Data Efficient Language-Supervised Zero-Shot Recognition with Optimal Transport Distillation

Ruizhe Cheng

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2021-58
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-58.html

May 12, 2021

Data Efficient Language-Supervised Zero-Shot Recognition with Optimal Transport Distillation

by Ruizhe Cheng

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Joseph E. Gonzalez

Professor Joseph E. Gonzalez
Research Advisor

5/11/21

(Date)

* * * * * * *

Kurt Keutzer

Professor Kurt Keutzer
Second Reader

May 6, 2021

(Date)

Acknowledgements

I would like to thank my advisor, Professor Joseph E. Gonzalez, for his extensive guidance and advice on both my research and career. I would also like to especially thank Dr. Bichen Wu for being an exceptional research mentor. Bichen not only has invaluable technical insights, but has also taught me how to define, investigate and solve research problems over the past three years. Additionally, I want to thank Professor Kurt Keutzer for his feedback on this project. Finally, I want to thank my mother, Peihua Zhang, my father, Hongqing Cheng, and my grandparents for their unconditional love and support.

Data Efficient Language-Supervised Zero-Shot Recognition with Optimal Transport Distillation

Ruizhe Cheng
UC Berkeley
chengruizhe@berkeley.edu

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 contrastive learning task to predict the pairings between images and text captions. CLIP, however, is data hungry and requires more than 400M image-text pairs for training. The inefficiency can be partially attributed to the fact that the image-text pairs are noisy. To mitigate this, we propose to use online entropic optimal transport to find a better image-text matching and using the matching as a soft training label for contrastive learning. Our model transfers knowledge from pretrained image and sentence encoders and achieves strong performance with only 3M image text pairs, 133x smaller than CLIP. We beat CLIP by 14% relatively on zero-shot evaluation on Google Open Images (19,958 classes). Our method also 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.

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.

Many works[46, 2, 1] in ZSL have focused on using attributes of unseen classes for knoweledge propagation. These work are limited in scope and application to real-world datasets due to their reliance on human-labeled attributes. Other traditional ZSL methods[20, 41] use the implicit image and text/word representations from pretrained models and learn a mapping into a common embedding space. More recent works[53, 32] have used graph convolutional networks and information from predefined class hierarchies, such as WordNet[19], to model inter-class relationships.

More recently, natural language has become a powerful source of supervision for image representation learning. [39] shows that pretraining by predicting hashtags on Instagram improves performance on ImageNet by over 5%. [15, 49, 60, 30] all demonstrate the effectiveness of transformer-based language modeling in learning image representation from text. CLIP [44] has applied natural language supervision to the domain of ZSL. It 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 paired image and text embeddings and minimize the similarity of unpaired ones. CLIP demonstrates good zero-shot classification results on a wide range of downstream image classification datasets. However, one main constraint of CLIP is that it is data hungry and requires over 400M image-text pairs for training. Collecting and training on such a huge

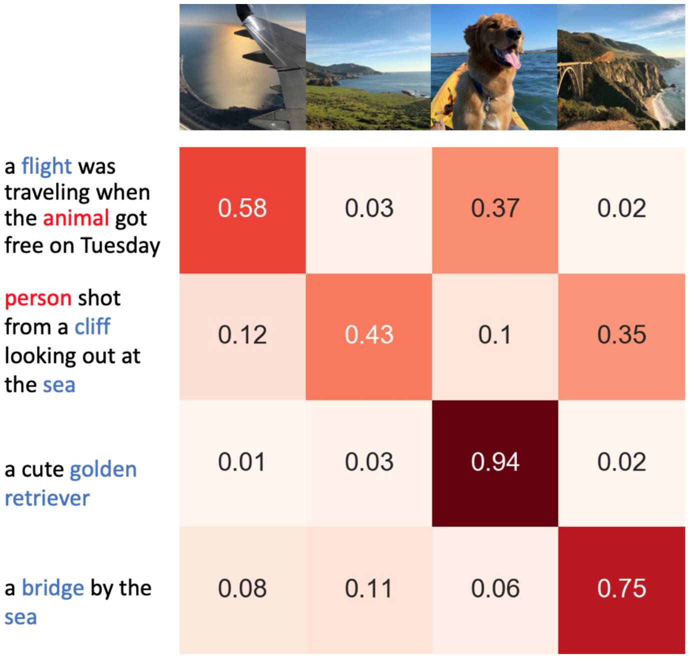

Figure 1: Caption and image pairings are noisy. Images may contain objects not mentioned in the caption, and captions have words not related to the image (colored red). There is a many-to-many relationship between a batch of images and captions, which is better modeled by soft probabilities than hard labels. We use optimal transport to compute soft labels and distillation from them to mitigate this noise. This enables us to achieve good performance with high data efficiency.

dataset is very expensive. The inefficiency can be partially attributed to the fact that the training signals from image-text pairs are noisy. As shown in Figure 1, in most of the datasets, we observe that images and captions are only loosely correlated. It is very common that one caption (image) can potentially match several images (captions), and the "ground-truth" pairings are not the only sensible, and sometimes not the optimal matchings between images and text captions. Despite this, CLIP uses the InfoNCE loss [23] to train the image and text embeddings, treating the the ground-truth pairings as hard labels. This ignores the many-to-many relationships between images and text captions, and leads to inefficiency.

To improve data efficiency and mitigate data noise, we propose a data-efficient ZSL training pipeline that enables any pretrained image encoders to generalize to unseen classes. We recognize the fact that there is considerable noise in the image-text pairings collected from the Internet. Whereas CLIP uses hard labels in the contrastive loss, we use a hybrid of hard contrastive and soft distillation losses. Furthermore, we propose using optimal transport as a natural solution to combat batch-level data noise under the contrastive learning setting. We use optimal transport to find the optimal coupling between a batch of image-text pairs, and use this soft coupling as the target for distillation. 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. We initialize our model with an image encoder pretrained on ImageNet[14] 1k and a pretrained text encoder. Then, we train our models on the public Conceptual Captions[50] dataset, which contains 3M loosely correlated image caption pairs. This framework significantly improves performance in zero-shot learning and is easily extensible to other domains such as contrastive self-supervised learning. Different from many other works using optimal transport, we use the optimal matching as labels for knowledge distillation, rather than directlying optimzing the Wasserstein loss.

With a ResNet50[25] image encoder and DeCLUTR[21] 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[44] for ZSL evaluation in section 4.3.2. To bypass these problems, we propose using Google Open Images[34], which contains 19,958 categories, as a benchmark for zero-shot knowledge transfer to common visual concepts. Our model exceeds CLIP on GOI by 14% relatively, while using a >100x fewer image-text pairs.

2 Related Works

2.1 Zero-Shot Learning

Zero-shot learning(ZSL) studies the generalization of knowledge to 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[20] projects image features from a pretrained CNN and word embeddings of labels into a common embedding space. ConSE[41] proposes a convex combination of the top-K most likely image embeddings. The second paradigm explicitly models class relationships as a graph, and use a graph convolutional network (GCN), or a predefined class hierarchy, such as WordNet[19], to learn the knowledge propagation between classes. GCNZ[53] and DGPZ[32] use a GCN to propagate knowledge into classifiers of unseen classes, while using a CNN and word embeddings to encode image and label features. HZSL[36] projects image and text embeddings into a hyperbolic space that groups together child and parent classes in the WordNet[19] class hierarchy. Lastly, [46, 2, 1] 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 pretrined CNNs and pretrained word embeddings such as GloVe[43]. 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[19], 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.

More recently, CLIP[44] applies large-scale language-supervision to ZSL by using over 400M image caption pairs collected from the Internet. CLIP trains an image encoder and a text encoder jointly with a contrastive loss to maximize the cosine similarity of paired image and text embeddings and minimize that of unpaired ones. However, CLIP has not published their image-caption dataset. It's also an expensive and daunting task to collect, maintain and train vision models on datasets of that size.

2.2 Optimal Transport

Optimal transport(OT) is a theory that enables comparison of two probability distributions whose supports may not overlap. We follow the definition of optimal transport in [13]. Let μ and ν be two probability measures defined on spaces X and Y, respectively. Define a cost function $c(x, y) : X \times Y \rightarrow [0, \infty]$ that measures the cost of transporting one unit of mass from $x \in X$ to $y \in Y$. Optimal Transport solves how to transport μ to ν while minimizing the cost c. In the discrete setting, optimal transport solves for the optimal strategy $T \in \mathbb{R}^{n_1 \times n_2}$ in the space of joint distributions $\Pi(\mu, \nu)$ that minimizes the Wasserstein loss:

$$W_c(\mu, \nu) = \min_{T \in \Pi(\mu,\nu)} \langle T\,, C\rangle_F \tag{1}$$

where $\langle \cdot\,, \cdot\rangle_F$ is the Frobenius dot product, $C \in \mathbb{R}^{n_1 \times n_2}$ is the cost matrix where $C_{ij} = c(x_i, y_j)$ and n_1, n_2 the size of the supports for μ and ν.

Recently, OT has been applied to many areas such as domain adaptation[11], and generative models[48]. [13] uses entropy regularized optimal transport to mitigate label noise in supervised learning. [7] applies OT to cross-domain alignment. It models the objects in an image and the words in a sentence as nodes in graphs, and tackles the problem of object and word alignment in a single image-text pair. [8] uses OT in local contrastive knowledge distillation, where it directly minimizes the Wasserstein loss between student and teacher embeddings in a batch. This, however, leads to mode collapse in the multi-modal setting, when both image and text encoders are end-to-end trainable. Instead, we keep an Exponential Moving Average(EMA) of the model, and calculate the optimal OT coupling from the outputs of the EMA model.

2.3 Knowledge Distillation

[26] first proposes knowledge distillation as a compression technique for deep neural networks, by matching the output logits of a teacher and a student model. It has then been applied to a vast number of different domains[47, 59, 54, 31, 28]. Distillation has long been used to address noise in data. [35] combines distillation with a label knowledge graph to learn from noisy data. [58, 3] repeatedly distill a student model with generated pseudo labels and show improved supervised learning performance. Recently, distillation has been also applied to self-supervised learning(SSL) where labeled data is scarce [10, 40, 18]. DINO[6] focuses on SSL with ViT[17] and proposes using a dynamically evolving teacher built from the Exponential Moving Average(EMA) of the student model, obviating the need for a pretrained fixed teacher during training. We extend upon the EMA distillation idea of DINO to multi-modal learning. Specifically, We feed the image and text embeddings output by the EMA teacher into the optimal transport module, and solve for the intra-batch optimal coupling, which is used as soft labels for knowledge distillation.

3 Methods

Our model has a two-tower structure with an image encoder and a text encoder that outputs fixed-sized embeddings for a batch of corresponding images and captions. Different from pervious ZSL works, our model assumes no class hierarchy. This makes our method more general, and easily extensible to datasets like Google Open Images[34].

3.1 Contrastive Learning

The contrastive learning[23] objective has been widely used in NLP and is at the core of several unsupervised[29, 57, 27] and self-supervised learning works[24, 9]. Similar to CLIP[44], we also use the contrastive loss, which measures the similarities of sample pairs in an embedding space. Specifically, we use the InfoNCE[52] loss where similarity is measured by dot product. Take a batch of N image and text pairs, the image and text encoders are joinly 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 ith image, and z_j^T that of the jth text. The probability of the ith image matching the jth text is:

$$P(z_i^I, z_j^T; \tau) = \frac{\exp(z_i^I \cdot z_j^T / \tau)}{\sum_{k=0}^{N} \exp(z_i^I \cdot z_k^T / \tau)} \tag{2}$$

The InfoNCE loss for images is defined as:

$$L_I = -\frac{1}{N} \sum_{i=0}^{N} \log P(z_i^I, z_i^T; \tau) \tag{3}$$

Symmetrically, we define the InfoNCE loss for texts:

$$L_T = -\frac{1}{N} \sum_{i=0}^{N} \log P(z_i^T, z_i^I; \tau) \tag{4}$$

The contrastive loss function thus becomes:

$$L_{\text{InfoNCE}} = \frac{1}{2}(L_I + L_T) \tag{5}$$

3.2 Optimal Transport

Image-text pairs collected from the Internet are usually only weakly correlated and noise is abundant. In a single batch, it's common for one caption to match objects in multiple images, and one image to match words in multiple captions. While the InfoNCE loss provides important learning signals, its supervision is noisy and fails to capture the many-to-many relationships in a batch of image-text pairs. Hence, it's not ideal to use hard labels as the only learning objective.

Algorithm 1: PyTorch Pseudocode

```
# gs: Model initialized with pretrained image and text encoders.
# gt: EMA teacher initialized with gs.
# tpi, tpk: temperature for InfoNCE and KLDiv losses.

for img, txt in loader:
  I_emb_s, T_emb_s = gs(img, txt) # Student embbeddings
  logits_s = I_emb_s @ T_emb_s.T

  I_emb_t, T_emb_t = gt(img, txt) # EMA embeddings
  sim_ii, sim_tt, sim_it, sim_ti = compute_similarities(I_emb_t, T_emb_t)

  # InfoNCE Loss
  labels = torch.arange(n)
  L_I = cross_entropy(logits_s * tpi, labels)
  L_T = cross_entropy(logits_s.T * tpi, labels)
  L_infoNCE = (L_I + L_T)/2

  # Optimal Transport
  I_cost = - (sim_ii + sim_tt + sim_it)
  T_cost = - (sim_ii + sim_tt + sim_ti)
  I_target = sinkhorn(I_cost, eps, iter)
  T_target = sinkhorn(T_cost, eps, iter)

  # KLDiv Loss
  L_KL = [KL(logits_s * tpk, I_target) + KL(logits_s.T * tpk, T_target)]/2

  loss = L_infoNCE + alpha * L_KL
  loss.backward()
  update_EMA(gs, gt)

def compute_similarities(I_emb, T_emb):
  sim_ii, sim_tt = I_emb @ I_emb.T, T_emb @ T_emb.T
  sim_it, sim_ti = I_emb @ T_emb.T, T_emb @ I_emb.T
  return sim_ii, sim_tt, sim_it, sim_ti
```

As a solution, we keep an Exponential Moving Average (EMA) of our model during training, and feed the output embeddings of the EMA model into an optimal transport (OT) module. The OT module finds the optimal coupling between a batch of image-text pairs, which we use as soft labels for knowledge distillation.

Solvers for the optimal transport problem defined in 1 are usually based on linear-programming. They have super-cubic complexity and are not differentiable. Instead, we use the Sinkhorn algorithm [12], which provides an efficient and differentiable way to solve the entropy regularized optimal transport problem. Let $\{(z_i^I, z_i^T)\}, i = 1, 2, \ldots, N$ be the image and text embeddings extracted from the EMA model in a batch of N image text pairs. Assuming a discrete uniform distribution μ over the batch, we use the sinkhorn algorithm to solve for the optimal coupling $T_I^* \in \mathbb{R}^{N\times N}$ from images to texts.

$$T_I^* = \underset{T\in\Pi(\mu,\mu)}{\arg\min} \langle T\,, C\rangle_F - \lambda H(T) \tag{6}$$

where

$$C_{ij} = -(z_i^I \cdot z_j^I + z_i^T \cdot z_j^T + z_i^I \cdot z_j^T) \tag{7}$$

and

$$H(T) = -\sum_{i,j} \log(T_{ij})T_{ij} \tag{8}$$

Symmetrically, we solve T_T^* as the optimal coupling from texts to images. When comparing the ith and jth image-text pairs, we take into account intra-domain and inter-domain embedding similarities. When the image embeddings in the two image-text pairs are close, it's more likely that there's a match in ith image and jth text. This formulation helps the model learn cross-modal connections based on the similarities of images and texts in two image-text pairs, when there's considerable noise in the data.

3.3 EMA Knowledge Distillation

The InfoNCE loss defined above is equivalent to the cross entropy loss with target probability of 1 for corresponding images and texts. When learning from soft labels, it's natural to use KL divergence as an extension of the InfoNCE loss, with logits computed from dot products. Additionally, Exponential Moving Average (EMA) has been empirically demonstrated to help models learn better representations in domains like self-supervised learning[6, 24]. During training, we keep an EMA of our model and use its output to solve for the soft targets for distillation.

Given T_I^* and T_T^* solved by the OT module, we use a KL divergence loss to match the outputs of our model with the optimal coupling. According to equation (2), define P_I as the probability distribution of images over texts in a batch for our model. Symmetrically, define P_T for texts over images.

$$L_{\text{KL}} = \frac{1}{2}[\text{KL}(P_I, T_I^*) + \text{KL}(P_T, T_T^*)] \tag{9}$$

The final loss we use is:

$$L = L_{\text{InfoNCE}} + \alpha L_{\text{KL}} \tag{10}$$

where α is set to 1.0 in our experiments.

4 Experiments

4.1 Visual and Language Pretraining

Pretraining has become a crucial procedure in many NLP tasks[16, 5, 37]. Likewise, BiT[33] and ViT[17] has shown that transfer of pretrained visual representations leads to significant performance gains. While image caption pairs are relatively expensive to collected, there are large-scale image or text datasets available with pretrained models. Therefore, we initialize our model with an image encoder pretrained on ImageNet[14] 1k and a pretrained text encoder, such as DeCLUTR[21], Sentence Transformers[45], or Bert[16]. Sentence Transformers are pretrained on SNLI[4] and MultiNLI[55]. DeCLUTR is pretrained on the OpenWebText Corpus[22] or the Semantic Scholar Open Research Corpus[38]. Bert is pretrained on the English Wikipedia and the BookCorpus[61].

4.2 Training

We apply a training schedule similar to the finetuning step of BiT[33]. 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. All of our models are trained on the Conceptual Captions[50] 3M dataset. We train the model on 4 GPUs using Pytorch[42] Distributed Data Parallel with a batch size of 128 per GPU for 30 epochs. While CLIP[44] 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.

4.3 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 neighors of its embedding based on cosine similarity.

4.3.1 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 overlap with any of the true labels. In ImageNet[14], each image is only labeled with one synset, but in Google Open Images[34], each

Dataset	Image Encoder	Text Encoder	Params	Flat Hit@k(%)			
				1	2	5	10
CLIP (400M)	ResNet50*	Bert Base*	102M	26.5	38.3	54.0	64.3
CLIP (400M)	ViT-B/32*	Bert Base*	151M	27.5	39.5	55.3	65.4
CC (3M)	FBNet C[56]	DeCLUTR Sci Base	114M	24.3	36.1	52.7	64.5
CC (3M)	EfficientNet B0[51]	DeCLUTR Sci Base	114M	26.1	38.6	56.1	68.5
CC (3M)	ResNet50	Sentence Bert Base	134M	27.6	38.9	53.4	63.3
CC (3M)	ResNet50	Bert Base	134M	27.5	39.1	54.5	64.6
CC (3M)	ResNet50	DeCLUTR Sci Base	135M	**30.2**	**43.1**	**59.3**	**70.5**

Table 1: Flat hit @k on Google Open Images. In the Dataset column, CC is the Conceptual Captions dataset. * means that the model is a modified version.

image is labeled with multiple classes. The formal definition of flat hit@k is:

$$\text{flat hit@k} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\{F(x_i)\}_K \cap L_i \neq \varnothing\} \tag{11}$$

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

4.3.2 Evaluation Dataset

We measure the ZSL performance mainly on Google Open Images [34]. 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[44]. We discuss our considerations below.

ImageNet 21K+1K: Despite its popularity, there are four main problems of using ImageNet[14] for ZSL evaluation. First, based on the WordNet[19] 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 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 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[34] 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[44] 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.

4.4 Results on Google Open Images

We evaluate the models on the test set of Google Open Images V6[34], with 125,436 images. Traditional ZSL baselines aren't evaluated on GOI due to the lack of a class structure. The image encoders are initialized with weights pretrained on ImageNet 1k. Sentence Bert[45] is pretrained

Dataset	Model	Image Encoder	Text Encoder	Flat Hit@k(%)			
				1	2	5	10
IN1k (1.2M)	DeViSE	ResNet50	skip-gram	0.3	0.9	2.2	3.6
IN1k (1.2M)	ConSE	ResNet50	skip-gram	0.1	1.5	3.5	4.9
IN1k (1.2M)	GCNZ	ResNet50	GloVe	1.0	2.3	5.3	8.1
IN1k (1.2M)	HZSL	ResNet50	GloVe*	2.2	4.6	9.2	12.7
CC (3M)	Ours	FBNet C	DeCLUTR Sci Base	2.8	4.3	8.0	11.9
CC (3M)	Ours	EfficientNet B0	DeCLUTR Sci Base	3.1	4.6	8.5	12.5
CC (3M)	Ours	ResNet50	Bert Base	3.2	**5.7**	**10.6**	**15.7**
CC (3M)	Ours	ResNet50	Sentence Bert Base	3.5	5.2	9.9	14.8
CC (3M)	Ours	ResNet50	DeCLUTR Sci Base	**3.7**	5.5	9.9	14.2
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.

on SNLI[4] and MultiNLI[55], Declutr Sci Base[21] is pretrained on the S2ORC[38], and Bert[16] on the English Wikipedia and the Book Corpus[61]. In table 1, we compare the flat hit@k of our models with pretrained CLIP[44]. Our ResNet50 and DeCLUTR Sci Base model trained with the joint contrastive and OT distillation loss exceeds CLIP ResNet50 and Bert[16] by 14% relatively in FH@k=1, while using $>$ 100x fewer image-text pairs.

4.5 Results on ImageNet 21k+1k

In this section, we present flat hit@k results on zero-shot transfer to the ImageNet 21k+1k[14] dataset, which contains 21841 classes in total. 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[44] vastly outperforms previous methods. Our method uses Conceptual Captions[50] 3M, which is on the same order of magnitude as ImageNet 1k, and outperforms the previous SoTA, HZSL[36], 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. Optimal transport distillation also encourages the model to output a softer probability output for multiple classes, which can be present but just not labeled in ImageNet.

5 Analysis

5.1 What contributes to performance gain?

In this section, we evaluate the performance of contrastive zero-shot learning under different modes of training, to demonstrate the effectiveness of OT distillation. In all three experiments, the image encoder is a ResNet50 pretrained on ImageNet1k and the text encoder is a DeCLUTR Sci Base pretrained on S2ORC. In the first experiment, we train the model with only contrastive loss using hard labels (same as CLIP[44]). In the second experiment, we train the model with a hybrid of contrastive loss and soft distillation loss, where we use the output of the EMA model directly as soft labels. In the third experiment, we also train the model with a hybrid of hard contrastive and soft distillation losses, but the soft labels are computed by the optimal transport module from the output of the EMA model. In table 3, we show that initializing with pretrained image and text encoders alone yields good results on GOI through joint end-to-end contrastive learning. Adding an EMA teacher and directly distilling from its output helps mitigate noise in image-text pairings and achieves an improvement of 1.2% on GOI F@K=1. Furthermore, we demonstrate that optimal transport under our cost formulation in equation 7 is effective in finding an optimal intra-batch coupling and improves performance by another 0.8% on GOI F@K=1.

Mode	Flat Hit@k(%)			
	1	2	5	10
CLIP (RN50+Bert)	26.5	38.3	54.0	64.3
Contrastive	28.2	40.6	57.6	68.7
EMA self-distillation	29.4	42.2	59.0	70.0
EMA + OT self-distllation	**30.2**	**43.1**	**59.3**	**70.5**

Table 3: Flat hit @k on Google Open Images under different modes of training. Here, the image encoder is ResNet 50 and the text encoder is DeCLUTR Sci Base.

Image Encoder	Text Encoder	Flat Hit@k(%)			
		1	2	5	10
RN50(CLIP)	Bert(CLIP)	26.5	38.3	54.0	64.3
RN50(IN1k)	DeCLUTR(S2ORC)	**30.2**	**43.1**	**59.3**	**70.5**
RN50(IN1k)	DeCLUTR(WebText)	12.2	19.1	30.7	40.0
RN50(IN1k)	Bert(Wiki)	27.5	39.1	54.5	64.6
RN50(Rand)	Bert(Rand)	fails	fails	fails	fails

Table 4: Flat hit @k on Google Open Images for models pretrained on different datasets.

5.2 Pretraining

In the section, we analyze the effects of using image and text encoders pretrained on different datasets. From Table 4, pretraining clearly has a significant effect on the performance of the model. The Open WebText Corpus[22] contains more than 8 million documents extracted from HTMLs crawled from the Internet, and S2ORC [38] consists of over 2 million scientific papers. While models trained from scratch struggles to converge, models pretrained on more structured data, such as S2ORC and Wikipedia, perform much better than those pretrained on crawled web texts.

6 Conclusion

Language-supervised zero-shot learning trained under a contrastive loss has shown impressive performance gains, but remains very data-hungry. CLIP, for example, requires 400M image-text pairs. To improve data efficiency and mitigate data noise, we propose a data-efficient ZSL training pipeline that enables any pretrained image encoders to generalize to unseen classes. We recognize the noisy nature of the image-text pairs collected from the Internet, and the many-to-many relationships in a batch of image-text pairs. While CLIP uses a hard InfoNCE loss, which ignores the potential image-text matchings in different image-text pairs, we use a hybrid of hard contrastive and soft distillation losses. The soft labels for distillation are solved by an optimal transport module with a specifically designed cost function to guide cross-modal matching. Furthermore, we recognize the many problems of previously used benchmarks such as ImageNet 21k+1k, and propose using Google Open Images as a new multi-label benchmark for ZSL. Using >100x fewer image-text pairs than CLIP, we demonstrate a highly efficient zero-shot learning method that exceeds CLIP's performance on Google Open Images, and achieves strong results on ImagetNet 21k+1k.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. *CVPR*, 2013.

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. *CVPR*, 2015.

[3] Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. Label refinery: Improving ima- genet classification through label progression. *arXiv preprint arXiv:1805.02641*, 2018.

[4] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. 2015.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[7] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. *ICML*, 2020.

[8] Liqun Chen, Zhe Gan, Dong Wang, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. *CVPR*, 2021.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020.

[10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, 2020.

[11] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *arXiv preprint arXiv:1507.00504v2*, 2016.

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 2013.

[13] Bharath Bhushan Damodaran, Rémi Flamary, Vivien Seguy, and Nicolas Courty. An entropic optimal transport loss for learning deep neural networks under label noise in remote sensing images. *arXiv preprint arXiv:1810.01163*, 2018.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255, 2009.

[15] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[18] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. SEED: self-supervised distillation for visual representation. *ICLR*, 2021.

[19] Ingo Feinerer and Kurt Hornik. *wordnet: WordNet Interface*, 2020. R package version 0.1-15.

[20] Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *NIPS*, 2013.

[21] John M Giorgi, Osvald Nitski, Gary D. Bader, and Bo Wang. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*, 2020.

[22] Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus, 2019.

[23] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. *CVPR*, 2006.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[26] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[27] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2019.

[28] Zeyi Huang, Yang Zou, Vijayakumar Bhagavatula, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. *NeurIPS*, 2020.

[29] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *ICML*, 2020.

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2020.

[31] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling BERT for natural language understanding. *EMNLP*, 2020.

[32] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing. Rethinking knowledge graph propagation for zero-shot learning. *CVPR*, 2019.

[33] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 2020.

[34] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[35] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Jia Li. Learning from noisy labels with distillation. *arXiv preprint arXiv:1703.02391*, 2017.

[36] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. *CVPR*, 2020.

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[38] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics.

[39] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. *ECCV*, 2018.

[40] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. *CVPR*, 2018.

[41] 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.

[42] 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'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[43] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. *EMNLP*, 2014.

[44] 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.

[45] 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, 11 2019.

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

[47] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *ICLR*, 2015.

[48] Tim Salimans, Han Zhang, Alec Radford, and Dimitris N. Metaxas. Improving gans using optimal transport. *ICLR*, 2018.

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

[50] 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.

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

[52] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1191.05722*, 2018.

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

[54] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Kdgan: Knowledge distillation with generative adversarial networks. *NeurIPS*, 2018.

[55] 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.

[56] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. *CVPR*, 2019.

[57] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. *CVPR*, 2018.

[58] Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *CVPR*, 2020.

[59] Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Distilling knowledge from graph convolutional networks. *CVPR*, 2020.

[60] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and texts. *arXiv preprint arXiv:2010.00747*, 2020.

[61] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *ICCV*, 2015.