorov, A. Physiognomy's new clothes. 2017. URL <https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a>.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.
- Bowker, G. C. and Star, S. L. *Sorting things out: Classification and its consequences*. MIT press, 2000.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.
- Browne, S. *Dark Matters: Surveillance of Blackness*. Duke University Press, 2015.
- Bulent Sarıyıldız, M., Perez, J., and Larlus, D. Learning visual representations with caption annotations. *arXiv e-prints*, pp. arXiv–2008, 2020.
- Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.
- Carreira, J., Noland, E., Hillier, C., and Zisserman, A. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020a.
- Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.
- Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.
- Cheng, G., Han, J., and Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.
- Church, K. W. and Hanks, P. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990. URL <https://www.aclweb.org/anthology/J90-1003>.
- Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223, 2011.
- Crawford, K. The trouble with bias. *NIPS 2017 Keynote*, 2017. URL https://www.youtube.com/watch?v=fMym_BKWQzk.
- Dai, A. M. and Le, Q. V. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pp. 3079–3087, 2015.
- D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- Deng, J., Berg, A. C., Satheesh, S., Su, H., Khosla, A., and Fei-Fei, L. ILSVRC 2012, 2012. URL <http://www.image-net.org/challenges/LSVRC/2012/>.
- Desai, K. and Johnson, J. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020.
- R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., và Zisserman, A. Mạng đa phương thức tự giám sát . bản in trước arXiv arXiv:2006.16228, 2020.
- Alcorn, MA, Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., và Nguyen, A. Strike (with) a pose: Các mạng nơ-ron dễ bị đánh lừa bởi các tư thế lạ của các vật thể quen thuộc. Trong Biên bản báo cáo của Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, trang 4845–4854, 2019.
- Assiri, Y. Tối ưu hóa ngẫu nhiên của mạng nơ-ron tích chập thông thường bằng các phương pháp đơn giản. Bản in trước arXiv arXiv:2001.08856, 2020.
- Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., và Katz, B. Objectnet: Một tập dữ liệu có kiểm soát độ lệch quy mô lớn để đẩy giới hạn của các mô hình nhận dạng đối tượng. Trong *Advances in Neural Information Processing Systems*, trang 9453–9463, 2019.
- Bechmann, A. và Bowker, GC Không được giám sát bởi bất kỳ tên gọi nào khác: Các lớp ẩn của quá trình sản xuất kiến thức trong trí tuệ nhân tạo trên phương tiện truyền thông xã hội. *Big Data & Society*, 6(1):205395171881956, tháng 1 năm 2019. doi: 10.1177/2053951718819569. URL <https://doi.org/10.1177/2053951718819569>.
- Blaise Aguera y Arcas, MM và Todorov, Bộ quần áo mới của tương số học. MỘT. 2017. <https://medium.com/@blaisea/physignomys-new-clothes-f2d4b59fdd6a>.
- Bolukbasi, T., Chang, K.-W., Zou, JY, Saligrama, V., và Kalai, AT Đàm ông đối với lập trình viên máy tính cũng giống như phụ nữ đối với người nội trợ? nhưng từ ngữ khứ thiên vị. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 29:4349–4357, 2016.
- Bowker, GC và Star, SL Phân loại mọi thứ: Phân loại và hậu quả của nó. Nhà xuất bản MIT, 2000.
- Brown, TB, Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Các mô hình ngôn ngữ là những người học ít lần. Bản in trước arXiv arXiv:2005.14165, 2020.
- Browne, S. Dark Matters: Giám sát người da đen. Nhà xuất bản Đại học Duke , 2015.
- Bulent Sarıyıldız, M., Perez, J. và Larlus, D. Học cách biểu diễn trực quan với chủ thích chú thích. Bản in điện tử arXiv, trang arXiv-2008, 2020.
- Buolamwini, J. và Gebru, T. Sắc thái giới tính: Sự chênh lệch độ chính xác giữa các phần trong phân loại giới tính thương mại. Trong Hội nghị về công bằng, trách nhiệm giải trình và minh bạch, trang 77–91, 2018.
- Carreira, J., Noland, E., Hillier, C. và Zisserman, A. Một ghi chú ngắn về tập dữ liệu hành động của con người kinetics-700. Bản in trước arXiv arXiv:1907.06987, 2019.
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M. và Hinton, G. Các mô hình tự giám sát lớn là những người học bán giám sát mạnh. Bản in trước arXiv arXiv:2006.10029, 2020a.
- Chen, X., Fan, H., Girshick, R., và He, K. Cải thiện đường cơ sở với phương pháp học tương phản động lượng. Bản in trước arXiv arXiv:2003.04297, 2020b.
- Chen, Y.-C., Li, L., Yu, L., Kholy, AE, Ahmed, F., Gan, Z., Cheng, Y., và Liu, J. Uniter: Học cách biểu diễn ảnh ảnh-văn bản phổ quát. Bản in trước arXiv arXiv:1909.11740, 2019.
- Cheng, G., Han, J., và Lu, X. Phân loại cảnh ảnh viễn thám : Điều chỉnh và tình trạng nghệ thuật. Biên bản báo cáo của IEEE, 105(10):1865–1883, 2017.
- Church, KW và Hanks, P. Chuẩn mực liên kết từ, thông tin lẩn nhau và từ điển học. Ngôn ngữ học tính toán, 16(1):22–29, 1990. URL <https://www.aclweb.org/anthology/J90-1003>.
- Coates, A., Ng, A., và Lee, H. Phân tích mạng một lớp trong học tính năng không giám sát. Trong Biên bản báo cáo của hội nghị quốc tế lần thứ mười bốn về trí tuệ nhân tạo và thống kê, trang 215–223, 2011.
- Crawford, K. Rắc rối với sự thiên vị. Bài phát biểu quan trọng tại NIPS 2017, 2017. URL https://www.youtube.com/watch?v=fMym_BKWQzk.
- Dai, AM và Le, QV Học trình tự bán giám sát. Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 3079–3087, 2015.
- D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Ali-panahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., và Hoffman, M. D., et al. Công sự. Việc không xác định rõ đặt ra những thách thức về độ tin cậy trong học máy hiện đại. bản in trước arXiv arXiv:2011.03395, 2020.
- Dặng, J., Đông, W., Socher, R., Lý, L.-J., Lý, K. và Fei-Fei, L. ImageNet: Cơ sở dữ liệu hình ảnh phân cấp quy mô lớn. Trong CVPR09, 2009.
- Deng, J., Berg, AC, Satheesh, S., Su, H., Khosla, A., và Fei-Fei, L. ILSVRC 2012, 2012. URL <http://www.image-net.org/challenges/LSVRC/2012/>.
- Desai, K. và Johnson, J. Virtex: Học cách biểu diễn trực quan từ chủ nghĩa văn bản. Bản in trước arXiv arXiv:2006.06666, 2020.

Learning Transferable Visual Models From Natural Language Supervision

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Divvala, S. K., Farhadi, A., and Guestrin, C. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3270–3277, 2014.
- Dodge, S. and Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pp. 1–7. IEEE, 2017.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Elhoseiny, M., Saleh, B., and Elgammal, A. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2584–2591, 2013.
- Fergus, R., Fei-Fei, L., Perona, P., and Zisserman, A. Learning object categories from google’s image search. In *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*, volume 2, pp. 1816–1823. IEEE, 2005.
- Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.
- Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., and Liu, J. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020.
- Gao, T., Fisch, A., and Chen, D. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.
- Garvie, C., May 2019. URL <https://www.flawedfacedata.com/>.
- Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.
- Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.
- Google. Google cloud api: Celebrity recognition. URL <https://cloud.google.com/vision/docs/celebrity-recognition>.
- Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.
- Hays, J. and Efros, A. A. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2008.
- He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016b.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.
- Devlin, J., Chang, M.-W., Lee, K., và Toutanova, K. Bert: Đào tạo trước các bộ biến đổi song hướng sâu để hiểu ngôn ngữ. Bản in trước arXiv arXiv:1810.04805, 2018.
- Divvala, SK, Farhadi, A., và Guestrin, C. Học mọi thứ về bất kỳ thứ gì: Học khái niệm trực quan được giám sát bởi Webly. Trong Biên bản báo cáo của Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, trang 3270–3277, 2014.
- Dodge, S. và Karam, L. Nghiên cứu và so sánh hiệu suất nhận dạng của con người và học sâu dưới sự biến dạng thị giác. Năm 2017, hội nghị quốc tế lần thứ 26 về truyền thông máy tính và mạng (ICCCN), trang 1–7. IEEE, 2017.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. Một hình ảnh có giá trị bằng 16x16 từ: Transformers để nhận dạng hình ảnh theo tỷ lệ. Bản in trước arXiv arXiv:2010.11929, 2020.
- Elhoseiny, M., Saleh, B., và Elgammal, A. Viết một bộ phân loại : Học không cần thực hiện chỉ sử dụng các mô tả hoàn toàn bằng văn bản. Trong Biên bản Hội nghị quốc tế IEEE về Tâm nhìn máy tính, trang 2584–2591, 2013.
- Fergus, R., Fei-Fei, L., Perona, P., và Zisserman, A. Học các danh mục đối tượng từ tìm kiếm hình ảnh của Google. Trong Hội nghị quốc tế lần thứ mươi của IEEE về thị giác máy tính (ICCV’05) Tập 1, tập 2, trang 1816–1823. IEEE, 2005.
- Frome, A., Corrado, GS, Shlens, J., Bengio, S., Dean, J., Ranzato, M., và Mikolov, T. Devise: Một mô hình nhúng ngôn ngữ thị giác sâu. Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh , trang 2121–2129, 2013.
- Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., và Liu, J. Đào tạo đối nghịch quy mô lớn để học biểu diễn ngôn ngữ và thị giác. Bản in trước arXiv arXiv:2006.06195, 2020.
- Gao, T., Fisch, A., và Chen, D. Làm cho các mô hình ngôn ngữ được đào tạo trước trở nên tốt hơn đối với người học ít lần. Bản in trước arXiv arXiv:2012.15723, 2020.
- Garvie, C., tháng 5 năm 2019. URL <https://www.flawedfacedata.com/>.
- Geiger, A., Lenz, P., và Urtasun, R. Chúng ta đã sẵn sàng cho việc lái xe tự động chưa? Bộ công cụ chuẩn mực Kitti Vision. Trong Hội nghị về Thị giác máy tính và Nhận dạng mẫu (CVPR), 2012.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, FA, và Brendel, W. Các cnn được đào tạo bằng Imagenet là thiên về kết cấu; tăng độ lệch về hình dạng sẽ cải thiện độ chính xác và độ bền. Bản in trước arXiv arXiv:1811.12231, 2018.
- Goodfellow, IJ, Erhan, D., Carrier, PL, Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. Những thách thức trong việc học biểu diễn : Báo cáo về ba cuộc thi học máy. Mạng nơ-ron, 64:59–63, 2015.
- Google. Google cloud api: Nhận dạng người nổi tiếng. URL <https://cloud.google.com/vision/docs/celebrity-recognition>.
- Grill, J.-B., Strub, F., Altche, F., Tallec, C., Richemond, PH, Buchatskaya, E., Doersch, C., Pires, BA, Guo, ZD, Azar, MG, et al. Khởi động tiềm năng của riêng bạn: Một cách tiếp cận mới để tự học có giám sát. Bản in trước arXiv arXiv:2006.07733, 2020.
- Harris, CR, Millman, KJ, van der Walt, SJ, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, NJ, Kern, R., Picus, M., Hoyer, S., van Kerkwijk, MH, Brett, M., Haldane, A., Fernandez del Rio, J., Wiebe, M., Peterson, P., Gerard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., và Oliphant, lập trình TE Array với NumPy. Thiên nhiên, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.
- Hays, J. và Efros, AA Im2gps: ước tính thông tin địa lý từ một hình ảnh duy nhất. Trong hội nghị ieee năm 2008 về thị giác máy tính và nhận dạng mẫu, trang 1–8. IEEE, 2008.
- He, K., Zhang, X., Ren, S., và Sun, J. Đưa sâu vào bộ chính lưu: Vượt qua hiệu suất ở cấp độ con người về phân loại imangenet. Trong Biên bản báo cáo của hội nghị quốc tế IEEE về thị giác máy tính, trang 1026–1034, 2015.
- He, K., Zhang, X., Ren, S., và Sun, J. Học du sâu để nhận dạng hình ảnh. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 770–778, 2016a.
- He, K., Zhang, X., Ren, S., và Sun, J. Học du sâu để nhận dạng hình ảnh. Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 770–778, 2016b.
- He, K., Fan, H., Wu, Y., Xie, S., và Girshick, R. Tương phản động lượng cho việc học biểu diễn trực quan không giám sát . Trong Biên bản báo cáo Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 9729–9738, 2020.

Learning Transferable Visual Models From Natural Language Supervision

- He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 558–567, 2019.
- Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.
- Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pp. 4182–4192. PMLR, 2020.
- Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.
- Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, W. M., Jaderberg, M., Teplyashin, D., et al. Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*, 2017.
- Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.
- Hongsuck Seo, P., Weyand, T., Sim, J., and Han, B. Cplanet: Enhancing image geolocation by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 536–551, 2018.
- Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*, 2014.
- Jaderberg, M., Simonyan, K., Zisserman, A., et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.
- Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.
- Joulin, A., Van Der Maaten, L., Jabri, A., and Vasilache, N. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pp. 67–84. Springer, 2016.
- Kalfaoglu, M., Kalkan, S., and Alatan, A. A. Late temporal modeling in 3d cnn architectures with bert for action recognition. *arXiv preprint arXiv:2008.01232*, 2020.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Keyes, O. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018.
- Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020.
- Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Large scale learning of general visual representations for transfer. *arXiv preprint arXiv:1912.11370*, 2019.
- Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2661–2671, 2019.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.
- Kärkkäinen, K. and Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age, 2019.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people, 2016.
- Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 951–958. IEEE, 2009.
- Larochelle, H., Erhan, D., and Bengio, Y. Zero-data learning of new tasks. 2008.
- He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., và Li, M. Túi mẹo để phân loại hình ảnh bằng mạng nơ-ron tích chập. Trong Biên bản báo cáo của Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, trang 558-567 , 2019.
- Helber, P., Bischke, B., Dengel, A., và Borth, D. Eurosat: Một tập dữ liệu mới và chuẩn mực học sâu để phân loại sử dụng đất và lớp phủ đất. Tạp chí IEEE về các chủ đề được chọn trong quan sát Trái đất ứng dụng và cảm biến từ xa , 12(7):2217-2226, 2019.
- Henaff, O. Nhận dạng hình ảnh hiệu quả với mã hóa dự đoán tương phản. Trong Hội nghị quốc tế về học máy, trang 4182-4192. PMLR, 2020.
- Hendrycks, D. và Gimpel, K. Đơn vị tuyển tính lỗi Gauss (gelus). Bản in trước arXiv arXiv:1606.08415, 2016.
- Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., và Song, D. Các máy biến áp được đào tạo trước cải thiện độ bền ngoài phân phối. Bản in trước arXiv arXiv:2004.06100, 2020.
- Hermann, KM, Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., Szepesvari, D., Czarnecki, WM, Jaderberg, M., Teplyashin, D., et al. Học ngôn ngữ cơ bản trong thế giới mô phỏng 3 chiều. Bản in trước arXiv arXiv:1706.06551, 2017.
- Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., và Chu, Y. Việc mở rộng quy mô học sâu có thể dự đoán được theo kinh nghiệm. Bản in trước arXiv arXiv:1712.00409, 2017.
- Hongsuck Seo, P., Weyand, T., Sim, J., và Han, B. Cplanet: Tăng cường định vị địa lý hình ảnh bằng cách phân vùng kết hợp các bản đồ. Trong Biên bản báo cáo của Hội nghị châu Âu về thị giác máy tính (ECCV), trang 536-551, 2018.
- Howard, J. và Ruder, S. Điều chỉnh mô hình ngôn ngữ chung để phân loại văn bản. Bản in trước arXiv arXiv:1801.06146, 2018.
- Ioffe, S. và Szegedy, C. Chuẩn hóa theo lô: Tăng tốc đào tạo mạng sâu bằng cách giảm sự dịch chuyển biến phụ thuộc nội bộ. Bản in trước arXiv arXiv:1502.03167, 2015.
- Jaderberg, M., Simonyan, K., Vedaldi, A., và Zisserman, A. Học đầu ra có cấu trúc sâu để nhận dạng văn bản không bị hạn chế. Bản in trước arXiv arXiv:1412.5903, 2014.
- Jaderberg, M., Simonyan, K., Zisserman, A., et al. Mạng lưới biến áp không gian. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, 28:2017-2025, 2015.
- Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., và Girshick, R. Clevr: Một tập dữ liệu chẩn đoán cho ngôn ngữ sáng tác và tiêu học
- He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., và Li, M. suy luận trực quan. Trong Biên bản báo cáo Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, trang 2901-2910, 2017.
- Joulin, A., Van Der Maaten, L., Jabri, A. và Vasilache, N. Học các đặc điểm trực quan từ dữ liệu lớn được giám sát yếu. Trong Hội nghị Châu Âu về Tầm nhìn Máy tính, trang 67-84. Springer, 2016.
- Kalfaoglu, M., Kalkan, S. và Alatan, AA Mô hình hóa thời gian muộn trong kiến trúc cnn 3d với bert để nhận dạng hành động. Bản in trước arXiv arXiv:2008.01232, 2020.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, TB, Chess, B., Child, R., Gray, S., Radford, A., Wu, J. và Amodei, D. Luật mở rộng cho các mô hình ngôn ngữ thần kinh. Bản in trước arXiv arXiv:2001.08361, 2020.
- Keyes, O. Những cỗ máy phân biệt giới tính sai: Những hàm ý của Trans/hci về nhận dạng giới tính tự động. Biên bản báo cáo của ACM về Tương tác giữa Người và Máy tính, 2(CSCW):1-22, 2018.
- Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., và Testuggine, D. Thủ thách về meme thù hận : Phát hiện ngôn từ kích động thù địch trong meme đa phương thức. Bản in trước arXiv arXiv:2005.04790, 2020.
- Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., và Houlsby, N. Học tập quy mô lớn các biểu diễn trực quan chung để chuyển giao. Bản in trước arXiv arXiv:1912.11370, 2019.
- Kornblith, S., Shlens, J., và Le, QV Các mô hình imangenet tốt hơn có truyền tài tốt hơn không? Trong Biên bản báo cáo của hội nghị IEEE về thị giác máy tính và nhận dạng mẫu, trang 2661-2671, 2019.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, DA, et al. Bộ gen thị giác: Kết nối ngôn ngữ và thị giác bằng cách sử dụng chú thích hình ảnh dày đặc do cộng đồng đóng góp. Tạp chí quốc tế về thị giác máy tính, 123(1):32-73, 2017.
- Kärkkäinen, K. và Joo, J. Fairface: Bộ dữ liệu thuộc tính khuôn mặt cân bằng về chủng tộc, giới tính và độ tuổi, 2019.
- Lake, BM, Ullman, TD, Tenenbaum, JB và Gersh-man, SJ Xây dựng những cổ máy có khả năng học hỏi và suy nghĩ như con người, 2016.
- Lampert, CH, Nickisch, H., và Harmeling, S. Học cách phát hiện các lớp đối tượng chưa nhìn thấy bằng cách chuyển giao thuộc tính giữa các lớp . Trong Hội nghị IEEE năm 2009 về Thị giác máy tính và Nhận dạng mẫu, trang 951-958. IEEE, 2009.
- Larochelle, H., Erhan, D., và Bengio, Y. Học dữ liệu bằng không nhiệm vụ mới. 2008.

Learning Transferable Visual Models From Natural Language Supervision

- LeCun, Y. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>.
- Lei Ba, J., Swersky, K., Fidler, S., et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4247–4255, 2015.
- Li, A., Jabri, A., Joulin, A., and van der Maaten, L. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4183–4192, 2017.
- Li, G., Duan, N., Fang, Y., Gong, M., and Jiang, D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. 2020a.
- Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*, 2020b.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.
- Linzen, T. How can we accelerate progress towards human-like linguistic generalization? *arXiv preprint arXiv:2005.00955*, 2020.
- Lippe, P., Holla, N., Chandra, S., Rajamanickam, S., Antoniou, G., Shutova, E., and Yannakoudakis, H. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*, 2020.
- Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celeb-faces attributes (celeba) dataset. *Retrieved August, 15 (2018):11*, 2018.
- Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.
- Lu, Z., Xiong, X., Li, Y., Stroud, J., and Ross, D. Leveraging weakly supervised data and pose representation for action recognition, 2020. URL <https://www.youtube.com/watch?v=KOQFxbPPLOE&t=1390s>.
- Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 181–196, 2018.
- McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.
- Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE international conference on computer vision*, pp. 2630–2640, 2019.
- Miech, A., Alayrac, J.-B., Laptev, I., Sivic, J., and Zisserman, A. Rareact: A video dataset of unusual interactions. *arXiv preprint arXiv:2008.01018*, 2020a.
- Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., and Zisserman, A. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9879–9889, 2020b.
- Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. *arXiv preprint arXiv:2004.14444*, 2020.
- Mishra, A., Alahari, K., and Jawahar, C. Scene text recognition using higher order language priors. 2012.
- Mori, Y., Takahashi, H., and Oka, R. Image-to-word transformation based on dividing and vector quantizing images with words. Citeseer, 1999.
- Muller-Budack, E., Pustu-Iren, K., and Ewerth, R. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 563–579, 2018.
- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.
- Noble, S. U. Algorithms of oppression: How search engines reinforce racism. 2018.
- Nosek, B. A., Banaji, M. R., and Greenwald, A. G. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101, 2002.
- Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pp. 3153–3160. IEEE, 2011.
- Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31:3235–3246, 2018.
- LeCun, Y. Cơ sở dữ liệu mnist về chữ số viết tay. <http://yann.lecun.com/exdb/mnist/>.
- Lei Ba, J., Swersky, K., Fidler, S., et al. Dự đoán sâu mạng nơ-ron tích chập zero-shot sử dụng văn bản mô tả. Trong Biên bản của IEEE Quốc tế Hội nghị về Tầm nhìn máy tính, trang 4247-4255, 2015.
- Li, A., Jabri, A., Joulin, A. và van der Maaten, L. Learning n-gram trực quan từ dữ liệu web. Trong Biên bản báo cáo Hội nghị quốc tế IEEE về thị giác máy tính, trang. 4183-4192, 2017.
- Li, G., Duan, N., Fang, Y., Gong, M. và Jiang, D. Unicoder-vl: Một bộ mã hóa phổ quát cho thị giác và ngôn ngữ bằng cách đào tạo trước da phương thức. 2020a.
- Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Đào tạo trước theo ngữ nghĩa đối tượng cho các nhiệm vụ ngôn ngữ thị giác. Bản in trước của arXiv arXiv:2004.06165, 2020b.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., và Zitnick, CL Microsoft coco: Các đối tượng chung trong ngữ cảnh. Trong hội nghị châu Âu về tầm nhìn máy tính, trang 740-755. Springer, 2014.
- Linzen, T. Làm thế nào chúng ta có thể đẩy nhanh tiến độ hướng tới khái quát ngôn ngữ giống con người? Bản in trước arXiv arXiv:2005.00955, 2020.
- Lippe, P., Holla, N., Chandra, S., Rajamanickam, S., Antoniou, G., Shutova, E., và Yannakoudakis, H. Một khuôn khổ đa thời gian để phát hiện các meme đáng ghét. Bản in trước arXiv arXiv:2012.12871, 2020.
- Liu, Z., Luo, P., Wang, X., và Tang, X. Bộ dữ liệu thuộc tính khuôn mặt người nổi tiếng quy mô lớn (celeba). Truy cập ngày 15 tháng 8 (2018):11, 2018.
- Lu, J., Batra, D., Parikh, D., và Lee, S. Vilbert: Đào tạo trước biểu diễn ngôn ngữ thị giác không phụ thuộc vào nhiệm vụ cho các nhiệm vụ thị giác và ngôn ngữ. Trong Những tiên bộ trong thông tin thần kinh Hệ thống xử lý, trang 13-23, 2019.
- Lu, Z., Xiong, X., Li, Y., Stroud, J., và Ross, D. Đòn bẩy dữ liệu được giám sát yếu và đặt ra biểu diễn cho hành động sự công nhận, 2020. URL <https://www.youtube.com/watch?v=KOQFxbPPLOE&t=1390s>.
- Mahajan, D., Girshick, R., Ramanathan, V., Anh Áy, K., Paluri, M., Li, Y., Bharambe, A., và van der Maaten, L. Khám phá giới hạn của quá trình đào tạo trước được giám sát yếu. Trong Biên bản Hội nghị Châu Âu về Máy tính Tầm nhìn (ECCV), trang 181-196, 2018.
- McCann, B., Keskar, N. S., Xiong, C., và Socher, R. Cuộc thi mười môn phối hợp ngôn ngữ tự nhiên: Học đa nhiệm như một cách trả lời câu hỏi. Bản in trước arXiv arXiv:1806.08730, 2018.
- Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., và Sivic, J. Howto100m: Học một đoạn văn bản-video bằng cách xem hàng trăm triệu đoạn video có lời bình. Trong Biên bản hội nghị quốc tế IEEE về tầm nhìn máy tính, trang 2630-2640, 2019.
- Miech, A., Alayrac, J.-B., Laptev, I., Sivic, J. và Zisserman, A. Rareact: Bộ dữ liệu video về các tương tác bất thường. Bản in trước của arXiv arXiv:2008.01018, 2020a.
- Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., và Zisserman, A. Học tập toàn diện các biểu diễn trực quan từ các video hướng dẫn chưa được biên tập. Trong Biên bản báo cáo Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng Mẫu, trang 9879-9889, 2020b.
- Miller, G. A. Wordnet: cơ sở dữ liệu từ vựng cho tiếng Anh. Truyền thông của ACM, 38(11):39-41, 1995.
- Miller, J., Krauth, K., Recht, B., và Schmidt, L. Hiệu ứng của sự dịch chuyển phân phối tự nhiên trên các mô hình trả lời câu hỏi. Bản in trước arXiv arXiv:2004.14444, 2020.
- Mishra, A., Alahari, K., và Jawahar, C. Nhận dạng văn bản cảnh bằng cách sử dụng các tiên nghiệm ngôn ngữ bậc cao. 2012.
- Mori, Y., Takahashi, H., và Oka, R. Chuyển đổi hình ảnh thành từ dựa trên việc chia và lượng tử hóa hình ảnh bằng vectơ bằng lời nói. Citeseer, 1999.
- Muller-Budack, E., Pustu-Iren, K., và Ewerth, R. Ước tính vị trí địa lý của ảnh bằng mô hình phân cấp và phân loại cảnh. Trong Biên bản của Hội đồng châu Âu Hội nghị về Tầm nhìn máy tính (ECCV), trang 563-579, 2018.
- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., và Ng, A. Y. Đọc các chữ số trong hình ảnh tự nhiên với học tính năng không giám sát. 2011.
- Noble, S. U. Thuật toán áp bức: Công cụ tìm kiếm như thế nào cũng có chủ nghĩa phân biệt chủng tộc. 2018.
- Nosek, B. A., Banaji, M. R., và Greenwald, A. G. Thu thập thái độ và niềm tin ngầm của nhóm từ một trang web trình duyệt. Động lực nhóm: Lý thuyết, Nghiên cứu và Thực hành, 6(1):101, 2002.
- Ø, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., và những người khác. Một tập dữ liệu chuẩn quy mô lớn để nhận dạng sự kiện trong video giám sát. Trong CVPR 2011, trang 3153-3160. IEEE, 2011.
- Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., và Goodfellow, I. Đánh giá thực tế về bán giám sát sâu thuật toán học tập. Những tiên bộ trong hệ thống xử lý tin thần kinh, 31:3235-3246, 2018.

Learning Transferable Visual Models From Natural Language Supervision

pandas development team, T. pandas-dev/pandas: Pandas, February 2020. URL <https://doi.org/10.5281/zenodo.3509134>.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., and Sacheti, A. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

Quattoni, A., Collins, M., and Darrell, T. Learning visual representations using images with captions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2007.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. Saving face: Investigating the ethical concerns of facial recognition auditing, 2020.

Ramanathan, V., Liang, P., and Fei-Fei, L. Video event understanding using natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 905–912, 2013.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in neural information processing systems*, pp. 901–909, 2016.

Scheuerman, M. K., Paul, J. M., and Brubaker, J. R. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–33, 2019.

Schwemmer, C., Knight, C., Bello-Pardo, E. D., Oklobdzija, S., Schoonvelde, M., and Lockhart, J. W. Diagnosing gender bias in image recognition systems. *Socius*, 6: 2378023120967171, 2020.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pp. 935–943, 2013a.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013b.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pp. 1857–1865, 2016.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., and Wang, J. Release strategies and the social impacts of language models, 2019.

Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên

nhóm phát triển gấu trúc, T. pandas-dev/pandas: Pan-das, tháng 2 năm 2020. URL <https://doi.org/10.5281/zenodo.3509134>.

Parkhi, OM, Vedaldi, A., Zisserman, A., và Jawahar, CV Mèo và chó. Trong Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, 2012.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., và Chintala, S. Pytorch: Một thư viện deep learning có phong cách bắt buộc, hiệu suất cao. Ở Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., và Garnett, R. (eds.), *Những tiến bộ trong Hệ thống xử lý thông tin thần kinh 32*, trang. 8024–8035, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., và Duchesnay, E. Scikit-learn: Học máy trong Python. *Tạp chí nghiên cứu học máy*, 12:2825–2830, 2011.

Pennington, J., Socher, R., và Manning, CD Glove: Các vectơ toàn cục để biểu diễn từ. Trong Biên bản báo cáo của hội nghị năm 2014 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên (EMNLP), trang 1532–1543, 2014.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., và Zettlemoyer, L. Biểu diễn từ ngữ theo ngữ cảnh sâu sắc. *Bản in trước arXiv arXiv:1802.05365*, 2018.

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., và Sacheti, A. Imagebert: Đào tạo trước đa phương thức với dữ liệu hình ảnh-văn bản giám sát yêu quý mô lớn. *Bản in trước arXiv arXiv:2001.07966*, 2020.

Quattoni, A., Collins, M., và Darrell, T. Học các biểu diễn trực quan bằng hình ảnh có chủ thích. Trong Hội nghị IEEE năm 2007 về Thị giác máy tính và Nhận dạng mẫu, trang 1–8. IEEE, 2007.

Radford, A., Narasimhan, K., Salimans, T., và Sutskever, I. Cải thiện khả năng hiểu ngôn ngữ bằng phương pháp đào tạo trước tao ra, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., và Sutskever, I. Các mô hình ngôn ngữ là những người học đa nhiệm không có giám sát. 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., và Liu, P. J. Khám phá giới hạn của việc học chuyển giao với bộ chuyển đổi văn bản sang văn bản thống nhất. *Bản in trước arXiv arXiv:1910.10683*, 2019.

Raji, ID, Gebru, T., Mitchell, M., Buolamwini, J., Lee, J. và Denton, E. Giữ gìn thể diện: Điều tra các vấn đề đạo đức của việc kiểm toán nhận dạng khuôn mặt, 2020.

Ramanathan, V., Liang, P., và Fei-Fei, L. Hiểu sự kiện video bằng cách sử dụng mô tả ngôn ngữ tự nhiên. Trong Biên bản báo cáo của Hội nghị quốc tế IEEE về thị giác máy tính, trang 905–912, 2013.

Recht, B., Roelofs, R., Schmidt, L., và Shankar, V. Các bộ phân loại im-agenet có tổng quát hóa thành imagenet không? *Bản in trước arXiv arXiv:1902.10811*, 2019.

Salimans, T. và Kingma, DP Chuẩn hóa trọng số: Một tham số hóa lại đơn giản để tăng tốc quá trình đào tạo mạng nơ-ron sâu. Trong *Những tiến bộ trong hệ thống xử lý thông tin nơ-ron*, trang 901–909, 2016.

Scheuerman, MK, Paul, JM và Brubaker, JR Máy tính nhìn nhận giới tính như thế nào: Đánh giá phân loại giới tính trong các dịch vụ phân tích khuôn mặt thương mại. *Biên bản báo cáo của ACM về Tương tác giữa người và máy tính*, 3(CSCW): 1–33, 2019.

Schwemmer, C., Knight, C., Bello-Pardo, ED, Oklobdzija, S., Schoonvelde, M., và Lockhart, JW Chẩn đoán sai lệch giới tính trong hệ thống nhận dạng hình ảnh. *Socius*, 6: 2378023120967171, 2020.

Sennrich, R., Haddow, B., và Birch, A. Bản dịch máy thần kinh của các từ hiếm có đơn vị từ phụ. *Bản in trước arXiv arXiv:1508.07909*, 2015.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., và Rohrbach, M. Hướng tới các mô hình vqa có thể đọc. Trong Biên bản báo cáo của Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu, trang 8317–8326, 2019.

Socher, R., Ganjoo, M., Manning, CD, và Ng, A. Học tập không bắn thông qua chuyển giao đa phương thức. Trong *Những tiến bộ trong hệ thống xử lý thông tin thần kinh*, trang 935–943, 2013a.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., và Potts, C. Các mô hình sâu đẽ quy cho tính tổng hợp ngữ nghĩa trên một ngân hàng câu lệnh cảm. Trong Biên bản báo cáo của hội nghị năm 2013 về các phương pháp thực nghiệm trong xử lý ngôn ngữ tự nhiên, trang 1631–1642, 2013b.

Sohn, K. Cải thiện việc học số liệu sâu với mục tiêu mất mát n-pair đa lớp. Trong *Advances in neural information processing systems*, trang 1857–1865, 2016.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, JW, Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., và Wang, J. Chiến lược phát hành và tác động xã hội của các mô hình ngôn ngữ, 2019.

Learning Transferable Visual Models From Natural Language Supervision

- Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- Speer, R. ftfy. Zenodo, 2019. URL <https://doi.org/10.5281/zenodo.2591652>. Version 5.5.
- Srivastava, N. and Salakhutdinov, R. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012.
- Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *IEEE International Joint Conference on Neural Networks*, pp. 1453–1460, 2011.
- Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.
- Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.
- Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.
- Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.
- Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., and Isola, P. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.
- Touvron, H., Vedaldi, A., Douze, M., and Jegou, H. Fixing the train-test resolution discrepancy. In *Advances in neural information processing systems*, pp. 8252–8262, 2019.
- Varadarajan, J. and Odobe, J.-M. Topic models for scene analysis and abnormality detection. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 1338–1345. IEEE, 2009.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. Rotation equivariant CNNs for digital pathology. June 2018.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.
- Tan, H. và Bansal, M. Lxmert: Học biểu diễn bộ mã hóa đa phương thức từ bộ biến áp. Bản in trước arXiv arXiv:1908.07490, 2019.
- Vo, N., Jacobs, N., and Hays, J. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2621–2630, 2017.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
- Wang, H., Lu, P., Zhang, H., Yang, M., Bai, X., Xu, Y., He, M., Wang, Y., and Liu, W. All you need is boundary: Toward arbitrary-shaped text spotting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12160–12167, 2020.
- Weyand, T., Kostrikov, I., and Philbin, J. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pp. 37–55. Springer, 2016.
- Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., and Girshick, R. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.
- Yang, Z., Lu, Y., Wang, J., Yin, X., Florencio, D., Wang, L., Zhang, C., Zhang, L., and Luo, J. Tap: Text-aware pre-training for text-vqa and text-caption. *arXiv preprint arXiv:2012.04638*, 2020.
- Soomro, K., Zamir, AR và Shah, M. Ucf101: Bộ dữ liệu gồm 101 lớp hành động của con người từ các video ngoài tự nhiên. Bản in trước arXiv arXiv:1212.0402, 2012.
- Speer, R. ftfy. Zenodo, 2019. URL <https://doi.org/10.5281/zenodo.2591652>. Phiên bản 5.5.
- Srivastava, N. và Salakhutdinov, R. Học đa phương thức với máy boltzmann sâu. Trong *NIPS*, 2012.
- Stallkamp, J., Schlipsing, M., Salmen, J., và Igel, C. Tiêu chuẩn nhận dạng biến báo giao thông của Đức: Cuộc thi phân loại nhiều lớp. Trong Hội nghị chung quốc tế về mạng nơ-ron của IEEE, trang 1453–1460, 2011.
- Szegedy, C., Ioffe, S., Vanhoucke, V., và Alemi, A. Inception-v4, inception-resnet và tác động của các kết nối còn lại đến việc học. Bản in trước arXiv arXiv :1602.07261, 2016.
- Tan, H. và Bansal, M. Lxmert: Học biểu diễn bộ mã hóa đa phương thức từ bộ biến áp. Bản in trước arXiv arXiv:1908.07490, 2019.
- Tan, M. và Le, QV Efficientnet: Xem xét lại khả năng mở rộng mô hình cho mạng nơ-ron tích chập. Bản in trước arXiv arXiv:1905.11946, 2019.
- Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., và Schmidt, L. Đo lường độ mạnh mẽ đối với sự thay đổi phân phối tự nhiên trong phân loại hình ảnh. Bản in trước arXiv arXiv :2007.00644, 2020.
- Thomee, B., Shamma, DA, Friedland, G., Elizalde, B., Ni, K., Ba Lan, D., Borth, D., và Li, L.-J. Yfcc100m: Dữ liệu mới trong nghiên cứu đa phương tiện. Truyền thông của ACM, 59(2):64–73, 2016.
- Tian, Y., Wang, Y., Krishnan, D., và Isola, P. Mã hóa đa góc nhìn tương phản. Bản in trước arXiv arXiv:1906.05849, 2019.
- Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, JB và Isola, P. Suy nghĩ lại về phân loại hình ảnh ít ảnh: tất cả những gì bạn cần là những tốt? Bản in trước arXiv arXiv:2003.11539, 2020.
- Touvron, H., Vedaldi, A., Douze, M., và Jegou, H. Sửa lỗi độ phân giải giữa bài kiểm tra và bài tập. Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 8252–8262, 2019.
- Varadarajan, J. và Odobe, J.-M. Các mô hình chủ đề để phân tích cảnh và phát hiện bất thường. Trong Hội nghị quốc tế lần thứ 12 về Thị giác máy tính của IEEE năm 2009, Hội thảo ICCV, trang 1338–1345. IEEE, 2009.
- Yang, Z., Lu, Y., Wang, J., Yin, X., Florencio, D., Wang, L., Zhang, C., Zhang, L., và Luo, J. Tap: Đào tạo trước nhận biết văn bản cho text-vqa và text-caption. Bản in trước arXiv arXiv:2012.04638, 2020.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN, Kaiser, Ł., và Polosukhin, I. Chú ý là tất cả những gì bạn cần. Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, trang 5998–6008, 2017.
- Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., và Welling, M. CNN tương đương biến thể quay cho bệnh lý kỹ thuật số. Tháng 6 năm 2018.
- Virtanen, P., Gommers, R., Oliphant, TE, Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, SJ, Brett, M., Wilson, J., Millman, KJ, Mayorov, N., Nelson, ARJ, Jones, E., Kern, R., Larson, E., Carey, CJ, Polat, İ., Feng, Y., Moore, EW, VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriquez, I., Quintero, EA, Harris, CR, Archibald, AM, Ribeiro, AH, Pedregosa, F., van Mulbregt, P., và những người đóng góp SciPy 1.0. SciPy 1.0: Thuật toán cơ bản cho tính toán khoa học bằng Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.
- Vo, N., Jacobs, N., và Hays, J. Xem lại im2gps trong kỹ nguyên học sâu. Trong Biên bản báo cáo của Hội nghị quốc tế IEEE về thị giác máy tính, trang 2621–2630, 2017.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., và Bowman, SR Glue: Nền tảng phân tích và chuẩn mực đa tác vụ để hiểu ngôn ngữ tự nhiên. Bản in trước arXiv arXiv:1804.07461, 2018.
- Wang, H., Lu, P., Zhang, H., Yang, M., Bai, X., Xu, Y., He, M., Wang, Y., và Liu, W. Tất cả những gì bạn cần là ranh giới: Hướng tới việc phát hiện văn bản có hình dạng tùy ý. Trong Biên bản báo cáo Hội nghị về Trí tuệ nhân tạo AAAI, tập 34, trang 12160–12167, 2020.
- Weyand, T., Kostrikov, I., và Philbin, J. Định vị địa lý ảnh hành tinh bằng mạng nơ-ron tích chập. Trong Hội nghị Châu Âu về Tầm nhìn Máy tính, trang 37–55. Springer, 2016.
- Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., và Girshick, R. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- Xie, Q., Luong, M.-T., Hovy, E., và Le, QV Tự đào tạo với học sinh nhiều cải thiện phân loại imagenet. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng mẫu, trang 10687–10698, 2020.

Learning Transferable Visual Models From Natural Language Supervision

Yogatama, D., d'Autume, C. d. M., Connor, J., Kociský, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., et al. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*, 2019.

Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., and Wang, H. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Zhang, R. Making convolutional networks shift-invariant again. *arXiv preprint arXiv:1904.11486*, 2019.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

Zuboff, S. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1):75–89, 2015.

Học các mô hình trực quan có thể chuyển giao từ giám sát ngôn ngữ tự nhiên

Yogatama, D., d'Autume, C. d. M., Connor, J., Kociský, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., et al. Học tập và đánh giá chung trí thông minh ngôn ngữ. bản in trước arXiv arXiv:1901.11373, 2019.

Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., và Wang, H. Ernie-vil: Kiến thức nâng cao biểu diễn ngôn ngữ thị giác thông qua đồ thị cảnh. arXiv bản in trước arXiv:2006.16934, 2020.

Zeiler, MD và Fergus, R. Hình dung và hiểu các mạng lưới tích chập. Trong hội nghị châu Âu về tầm nhìn máy tính, trang 818-833. Springer, 2014.

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, AS, Neumann , M., Dosovitskiy, A., et al. Một nghiên cứu quy mô lớn về học biểu diễn với sự thích nghi nhiệm vụ trực quan chuẩn mực. bản in trước arXiv arXiv:1910.04867, 2019.

Zhang, R. Làm cho mạng lưới tích chập bắt biến dịch chuyển một lần nữa. Bản in trước arXiv arXiv:1904.11486, 2019.

Zhang, Y., Jiang, H., Miura, Y., Manning, CD và Langlotz, CP Học tương phản các biểu diễn trực quan y tế từ hình ảnh ghép và văn bản. Bản in trước arXiv arXiv:2010.00747, 2020.

Zuboff, S. Big khác: chủ nghĩa tư bản giám sát và triển vọng của nền văn minh thông tin. Tạp chí Công nghệ thông tin, 30(1):75-89, 2015.