Data-Efficient Language-Supervised Zero-Shot Learning with Self-Distillation

Ruizhe Cheng\textsuperscript{1}, Bichen Wu\textsuperscript{2}, Peizhao Zhang\textsuperscript{2}, Peter Vajda\textsuperscript{2}, Joseph E. Gonzalez\textsuperscript{1}
\textsuperscript{1}UC Berkeley, \textsuperscript{2}Facebook Reality Labs
\{chengruizhe,jegonzal\}@berkeley.edu, \{wbc,stzpz,vajdap\}@fb.com

Abstract

Traditional computer vision models are trained to predict a fixed set of predefined categories. Recently, natural language has been shown to be a broader and richer source of supervision that provides finer descriptions to visual concepts than supervised “gold” labels. Previous works, such as CLIP, use a simple pretraining task of predicting the pairings between images and text captions. CLIP, however, is data hungry and requires more than 400M image text pairs for training. We propose a data-efficient contrastive distillation method that uses soft labels to learn from noisy image-text pairs. Our model transfers knowledge from pre-trained image and sentence encoders and achieves strong performance with only 3M image text pairs, 133x smaller than CLIP. Our method exceeds the previous SoTA of general zero-shot learning on ImageNet 21k+1k by 73% relatively with a ResNet50 image encoder and DeCLUTR text encoder. We also beat CLIP by 10.5% relatively on zero-shot evaluation on Google Open Images (19,958 classes).

1. Introduction

In real-world image recognition tasks, input images can come from a broad range of distributions, spanning tens of thousands of object categories unknown during training. It is thus important for computer vision models to generalize to a large number of visual concepts that may or may not be present in the training data. This problem is called zero-shot learning (ZSL), which aims to transfer knowledge from some known classes with training data to a much larger number of unfamiliar classes. In this paper, we focus on the general zero-shot learning scenario where, at test time, the labels can be either seen or unseen classes.

Traditional ZSL methods mainly follow three paradigms. The first paradigm uses pretrained word embedding vectors to represent different categories and implicitly model their relationships. DeViSE\cite{frome2013devise} projects image features from a pretrained CNN and label’s word embeddings into a common embedding space. ConSE\cite{song2018conse} proposes a convex combination of the top k most likely image embeddings.

The second explicitly models class relationships as a graph, and use a graph convolutional network (GCN), or a predefined class hierarchy, such as WordNet\cite{miller1995wordnet}, to learn the knowledge propagation between classes. GCNZ\cite{zhao2019generalized} and DGPZ\cite{jiang2019decoupled} use a GCN to propagate knowledge into classifiers of unseen classes, while using CNN and word embeddings to encode image and label features. HZSL\cite{vinyals2016matching} projects image and text embeddings into a hyperbolic space that groups together child and parent classes in the WordNet\cite{miller1995wordnet} class hierarchy. Lastly, \cite{manmatha2016fine, chen2008transforming, ganea2018zero} rely on human-labeled attributes to model semantics of classes.

These works, however, have several drawbacks. First, they focus on finding a better mapping between image features extracted from pretrained CNNs and pretrained word embeddings such as GloVe\cite{pennington2014glove}. The image and text embeddings are not trained end-to-end jointly, limiting the generalization power and the quality of feature representations.
Second, predefined class hierarchies, such as WordNet[11], model categories in a tree structure, which fails to capture the complicated inter-class relationships present in real-world objects. Third, reliance on class hierarchies also limits the scope of classifiable objects to those present in the hierarchy. Fourth, methods that depend on attributes cannot generalize to categories that do not have known attributes.

In recent years, natural language has become a powerful source of supervision for image representation learning. [28] shows that pretraining by predicting hashtags on Instagram improves performance on ImageNet by over 5%. [8, 35, 44] all demonstrate the effectiveness of transformer-based language modeling in learning image representation from text. CLIP[32] and ALIGN[21] apply natural language supervision to the domain of ZSL. CLIP collects an enormous dataset with over 400M image caption pairs from the Internet, and trains an image encoder and a text encoder jointly with a contrastive loss to maximize the cosine similarity of corresponding image text embeddings and minimize those of others. CLIP demonstrates good zero-shot classification results on 27 downstream image classification datasets. However, neither CLIP nor ALIGN has published their image-caption datasets. It’s also an expensive and daunting task to collect, maintain and train vision models on datasets of that size.

We propose a data-efficient ZSL training pipeline that enables any pretrained image encoders to generalize to unseen classes. We initialize our model with an image encoder pretrained on ImageNet[7] 1k and a pretrained universal sentence encoder. We train our models on the public Conceptual Captions[36] dataset, which contains 3M loosely correlated image caption pairs. As seen in figure 1, there is considerable noise in the image-text pairings collected from the Internet. CLIP uses hard labels in the contrastive loss and account for the noise with a lot of data. Instead, we propose to use a hybrid of hard contrastive and soft distillation losses. We distill the model from its running Exponential Moving Average(EMA) with soft labels, as a method of denoising. Learning from soft labels enables better modelling of the rich correlations between vision and language and effectively account for cases where one caption matches objects in multiple images and vice versa. EMA is used as a continuous version of repeated self-distillation[18, 3].

With a ResNet50[17] image encoder and DeCLUTR[13] text encoder, we outperform the current SoTA of general ZSL on ImageNet 21k+1k by 73% relatively. In addition, we recognize issues with ImageNet21k and the 27 datasets used by CLIP[32] for ZSL evaluation in section 3.2.2. To bypass these problems, we propose using Google Open Images[24], which contains 19,958 categories, as a benchmark for zero-shot knowledge transfer to common visual concepts. Our model also exceeds CLIP on GOI by 10.5% relatively, while using a >100x smaller dataset.

2. Methods

Our model has a two-tower structure with an image encoder and a sentence encoder that outputs fixed-sized embeddings for a batch of corresponding images and captions. Different from previous ZSL works, our model assumes no class hierarchy. This makes our method more general, and easily extensible to datasets like Google Open Images[24].

2.1. Visual and Language Pretraining

Pretraining has become a crucial procedure in many NLP tasks[9, 5, 26]. Likewise, BiT[23] and ViT[10] has shown that transfer of pretrained visual representations leads to significant performance gains. Therefore, we initialize our model with an image encoder pretrained on ImageNet[7] 1k and a pretrained universal sentence encoder, such as Sentence Transformers[33] or DeCLUTR[13]. Sentence Transformers are pretrained on SNLI[4] and MultiNLI[41], whereas DeCLUTR is pretrained on the OpenWebText Corpus[14] or the Semantic Scholar Open Research Corpus[27].

2.2. Contrastive Learning

The contrastive learning[15] objective has been widely used in NLP and is at the core of several unsupervised[20, 43, 19] and self-supervised learning works[16, 6]. Similar to CLIP[32], we also use the contrastive loss, which measures the similarities of sample pairs in an embedding space. Specifically, we use the InfoNCE[39] loss where similarity is measured by dot product. Take a batch of \(N\) image and text pairs, the image and text encoders are jointly trained to maximize the cosine similarity of the \(N\) positive image and text pairings while minimizing the cosine similarity of the other \(N^2 - N\) negative image text pairings. In a batch of \(N\) image text pairs, let \(z_i^I\) be the embedding of the \(i\)th image, and \(z_j^T\) that of the \(j\)th text. The probability of the \(i\)th image matching the \(j\)th text is:

\[
P(z_i^I, z_j^T; \tau) = \frac{\exp(z_i^I \cdot z_j^T / \tau)}{\sum_{k=0}^{N} \exp(z_i^I \cdot z_k^T / \tau)} \quad (1)
\]

The InfoNCE loss for images is defined as:

\[
L_I = -\frac{1}{N} \sum_{i=0}^{N} \log P(z_i^I, z_i^T; \tau) \quad (2)
\]

We define the probability in (1) similarly for texts, and compute the InfoNCE loss symmetrically to get \(L_T\). The contrastive loss function thus becomes:

\[
L_{\text{InfoNCE}} = \frac{1}{2} (L_I + L_T) \quad (3)
\]
Table 1. Flat hit @k on Google Open Images. In the Model column, C means trained using contrastive loss only, and C+D means trained with contrastive and distillation loss jointly. * means that the model is a modified version.

| Dataset | Size | Model | Image Encoder | Text Encoder | Params | Flat Hit@k(%)  |
|---------|------|-------|---------------|--------------|--------|----------------|
|         |      |       |               |              |        | 1 2 5 10       |
| CLIP    | 400M | CLIP  | ResNet50      | Bert Base*    | 102M   | 26.5 38.3 54.0 64.3 |
| CLIP    | 400M | CLIP  | ViT-B/32      | Bert Base*    | 151M   | 27.5 39.5 55.3 65.4 |
| CC      | 3M   | C     | FBNet C       | DeCLUTR Sci Base | 114M   | 20.8 31.5 47.7 60.0 |
| CC      | 3M   | C     | EfficientNet B0 | DeCLUTR Sci Base | 114M   | 23.3 34.8 51.4 63.5 |
| CC      | 3M   | C     | ResNet50      | Sentence Bert Base | 134M   | 22.5 33.1 47.8 58.2 |
| CC      | 3M   | C     | ResNet50      | Bert Base     | 134M   | 24.6 35.4 50.0 60.2 |
| CC      | 3M   | C     | ResNet50      | DeCLUTR Sci Base | 135M   | 28.2 40.6 57.6 68.7 |
| CC      | 3M   | C+D   | ResNet50      | DeCLUTR Sci Base | 135M   | 29.3 42.0 58.6 69.4 |

2.3. EMA Self-Distillation

Image-text pairs collected from the Internet are usually only weakly correlated and noise is abundant. Often, images contain objects not mentioned in their captions, and captions contain words unrelated to their images. It’s also common for one caption to match objects in multiple images in a single batch. Hence, it’s not ideal to use hard labels as the only learning objective. We keep an Exponential Moving Average (EMA) of our model during training and use it as a continuously evolving teacher for self-distillation. We use a KL divergence loss to match the outputs of our model and its EMA teacher. According to equation (1), define $P_M$ and $P_{EMA}$ as the probability distribution of images over texts in a batch, for our model and its EMA teacher, respectively. Symmetrically, define $P_M^T$ and $P_{EMA}^T$.

$$L_{KL} = \frac{1}{2}[KL(P_M^T, P_{EMA}^T) + KL(P_M, P_{EMA})]$$ (4)

The final loss we use is:

$$L = L_{InfoNCE} + \alpha L_{KL}$$ (5)

where $\alpha$ is set to 1.0 in our experiments.

3. Experiments

3.1. Training

We apply a training schedule similar to the finetuning step of BiT[23]. We use SGD with an initial learning rate of 3e-3, a cosine annealing lr scheduler, momentum 0.9, and no weight decay. Input images are resized to 256x256 and random cropped to 224x224. We train the model on 4 GPUs using Pytorch[30] Distributed Data Parallel with a batch size of 128 per GPU for 30 epochs. While CLIP[32] computes the contrastive loss using only the batch on each GPU, we find that it’s important to all gather logits from the other GPUs and use them as negative samples.

3.2. Evaluation

During evaluation, we use a prompt template of “a photo of {label}” to augment the text labels of the target categories. We then compute the text embeddings of test categories with the trained text encoder, and fit a KNN using the embeddings. Given an image, we find the top k nearest neighbors of its embedding based on cosine similarity.

3.2.1 Evaluation Metric

The main metric we use for evaluating performance of ZSL is flat hit@k. Flat hit@k is the percentage of test images such that the top k predictions the model returns overlaps with any of the true labels. In ImageNet[7], each image is only labeled with one synset, but in Google Open Images[24], each image is labeled with multiple classes. The formal definition of flat hit@k is:

$$flat \text{hit@}k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{\{F(x_i)\}_k \cap L_i \neq \emptyset\}$$ (6)

where $\{F(x_i)\}_k$ is the top k predictions for the $i$th image and $L_i$ is the set of true labels.

3.2.2 Evaluation Dataset

We measure the ZSL performance mainly on Google Open Images [24]. And for backward compatibility to compare with prior work, we also report the results on ImageNet 21K+1K benchmark. We do not report results on the 27 datasets benchmark used by CLIP[32]. We discuss our considerations below.

**ImageNet 21K+1K:** Despite its popularity, there are four main problems of using ImageNet[7] for ZSL evaluation. First, based on the WordNet[11] structure, ImageNet has many repeated or trivially different classes. For example, "sunglass" and "sunglasses" are two different classes. Out of 22843 synsets, 1128 of them have names identical to at least another synset. Second, ImageNet labels don’t distinguish words with multiple meanings. For example, the word "crane" can mean either a type of bird or machine. Both classes are in ImageNet but have the same label. This happens for many words such as "ball". Third, each image in ImageNet is only labeled with exactly


| Dataset     | Size | Model       | Image Encoder | Text Encoder | Flat Hit@k(%) |
|-------------|------|-------------|---------------|--------------|---------------|
|             |      |             |               |              | 1  | 2  | 5  | 10 |
| ImageNet1k  | 1.2M | DeViSE      | ResNet50      | skip-gram    | 0.3 | 0.9 | 2.2 | 3.6 |
| ImageNet1k  | 1.2M | ConSE       | ResNet50      | skip-gram    | 0.1 | 1.5 | 3.5 | 4.9 |
| ImageNet1k  | 1.2M | GCNZ        | ResNet50      | GloVe        | 1.0 | 2.3 | 5.3 | 8.1 |
| ImageNet1k  | 1.2M | HZSL        | ResNet50      | GloVe*       | 2.2 | 4.6 | 9.2 | 12.7 |
| CC          | 3M   | C           | FBNet C[42]   | DeCLUTR Sci Base | 2.7 | 4.0 | 7.5 | 11.1 |
| CC          | 3M   | C           | EfficientNet B0[37] | DeCLUTR Sci Base | 3.0 | 4.6 | 8.4 | 12.2 |
| CC          | 3M   | C           | ResNet50      | Bert Base    | 3.2 | 5.7 | 10.5 | 15.3 |
| CC          | 3M   | C           | ResNet50      | Sentence Bert Base | 3.6 | 5.4 | 10.1 | 14.7 |
| CC          | 3M   | C           | ResNet50      | DeCLUTR Sci Base | 3.8 | 5.5 | 9.8 | 13.9 |
| CC          | 3M   | C+D         | ResNet50      | DeCLUTR Sci Base | 3.7 | 5.4 | 9.5 | 13.6 |
| CC          | 3M   | C           | ViT-Deit-B/16[38] | DeCLUTR Sci Base | 4.0 | 6.0 | 10.9 | 15.5 |
| CLIP        | 400M | CLIP        | ResNet50      | Bert Base*   | 13.5 | 19.7 | 30.5 | 39.4 |
| CLIP        | 400M | CLIP        | ViT-B/32      | Bert Base*   | 15.3 | 22.2 | 33.9 | 43.3 |

Table 2. Flat hit @k on ImageNet 21k+1k.

one class. When there are 2 or more visual concepts in the image, the model is forced to guess which object to classify. Fourth, ImageNet lacks the interactions between different visual concepts. About 90% of the images in ImageNet have only 1 distinct class, and almost no images have more than 4 distinct classes.

**Google Open Image:** Compared to ImageNet, Google Open Images[24] also contains a wide range of concepts, and it fixes all four problems outlined above. There are no repeated labels for different classes in GOI. Words with multiple meanings are also differentiated. For example, “crane” is labeled with “Crane (Machine)” and “Crane (Bird)”. More importantly, GOI labels each image with multiple classes, largely eliminating false negatives. In addition, GOI contains much more interactions between distinct classes per image, where more than 60% of images have 2 or more distinct classes. Inter-class interactions are especially useful in zero-shot learning, when we aim to transfer knowledge from seen to unseen classes.

**CLIP benchmark with 27 datasets:** CLIP[32] evaluates their model on 27 image classification datasets. However, many of these datasets are domain specific, such as Stanford Cars and FGVC aircraft, which have specific models of cars or planes as categories. This makes evaluation on them a test of knowledge memorization, rather than generalization. Similar to ImageNet, very few of these datasets contain multiple distinct classes in the same image, reflecting a lack of visual richness. Lastly, with only 3896 total categories, the 27 datasets altogether don’t cover nearly as many common visual concepts as GOI.

### 3.3. Results on Google Open Images

We evaluate the models on the test set of Google Open Images V6[24], with 125,436 images. Traditional ZSL baselines aren’t evaluated on GOI due to the lack of a class structure. In table 1, we compare the flat hit@k of our models with pretrained CLIP[32]. Our ResNet50 and DeCLUTR Sci Base model trained with the joint contrastive and distillation loss exceeds CLIP ResNet50 and Bert[9] by 10.5% relatively in FH@k=1, while being > 100x more data efficient.

### 3.4. Results on ImageNet 21k+1k

In this section, we present flat hit@k results on zero-shot transfer to the ImageNet 21k+1k[7] dataset, which contains 21841 classes in total. The image encoders are initialized with weights pretrained on ImageNet 1k. Sentence Bert[33] is pretrained on SNLI[4] and MultiNLI[41], while Declutr Sci Base[13] is pretrained on the S2ORC[27].

Many traditional ZSL methods rely on a predefined class hierarchy for explicit knowledge propagation. ImageNet, whose classes are a subset of WordNet, becomes the ideal benchmark for these works. With 400M image text pairs, CLIP[32] vastly outperforms previous methods. Our method uses Conceptual Captions[36] 3M, which is on the same order of magnitude as ImageNet 1k, and outperforms the previous SoTA, HZSL[25], by 73% relatively. In table 2, we demonstrate good performance on a variety of image and sentence encoder architectures. The gap between our method and CLIP may be caused by the fact that ImageNet classes contain many uncommon words, such as scientific names of animals or medical terms. CLIP’s dataset is much larger and thus covers much more uncommon words. EMA distillation also slightly decreases the performance compared to using only contrastive loss. We hypothesize that this is because ImageNet only has one "gold" label per image during evaluation. However, EMA distillation encourages the model to output a softer probability output for multiple classes, which can be present but just not labeled.
Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S. Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *ICLR*, 2014.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. *EMNLP*, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.

Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. *ICML*, 2015.

Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. *arXiv preprint arXiv:2008.01392*, 2020.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *ICML*, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1911.05722*, 2018.

Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. *CVPR*, 2018.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018.