ATENEO DE MANILA UNIVERSITY

DISTILLCLIP - KNOWLEDGE DISTILLATION OF CONTRASTIVE LANGUAGE-IMAGE PRETRAINED MODELS

A THESIS SUBMITTED TO
THE GRADUATE FACULTY OF
THE SCHOOL OF SCIENCE AND ENGINEERING
IN CANDIDACY FOR THE DEGREE OF
MASTER OF SCIENCE IN
COMPUTER SCIENCE

DEPARTMENT OF INFORMATION SYSTEMS
AND COMPUTER SCIENCE

BY
PATRICK JOHN C. RAMOS

QUEZON CITY, PHILIPPINES
JUNE 2023

The THESIS entitled:

DISTILLCLIP - KNOWLEDGE DISTILLATION OF CONTRASTIVE LANGUAGE-IMAGE PRETRAINED MODELS

submitted by Patrick John C. Ramos has been examined and is recommended for **Oral Defense**.

PATRICIA ANGELA R. ABU, Ph.D.
Chair

RAPHAEL B. ALAMPAY, Ph.D.
Adviser

PATRICIA ANGELA R. ABU, Ph.D.
Co-Adviser

RAPHAEL A. GUERRERO, Ph.D.
Dean
School of Science and Engineering

The Faculty of the Department of Information Systems and Computer Science, School of Science and Engineering, Ateneo de Manila University ACCEPTS THE THESIS entitled:

DISTILLCLIP - KNOWLEDGE DISTILLATION OF CONTRASTIVE LANGUAGE-IMAGE PRETRAINED MODELS

submitted by Patrick John C. Ramos in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

RAPHAEL B. ALAMPAY, Ph.D.
Member

JOHN PAUL C. VERGARA, Ph.D.
Member

JANN RAILEY E. MONTALAN, M.Sc.
Member

RAPHAEL B. ALAMPAY, Ph.D.
Adviser

PATRICIA ANGELA R. ABU, Ph.D.
Co-Adviser

RAPHAEL A. GUERRERO, Ph.D.
Dean
School of Science and Engineering

Grade: A-
Date: JUNE 29, 2023

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES vii

CHAPTER

ABSTRACT viii

ACKNOWLEDGMENTS ix

I INTRODUCTION 1
1.1 Statement of the Problem 1
1.2 Research Objectives 2
1.3 Research Questions 3
1.4 Scope and Limitations 3
1.5 Significance of the Study 4

II REVIEW OF RELATED LITERATURE 6

III METHODOLOGY 9
3.1 Preliminaries: CLIP 9
3.2 CLIP Knowledge Distillation 10
3.2.1 Similarity Map Distillation 11
3.2.2 Training Objective 13
3.2.3 Training 13
3.3 Evaluation 14
3.3.1 Datasets 15
3.3.2 Zero-Shot Transfer 16
3.3.3 Linear Probe Evaluation 17
3.3.4 Robustness to Natural Distribution Shift 17
3.3.5 Video Accident Recognition 18
3.4 Comparison to CLIP 19

IV RESULTS 20
4.1 Zero-Shot Transfer 20
4.2 Linear Probe Evaluation 21
4.3 Robustness to Natural Distribution Shift 22
4.4 Video Accident Recognition 24

4.5 Size and Speed Comparison 24

V CONCLUSION 25

Appendix A MODEL CARD 34

Appendix B MODEL ARCHITECTURE CONFIGURATIONS . . 38

Appendix C WEB DEMO 39

LIST OF FIGURES

Figures Page

3.1 Overview of the CLIP knowledge distillation scheme. 11
3.2 Overview of datasets and tasks. 14

C.1 DistillCLIP zero-shot image classification web demo. 39

LIST OF TABLES

Tables Page

4.1 Zero-shot image classification results. 21
4.2 Linear probe evaluation results. 21
4.3 Zero-shot transfer results on distribution shift datasets. . . . 22
4.4 Video accident recognition results. 23
4.5 Size and speed of CLIP and DistillCLIP 24

5.1 Summary of results. 27

A.1 Model Card 37

B.1 Teacher and student model configurations. 38

ABSTRACT

Despite CLIP's performance on vision-language tasks, CLIP's size limits its deployment in low resource environments. We propose a knowledge distillation scheme to compress a teacher CLIP into a smaller student model we term DistillCLIP. Our framework consists of distilling both intra-modal and inter-modal similarity maps between and within image and text embeddings. DistillCLIP is 43.69% the size of CLIP and has 82.43% its FLOPs. We show that the ability of DistillCLIP to retain teacher performance on zero-shot transfer tasks may depend on the semantic granularity of class labels, preserving only 63.81% of teacher accuracy on average. Meanwhile DistillCLIP's linear probe performance matches and on some datasets surpasses that of the teacher CLIP with an average retention rate of 100.53%. However, DistillCLIP retains only 12.28% teacher accuracy on average on distribution shift datasets. We also demonstrate that DistillCLIP is able to preserve 99.34% teacher accuracy on video accident recognition in dashcam videos.

ACKNOWLEDGMENTS

I would like to acknowledge advisors Raphael Alampay, Ph.D. and Patricia Abu, Ph.D. for their invaluable guidance; panelists John Vergara, Ph.D. and Jann Montalan, M.Sc. for their useful comments and suggestions; fellow researcher Ryan Ramos for his indispensable input throughout the thesis; and Ateneo de Manila University for its continued support throughout my studies.

CHAPTER I

INTRODUCTION

Vision-language pretrained models encode images and text to a shared semantic space. These models can then be fine-tuned on various downstream vision-language tasks such as visual question answering [3], visual entailment [54], and text-to-image retrieval. These tasks require models to understand both visual and linguistic information. For example, in text-to-image retrieval, a model must extract the linguistic content of the text query and the visual content of each candidate image to identify the image best described by the query.

Popularized by CLIP [38], recent success in vision-language pretraining has been achieved by multi-encoder models that employ or incorporate a contrastive learning objective [24, 37, 46, 55]. Motivated by language-supervised image representation learning, CLIP exhibits zero-shot transfer performance on computer vision benchmarks competitive with fully supervised task-specific baselines. Although originally evaluated on computer vision tasks, CLIP is a vision-language model by design, and has seen further applications in visual question answering [45], visual entailment [45], text-guided object detection [56], text-guided image segmentation [52], and text-to-image synthesis [10, 34, 39, 40].

1.1 Statement of the Problem

Despite state-of-the-art performance across a variety of vision-language tasks, the computational requirements of CLIP make it difficult to deploy in resource constrained environments such as edge devices. The more performant CLIP models use a ViT [14] as an image encoder and a Transformer [48] model structurally similar to BERT [13] as a text encoder. These encoders have individually already been reported to be computationally expensive for low-resource devices [18, 25, 27, 30, 43, 57]. CLIP becomes even more expensive compared to these individual encoders as it uses both of them in its dual-encoder architecture, motivating the need for compressing models into smaller and faster versions that still perform close to their uncompressed ones.

Various methods exist to compress models, such as pruning [21] and quantization [17]. Of particular interest to this paper is knowledge distillation [23], which trains a smaller student model to match the outputs of a larger teacher model. While knowledge distillation has been applied to BERT [26, 43] and ViT [25], knowledge distillation for CLIP is still an emerging field [42, 49].

1.2 Research Objectives

Knowledge distillation is non-trivial, as there is no fixed framework for performing distillation. There exist multiple supervisory signals from teacher models, and the choice of these signals and how these signals are used may affect the student's performance. Therefore, we plan to design a knowledge distillation scheme for CLIP.

We will then evaluate the performance of DistillCLIP on tasks CLIP was originally evaluated on, along with a practical use case of CLIP's expressive encoders. Because the student is expected to reproduce the behavior of the teacher to a certain extent, it is intuitive to evaluate the distilled

CLIP on the same benchmarks used for the teacher CLIP, which include zero-shot transfer, linear probe evaluation, and robustness to natural distribution shift. Additionally, we evaluate DistillCLIP on traffic accident recognition [6], a practical application of CLIP.

Lastly, we aim to compare the performance, parameter count, and inference speed of DistillCLIP to the teacher CLIP. Given that the purpose of knowledge distillation is to have a student that is smaller and faster than its teacher but still competitive in performance, it is important to evaluate the performance, size, and speed of the student with respect to the teacher.

1.3 Research Questions

1. What supervisory signals can be used to distill CLIP into a smaller model?
2. How does the distilled CLIP perform on tasks CLIP was evaluated on?
   (a) How does the distilled CLIP perform on zero-shot transfer?
   (b) How does the distilled CLIP perform on linear probe evaluation?
   (c) How robust is the distilled CLIP to natural distribution shift?
3. How does the distilled CLIP perform on traffic accident recognition?
4. How does the performance, parameter count, and FLOPS of the distilled CLIP compare to those of CLIP?

1.4 Scope and Limitations

Our training and evaluation experiments will mostly be constrained by computational and storage limitations. Although CLIP was pretrained on a large-scale dataset of 400 million image-text pairs, we will train on a much smaller dataset as a result of storage constraints. Furthermore, while CLIP was evaluated on over 30 computer vision benchmarks, we will only evaluate the distilled CLIP on a smaller set of tasks due to computational limitations. However, we will expand the scope of tasks of the original CLIP paper by also evaluating on traffic accident recognition.

1.5 Significance of the Study

The significance of a knowledge-distilled CLIP comes from 1) the value of CLIP itself, and 2) the value of a small, fast, but performant CLIP.

While many vision-language models are task-specific in the sense that they are explicitly trained to perform only one type of task (ex. image captioning, visual question answering), CLIP can be considered a more general pretrained model that can be applied to different vision-language use cases. CLIP is one of several models in recent years [5, 13] termed “foundation models” [4], which demonstrate the ability of models pretrained on large amounts of data to be fine-tuned on several downstream tasks.

Because CLIP in particular can jointly encode both visual and linguistic information, it can be applied to a plethora of tasks which require understanding such information. For example, CLIP can be a component for visual question answering for low vision users [19], image captioning models for those with low vision [20], or interactive technologies that describe or answer students’ questions about images [1]. The practical use case we evaluate on, accident recognition, is especially useful for driving environments as it can be used in advanced driver-assistance systems and autonomous driving scenarios.

However, many of these applications work best when they are deployed on embedded or edge devices, such as a mobile phone or an on-vehicle device attached to a dashcam. Because CLIP is considered computationally expensive, it might not perform well on such devices. Therefore, end-users would benefit from a smaller and faster version of CLIP whose performance is still acceptable for their use case.

CHAPTER II

REVIEW OF RELATED LITERATURE

Knowledge distillation is a model compression method where a model is made a teacher whose behavior a smaller student model is trained to reproduce. The intuition behind knowledge distillation is that the outputs of a trained model often contain more information than the supervisory signals it was originally trained with. For example, while the output probability distribution for a vision encoder that correctly identifies a cat gives a much higher probability to the cat class than any other class, it might still give higher probabilities to other animal classes compared to non-animal classes. These minute differences offer a much richer training signal than the original one-hot vector indicating the ground truth label. Training a student to not only match the ground truth labels but also the output distributions of a teacher model can help it learn better than relying solely on the ground truth labels.

The standard knowledge distillation setup exists in a classification setting, and trains the student model with a linear combination of a cross entropy loss with ground truth labels and a cross entropy loss with “soft” labels from the teacher’s output probability distributions. Subsequent work into knowledge distillation revolve around proposing novel “novel distillation schemes”, which typically involve identifying other outputs of the teacher model that can be used as supervision signals for the student.

Precursory work can be found in knowledge distillation of the Transformer encoders that comprise CLIP[1] are still of relevance. Prior works have

[1] CLIP was also proposed with ResNet-50 vision encoder configurations, but in this work

distilled these encoders using a variety of learning signals from the teacher, such as hidden states [26, 43], attention matrices [26, 50], and token-level manifolds [25] with cosine, MSE, and Kullback-Leibler divergence losses.

Of prior work on CLIP knowledge distillation, few focus on compressing CLIP but rather on transferring knowledge between CLIP and non-CLIP models. Prior studies have distilled unimodal encoders to the individual CLIP encoders [51], from architecturally different multimodal models to CLIP [53], or from individual CLIP encoders to unimodal generative models [11].

Furthermore, several studies that do distill between CLIP architectures are not motivated by model compression. Some studies in CLIP distillation focus specifically on self-distillation [2, 8], a variant of knowledge distillation where the teacher is based on the student, to improve data efficiency. This is not used for model compression as the teacher and the student are of the same size.

Of existing work into knowledge distillation for CLIP model compression, MoTIS [42] and ConaCLIP [49], the latter of which is concurrent to our study, are the most similar to our work. Both perform distillation through a two-stage pre-training-and-fine-tuning framework. MoTIS first individually compresses the image and text encoders with an intra-modal contrastive objective, then performs task-specific fine-tuning by distilling within and across modalities using contrastive and Kullback-Leibler divergence losses, respectively. ConcaCLIP follows a fully-connected "knowledge interaction graph" and distills each student from itself, its teacher, and both the student and teacher of the opposite modality using contrastive, squared l_2 norm, and

we focus only on CLIP using Transformer encoders for both images and text.

Kullback-Leibler divergence losses.

Our study differs from these studies however as our method is simpler; our proposed knowledge distillation scheme is only a one-stage framework and with only one type of loss for distilling intra and inter-modal knowledge. Furthermore, while they evaluate their model primarily on image-text retrieval, our evaluations focus on natural language supervised image classification.

CHAPTER III

METHODOLOGY

After an overview of preliminary concepts surrounding CLIP, we discuss our approach to distilling CLIP into DistillCLIP, evaluating DistillCLIP, and comparing DistillCLIP to CLIP.

3.1 Preliminaries: CLIP

CLIP (Contrastive Language-Image Pre-training) [38] is a vision-language model pretrained with a contrastive loss. It is trained to jointly embed images and texts such that images and texts with similar semantic content have similar vector representations. Opposed to other multimodal fusion-encoder architectures which encode both visual and linguistic information with a shared encoder, CLIP follows a dual-encoder architecture and separately encodes texts and images with individual modality-specific encoders.

Given a batch of image-text pairs where the each text describes its corresponding image, CLIP independently embeds the images and texts with an image encoder and a text encoder, respectively. The hidden image and text representations are then linearly projected to a contrastive multimodal embedding space. CLIP is trained with an InfoNCE loss [35] on the cosine similarities of the image and text embeddings. This trains the model to align the embeddings of images and texts belonging to the same pair and contrast the embeddings of images and texts belonging to different pairs.

In practice, the image and text embedders are Transformer encoders. Images are tensors of shape $channel \times height \times width$ which are reshaped into a list of patches each of shape $patch\_height \times patch\_width$. These patches are then embedded into d_v-dimensional vectors, and the list of patch embeddings is prepended with a `[CLS]` token of the same dimensionality. After processing all tokens with the vision Transformer encoder, the `[CLS]` is projected into d dimensions. Meanwhile, texts are tokenized into sequences of token IDs, and each token is projected into d_t-dimensional vectors with each sequence prepended with its own d_t-dimensional `[CLS]` token. Similar to the images, the whole sequence of embeddings is processed by the text Transformer and the `[CLS]` token is projected one last time to d dimensions. These `[CLS]` tokens serve as each data point's image and text vector representations.

Processing a batch of image-texts pairs, where each text describes its corresponding image, produces a batch of image and text representations $z^v, z^t \in \mathbb{R}^{b \times d}$, where z_i^v and z_i^t correspond to the same data point and b refers to the batch size. Afterwards, an inter-modal similarity map $S_{inter} \in [-1, 1]^{b \times b}$ is produced, where $S_{i,j}$ refers to the similarity of image i and text j. This is computed by taking the dot product $\widetilde{z^v}\widetilde{z^{t\intercal}}$, where $\widetilde{x}$ refers to the row-wise l_2-normalized x. This essentially computes a cosine similarity map between the image and text vectors and represents each similarity as a scalar in $[-1, 1]$. InfoNCE loss is then used to maximize the diagonal of S and minimize the off-diagonal of S, training CLIP to only produce similar vectors for images and texts that are semantically similar.

3.2 CLIP Knowledge Distillation

Figure 3.1 provides an overview of the proposed CLIP distillation approach. In this particular figure, assume all model outputs are already l_2-normalized.

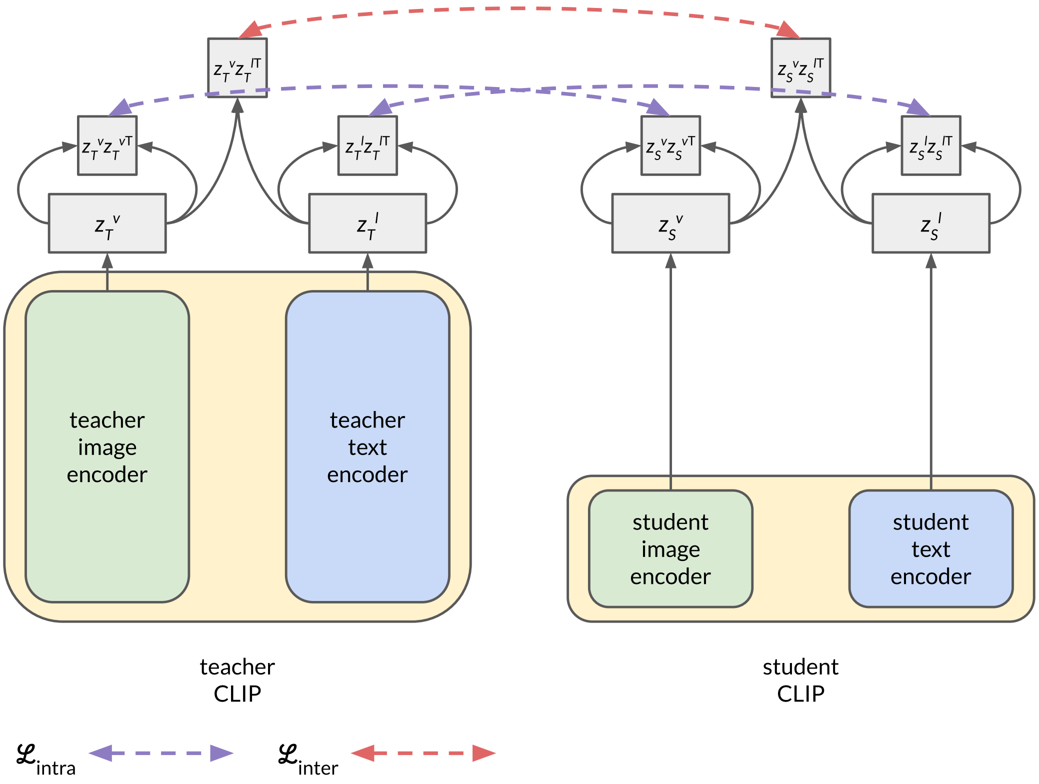


Figure 3.1: Overview of the CLIP knowledge distillation scheme.

We distill a teacher CLIP T with image and text Transformer encoders T^v and T^l of embedding sizes d_T into a student CLIP S with image and text Transformer encoders S^v and S^l with embedding sizes $d_S < d_T$. The teacher image and text encoders produce outputs z_T^v and z_T^t while the student image and text encoders produce representations z_S^v and z_S^t.

3.2.1 Similarity Map Distillation

The main intuition behind our knowledge distillation scheme lies in the similarity map $S_{inter} \in [-1, 1]^{b \times b}$ produced by CLIP. These similarity scores can be interpreted as logits representing the probability that image i and text j match, or share the semantic content. As such, the similarity map contains useful information about how to embed inputs. Therefore, we use the logits of these inter-modal image-text similarity maps as supervisory signals for distillation. As discussed in Section 3.2.1.2, we also use the logits of intra-modal image-image and text-text similarity maps.

3.2.1.1 Inter-modal Similarity Map Distillation

We align the logits of the image-text similarity map of the student encoders with those of the teacher encoders with a squared l_2-norm loss $\mathcal{L}_{inter}$:

$$\mathcal{L}_{inter} = \|\widetilde{z_T^v}\widetilde{z_T^{t\intercal}} - \widetilde{z_S^v}\widetilde{z_S^{t\intercal}}\|_2^2 \tag{3.1}$$

3.2.1.2 Intra-modal Similarity Map Distribution

We observe that it is not only possible to compute inter-modal (i.e. image-text) similarity maps with CLIP but also intra-modal ones (i.e. image-image, text-text) too. While CLIP was not originally trained with intra-modal similarity maps, derivative works [16, 33] have employed them to improve CLIP performance. Additionally, intra-modal similarity maps have been used in unimodal contrasive learning for images [7] and text [15].

For embeddings z^m of modality $m \in v, t$, the intra-modal similarity map can be produced by computing the cosine similarity map $S^m_{intra} = \widetilde{z_T^m}\widetilde{z_T^{m\intercal}} \in \mathbb{R}^{b \times b}$. Each value in S^m_{intra} represents a probability that image/text i matches image/text j. We align image-image and text-text similarity maps using a squared l_2-norm loss $\mathcal{L}_{intra}$:

$$\mathcal{L}_{intra} = \|\widetilde{z_T^v}\widetilde{z_T^{v\intercal}} - \widetilde{z_S^v}\widetilde{z_S^{v\intercal}}\|_2^2 + \|\widetilde{z_T^t}\widetilde{z_T^{t\intercal}} - \widetilde{z_S^t}\widetilde{z_S^{t\intercal}}\|_2^2 \tag{3.2}$$

3.2.2 Training Objective

The final loss is a weighted sum of the inter and intra-modal losses:

$$\mathcal{L} = \lambda_{inter}\mathcal{L}_{inter} + \lambda_{intra}\mathcal{L}_{intra}, \tag{3.3}$$

where λ_r and λ_s are the weights of the relative representation loss and the image-text similarity loss, respectively.

3.2.3 Training

We use a CLIP-ViT-B/32 as our teacher model. Following MoTIS and Cona-CLIP, for our student model we create a CLIP with a ViT-S/16 image encoder, a 6-layer text Transformer encoder with the same dimensionality as the teacher text encoder, and $d_S = 256$. We call this student DistillCLIP. The image encoder is initialized with pretrained ImageNet-21k weights [47] while each layer i in the text encoder is initialized witht he weights of layer $2i$ of the teacher text encoder. We train the student using our CLIP distillation scheme on Conceptual Captions 3M [44] for 33,513 steps, or approximately one epoch, using AdamW [31] with a learning rate of 3e-5, β_1=0.9, β_2=0.98, ϵ=$1\text{e}{-6}$, weight decay of $1\text{e}{-1}$, cosine learning rate decay, 10,000 warm-up steps, and a batch size of 84. These hyperparameters are modified from the MoTIS training protocol. We set λ_{inter}=λ_{intra}=1. Additionally, although Conceptual Captions 3M was originally created with 3M data points, preparing the dataset requires downloading all images from scratch, however due to data unavailability after the publication of Conceptual Captions 3M, we were only able to prepare 2.8M image-text pairs.

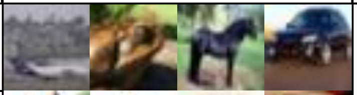
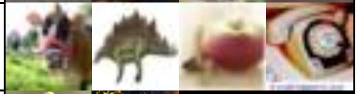
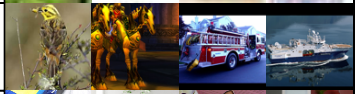
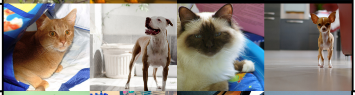
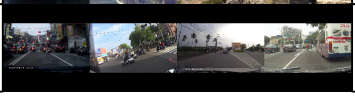
Dataset	Examples	Task/s
CIFAR-10		Zero-Shot Transfer, Linear Probe Evaluation
CIFAR-100		
STL-10		
Oxford-IIIT Pet		
ImageNetV2		Robustness to Natural Distribution Shift
ImageNet-A		
Aadv		Video Accident Recognition

Figure 3.2: Overview of datasets and tasks.

3.3 Evaluation

After training DistillCLIP, we evaluate the model on a series of benchmarks. We evaluate on three tasks CLIP was originally tested on (zero-shot transfer, linear probe evaluation, and robustness to natural distribution shift), and on, accident recognition in videos. Note that in each evaluation, the goal is not to maximize student performance but to maximize the *retention* of teacher performance, or the capacity of the student to match teacher performance. As all evaluation tasks are measured with accuracy, we define the retention rate as the student accuracy divided by the teacher accuracy.

3.3.1 Datasets

We present the datasets used in each evaluation in Figure 3.2. We describe each dataset in detail below.

CIFAR-10 Canadian Institute For Advanced Research 10 [28] is an image classification dataset of 60,000 32×32 pixel RGB images split across 10 semantically distinct classes i.e. cat, truck, ship. Each class contains 6,000 images, with 5,000 for training and 1,000 for testing.

CIFAR-100 Canadian Institute For Advanced Research 100 [28] is an image classification dataset of 60,000 32×32 pixel RGB images split across 100 semantically distinct classes i.e. apple, lion, keyboard. Each class contains 600 images, with 500 for training and 100 for testing.

STL-10 Self-Taught Learning 10 [9] is a dataset for self-supervised representation learning for image classification. The dataset consists of 10 semantically distinct classes i.e. bird, car, truck. Each class consists of 500 training images and 800 testing images. Although STL-10 contains 100,000 unlabelled images, we ignore these and only use the train and test splits. Each image is 96×96 pixels with RGB channels.

Oxford-IIIT Pet Oxford-IIIT Pet is an image classification dataset of 37 classes of cat and dog breeds. The dataset comes with species (cat and dog) annotations, which we use as “coarse labels” to create a binary classification version of the dataset. Evaluations are performed with both the fine-grained 37-class dataset and the coarse-grained 2-class dataset.

ImageNetV2 ImageNetV2 [41] is an alternative ImageNet test set of 10,000 images created after the creation of the original ImageNet [12]. It contains examples which are more difficult for ImageNet-trained models to generalize to.

ImageNet-A ImageNet-A [22] is an ImageNet test set containing natural adversarial examples that image classifiers tend to misclassify. It consists of 7,500 images across a 200-class subset of ImageNet classes.

Aadv Aadv [6] is originally a video accident anticipation dataset, where the task is to predict whether an accident will occur in succeeding frames. However, we convert this into a binary classification video accident recognition dataset, where the task is to predict whether or not a video contains an accident. Each video is 100 frames, with each frame being a 1280×720 RGB image. If a video contains an accident, it occurs in the last 10 frames of the video.

3.3.2 Zero-Shot Transfer

Radford et al. [38] propose zero-shot transfer as a benchmark to evaluate CLIP's task-learning capabilities, or the ability to generalize to unseen tasks or datasets. These tasks are usually image classification tasks. To use DistillCLIP for zero-shot transfer on a image classification dataset, we first perform prompt engineering and ensembling to create text prompts for the classes. For each label, we create multiple prompts e.g. "`A photo of a {label}.`", "`A blurry photo of a {label}.`", etc.[1] We then encode the

[1] We use the prompts used by Radford et al. [38] provided in `https://github.com/openai/CLIP/tree/main`.

prompts and the images with the text and image encoders of DistillCLIP, respectively. Each class is an ensemble of prompts represented as the l_2-normalized average of its corresponding text prompt embeddings, which are also l_2-normalized. The prompt ensemble with which an image has the highest cosine similarity is taken as the model's prediction of the image's label. While Radford et al. [38] perform zero-shot transfer on 27 tasks, we focus on the following datasets: CIFAR10 [28], CIFAR100 [28], STL-10 [9], and Oxford-IIIT Pet [36]. We center crop each image to 224×224.

3.3.3 Linear Probe Evaluation

Linear probe evaluation measures the representation learning capabilities of the model. We conduct linear probe evaluation by freezing the model, attaching a single trainable linear layer on top of it, and fine-tuning on a target image classification dataset. Although CLIP's linear probe evaluation was conducted on 12 datasets, we focus on the same datasets from the zero-shot transfer experiments for DistillCLIP, with the same cropping method.

3.3.4 Robustness to Natural Distribution Shift

While robustness to natural distribution shift typically refers to a model's ability to generalize to data that does not fit its training distribution, distribution shift robustness in terms of CLIP evaluation refers to generalization to data that does not fit the ImageNet training distribution. Neither CLIP nor DistillCLIP were trained on ImageNet, however standard practice is to train and evaluate models trained on ImageNet, hence it is of particular interest to image classifiers to exhibit robustness to ImageNet distribution shifts. We measure DistillCLIP's robustness to these shifts by performing zero-shot transfer on ImageNetV2 [41] and ImageNet-A [22].

3.3.5 Video Accident Recognition

We extend the existing CLIP evaluations to video accident recognition, a more practical benchmark. In this task, given a video we are tasked to determine whether an accident has occurred or not.

To extend CLIP and DistillCLIP to videos, we use EVL [29], a video classification architecture using a frozen pretrained CLIP backbone. Given a video, EVL first extracts multi-level spatiotemporal features by embedding each frame with a frozen CLIP backbone and taking the stacked frame embeddings at different layers of the backbone. EVL then aggregates information from these features using a Transformer decoder with a learnable `[CLS]` token. To incorporate temporal information into the spatial features extracted with CLIP, EVL employs local temporal modules composed of depthwise convolutions along the temporal dimension, temporal positional embeddings, and cross-attention between frames. We choose EVL for this task as it does not fine-tune the CLIP backbone, thereby isolating and measuring the repesentative power of CLIP, akin to linear probe evaluation.

We adopt a video accident anticipation dataset Aadv [6] for this task. As each video is 100 frames, we sample every 14th frame starting with the 0th frame to create 10-frame videos. Instead of center cropping, we shrink and pad each video frame to 224×224 so that no accident-related information is cut off.

We train EVL using AdamW with learning rate $4\mathrm{e}{-4}$, weight decay $5\mathrm{e}{-2}$, cosine learning rate decay, batch size 8, and 16 gradient accumulation steps.

3.4 Comparison to CLIP

We investigate the performance/size/speed trade-off between DistillCLIP and CLIP. Specifically, we compare the two models in terms of performance on the tasks outlined in Section 3.3, parameter count, and FLOPS.

CHAPTER IV

RESULTS

4.1 Zero-Shot Transfer

Results for zero-shot transfer experiments are presented in Table 4.1. The zero-shot ability of DistillCLIP is variable, ranging from as low as 3.38% accuracy on Oxford-IIIT Pet (3.88% retention of teacher performance) to up to 97.83% accuracy on Oxford-IIIT Pet with coarse labels (97.83% retention). We observe that DistillCLIP's zero-shot performance is dependent on the complexity of the dataset. Upon initial inspection it appears that the ability to match teacher performance degrades as the class labels increase, as DistillCLIP achieves 76.78% (85.47%) CIFAR-10 and 84.3% (86.79%) STL-10 accuracy but 29.35% (45.1%) on CIFAR-100. However, we note that DistillCLIP achieves only 3.38% (3.88%) on Oxford-IIIT Pet despite it having fewer classes compared to CIFAR-100. We believe that DistillCLIP's zero-shot performance is therefore dependent on the granularity or similarity of classes i.e. it is easier to classify between dogs and cats than it is to classify between an American Bulldog and an American Pit Bull Terrier. CIFAR-10 and STL-10 have relatively more easily separable classes (e.g. airplane vs cat) compared to CIFAR-100 (e.g. shrew vs squirrel). To test our hypothesis, we perform an additional evaluation on Oxford-IIIT Pet but using coarse labels. Rather than classify between 37 breeds of cats and dogs, we group all labels according to cat and dog superclasses. With more conceptually easily separable labels, we improve zero-shot performance to 97.83%, with is also the amount of retention of teacher accuracy on the dataset using the same

Dataset	**# Classes**	**CLIP**	**DistillCLIP**
CIFAR-10 [28]	10	89.83	76.78
CIFAR-100 [28]	100	65.08	29.35
STL-10 [9]	10	97.13	84.3
Oxford-IIIT Pet [36]	37	87.21	3.38
Oxford-IIIT Pet (coarse) [36]	2	100.00	97.83

Table 4.1: Zero-shot image classification results.

Dataset	**# Classes**	**CLIP**	**DistillCLIP**
CIFAR-10 [28]	10	94.82	95.71
CIFAR-100 [28]	100	79.43	81.86
STL-10 [9]	10	98.56	98.63
Oxford-IIIT Pet [36]	37	93.44	92.29
Oxford-IIIT Pet (coarse) [36]	2	99.93	99.73

Table 4.2: Linear probe evaluation results.

superclasses.

4.2 Linear Probe Evaluation

Table 4.2 presents classification results for linear probe evaluation. We observe that DistillCLIP achieves better classification results with linear probe evaluation compared to zero-shot transfer on all datasets. Even when DistillCLIP attains only 3.38% accuracy with zero-shot transfer on Oxford-IIIT Pet, it achieves 92.29% accuracy on the same dataset in linear probe evaluation. Retention of teacher performance is also much higher under

Dataset	# Classes	CLIP	DistillCLIP
ImagenetV2 [41]	1000	55.79	6.17
ImageNet-A [22]	200	31.37	4.55

Table 4.3: Zero-shot transfer results on distribution shift datasets.

linear probe evaluation, with retention only going as low as 98.77% on the evaluated benchmarks. Furthermore, on CIFAR-10, CIFAR-100, and STL-10, DistillCLIP actually achieves higher linear probe classification that the full-sized CLIP. This implies that although DistillCLIP embeddings are difficult to use straight out of the box to use for zero-shot image classification, they are informative enough to be linearly separable, even more so than embeddings from the original CLIP.

4.3 Robustness to Natural Distribution Shift

We present our results for zero-shot classification on natural distribution shift datasets in Table 4.3. We achieve relatively low accuracies on both datasets, only retaining 11.06% and 14.5% of teacher accuracy on ImagenetV2 and Imagenet-A, respectively. Radford et al. [38] pose the question of whether CLIP's robustness to ImageNet distribution shift is attributable to its nature as a zero-shot model, its own large training distribution, or its natural language supervision. As DistillCLIP is also a zero-shot model with implicit natural language supervision,[1] we believe that the disparity in teacher and student performance may be attributed to the 400M image-

[1] Although DistillCLIP is not explicitly trained with natural language supervision using an InfoNCE loss like CLIP, it is still implicitly trained with such as the intra-modal similarity maps distilled during training are a result of natural language supervision.

Model	**Accuracy**
CLIP	65.02
DistillCLIP	64.59

Table 4.4: Video accident recognition results.

text dataset used to train CLIP. Such a scale captures a large distribution of images and texts and contributes to CLIP's robustness. Although knowledge distillation aims to teach the student to generalize to data in the same manner as the teacher, it is possible that using only the ∼3M data points in Conceptual Captions 3M was not enough to teach the student. However, this is only a hypothesis; as we currently do not have the computational resources to use similarly large datasets, we leave it to future works to investigate the relationship between the amount of training data and distillation performance.

The difference in robustness may also explain why DistillCLIP has more easily linearly separable image and text vectors than CLIP. In CLIP's attempt to be robust and account for a wider distribution of images and texts, its embeddings occupy several spaces in semantic space to the point of being difficult to separate with a multidimensional plane. Meanwhile, DistillCLIP does not cater to this wider distribution and instead produces simpler representations, resulting in poorer zero-shot classification performance. However because of this, it is much easier to linearly separate its embeddings.

	Params			FLOPs		
Model	**Vision Model**	**Text Model**	**Total**	**Vision Model**	**Text Model**	**Total**
CLIP	87.5M	66.2M	151.3M	8.8 GFLOPs	6 GFLOPs	14.8 GFLOPs
DistillCLIP	21.7M	44.2M	66.1M	9.2 GFLOPs	3 GFLOPs	12.2 GFLOPs

Table 4.5: Size and speed of CLIP and DistillCLIP

4.4 Video Accident Recognition

The classification accuracy on the video accident recognition dataset are shown in Table 4.4. Although DistillCLIP only achieves 64.59%, the full-size CLIP attains 65.02%, resulting in 99.34% retention of teacher performance. Again we note that the goal is not the maximize student performance, but the ability of the student to match teacher performance.

4.5 Size and Speed Comparison

We compare the parameter counts and FLOPs of CLIP and DistillCLIP in Table 4.5. We observe that DistillCLIP has around 43.69% the parameters and 82.43% the FLOPs of the teacher CLIP. A large part of the parameter saving can be attributed to the use of ViT-S/16 student vision encoder, which is 24.8% the size of the teacher's ViT-B/32 vision encoder. However, we observe that the savings in FLOPs are instead more attributable to the text encoder, which has 50% the FLOPs of the teacher's. We believe that this is due to the ViT-S/16 using a patch size of 16, which creates a longer sequence of tokens during inference compared to the teacher.

CHAPTER V

CONCLUSION

We propose a knowledge distillation scheme for CLIP. Our method distills the inter and intra-modal similarity maps of the teacher with squared l_2 norm losses.

Table 5.1 summarizes our results. ↑ and ↓ respectively indicate "high better" and "lower better" for evaluation metrics. Results highlighted in blue and red indicate outcomes we found to be satisfactory and unsatisfactory, respectively. Our distilled CLIP, DistillCLIP, is successfully smaller and faster than the teacher CLIP by being 43.69% the teacher's size teacher and having 82.43% the FLOPs of the teacher. We observe that the zero-shot performance of DistillCLIP, our distilled CLIP, is sensitive to the granularity of class labels and better matches teacher performance when using conceptually easily separable coarse labels. As a result, DistillCLIP only retains 63.81% of teacher performance on zero-shot transfer averaged across our evaluation datasets. However, the embeddings produced by DistillCLIP are still comparably informative. On linear probe evaluation, DistillCLIP has a considerably smaller deficit compared to the teacher and can even surpass the teacher on some datasets, leading to a 100.53% retention rate averaged across evaluation datasets. DistillCLIP does not contain the robustness properties of CLIP, preserving only 12.28% of teacher performance averaged across two distribution shift datasets. This may be related to its better performance in linear probe evaluation compared to zero-shot transfer, although it is currently beyond the scope of this study to thoroughly investigate. Lastly, we show DistillCLIP can closely follow the performance of CLIP on a video accident recognition task with a 99.34% retention rate.

Future studies can further this work by investigating other possible distillation signals and losses, expanding the dataset benchmarks used in evaluations, evaluating on other vision-language tasks such as visual question answering, and exploring the role of dataset size during distillation.

	CLIP	DistillCLIP	Retention
Parameters ↓	151.3M	66.1M	43.69
FLOPs ↓	14.8 GFLOPs	12.2 GFLOPs	82.43
Zero-Shot Transfer ↑			
CIFAR-10 [28]	89.83	76.78	85.47
CIFAR-100 [28]	65.08	29.35	45.1
STL-10 [9]	97.13	84.3	86.79
Oxford-IIIT Pet [36]	87.21	3.38	3.88
Oxford-IIIT Pet (coarse) [36]	100.00	97.83	97.83
Average			63.81
Linear Probe Evaluation ↑			
CIFAR-10 [28]	94.82	95.71	101.25
CIFAR-100 [28]	79.43	81.86	102.81
STL-10 [9]	98.56	98.63	100.001
Oxford-IIIT Pet [36]	93.44	92.29	98.77
Oxford-IIIT Pet (coarse) [36]	99.93	99.73	99.80
Average			100.53
Robustness to Natural Distribution Shift ↑			
ImagenetV2 [41]	55.79	6.17	11.06
ImageNet-A [22]	31.37	4.55	14.5
Average			12.28
Video Accident Prediction ↑	65.02	64.59	99.34

Table 5.1: Summary of results.

REFERENCES

[1] Agrawal, A., Teney, D. and Nematzadeh, A. [2022], Vision-language pretraining: Current trends and the future, *in* ‘Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts’, pp. 38–43.

[2] Andonian, A., Chen, S. and Hamid, R. [2022], Robust cross-modal representation learning with progressive self-distillation, *in* ‘Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition’, pp. 16430–16441.

[3] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L. and Parikh, D. [2015], Vqa: Visual question answering, *in* ‘Proceedings of the IEEE international conference on computer vision’, pp. 2425–2433.

[4] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E. et al. [2021], ‘On the opportunities and risks of foundation models’, *arXiv preprint arXiv:2108.07258* .

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. [2020], ‘Language models are few-shot learners’, *Advances in neural information processing systems* **33**, 1877–1901.

[6] Chan, F.-H., Chen, Y.-T., Xiang, Y. and Sun, M. [2016], Anticipating accidents in dashcam videos, *in* ‘Asian Conference on Computer Vision’, Springer, pp. 136–153.

[7] Chen, T., Kornblith, S., Norouzi, M. and Hinton, G. [2020], A simple framework for contrastive learning of visual representations, *in* ‘International conference on machine learning’, PMLR, pp. 1597–1607.

[8] Cheng, R., Wu, B., Zhang, P., Vajda, P. and Gonzalez, J. E. [2021], Data-efficient language-supervised zero-shot learning with self-distillation, *in* ‘Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition’, pp. 3119–3124.

[9] Coates, A., Ng, A. and Lee, H. [2011], An analysis of single-layer networks in unsupervised feature learning, *in* 'Proceedings of the fourteenth international conference on artificial intelligence and statistics', JMLR Workshop and Conference Proceedings, pp. 215–223.

[10] Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L. and Raff, E. [2022], 'Vqgan-clip: Open domain image generation and editing with natural language guidance', *arXiv preprint arXiv:2204.08583* .

[11] Dai, W., Hou, L., Shang, L., Jiang, X., Liu, Q. and Fung, P. [2022], 'Enabling multimodal generation on clip via vision-language knowledge distillation', *arXiv preprint arXiv:2203.06386* .

[12] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. [2009], Imagenet: A large-scale hierarchical image database, *in* '2009 IEEE conference on computer vision and pattern recognition', Ieee, pp. 248–255.

[13] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. [2018], 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805* .

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. [2020], 'An image is worth 16x16 words: Transformers for image recognition at scale', *arXiv preprint arXiv:2010.11929* .

[15] Gao, T., Yao, X. and Chen, D. [2021], 'Simcse: Simple contrastive learning of sentence embeddings', *arXiv preprint arXiv:2104.08821* .

[16] Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V. and Grover, A. [2022], 'Cyclip: Cyclic contrastive language-image pretraining', *Advances in Neural Information Processing Systems* **35**, 6704–6719.

[17] Gong, Y., Liu, L., Yang, M. and Bourdev, L. [2014], 'Compressing deep convolutional networks using vector quantization', *arXiv preprint arXiv:1412.6115* .

[18] Gordon, M. A., Duh, K. and Andrews, N. [2020], 'Compressing bert: Studying the effects of weight pruning on transfer learning', *arXiv preprint arXiv:2002.08307* .

[19] Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J. and Bigham, J. P. [2018], Vizwiz grand challenge: Answering visual questions from blind people, *in* ‘Proceedings of the IEEE conference on computer vision and pattern recognition’, pp. 3608–3617.

[20] Gurari, D., Zhao, Y., Zhang, M. and Bhattacharya, N. [2020], Captioning images taken by people who are blind, *in* ‘European Conference on Computer Vision’, Springer, pp. 417–434.

[21] Han, S., Pool, J., Tran, J. and Dally, W. [2015], ‘Learning both weights and connections for efficient neural network’, *Advances in neural information processing systems* **28**.

[22] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J. and Song, D. [2021], Natural adversarial examples, *in* ‘Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition’, pp. 15262–15271.

[23] Hinton, G., Vinyals, O., Dean, J. et al. [2015], ‘Distilling the knowledge in a neural network’, *arXiv preprint arXiv:1503.02531* **2**(7).

[24] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z. and Duerig, T. [2021], Scaling up visual and vision-language representation learning with noisy text supervision, *in* ‘International Conference on Machine Learning’, PMLR, pp. 4904–4916.

[25] Jia, D., Han, K., Wang, Y., Tang, Y., Guo, J., Zhang, C. and Tao, D. [2021], ‘Efficient vision transformers via fine-grained manifold distillation’, *arXiv preprint arXiv:2107.01378* .

[26] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F. and Liu, Q. [2019], ‘Tinybert: Distilling bert for natural language understanding’, *arXiv preprint arXiv:1909.10351* .

[27] Kim, S., Gholami, A., Yao, Z., Mahoney, M. W. and Keutzer, K. [2021], I-bert: Integer-only bert quantization, *in* ‘International conference on machine learning’, PMLR, pp. 5506–5518.

[28] Krizhevsky, A., Hinton, G. et al. [2009], ‘Learning multiple layers of features from tiny images’.

[29] Lin, Z., Geng, S., Zhang, R., Gao, P., de Melo, G., Wang, X., Dai, J., Qiao, Y. and Li, H. [2022], Frozen clip models are efficient video learners, *in* 'European Conference on Computer Vision', Springer, pp. 388–404.

[30] Liu, Z., Wang, Y., Han, K., Zhang, W., Ma, S. and Gao, W. [2021], 'Post-training quantization for vision transformer', *Advances in Neural Information Processing Systems* **34**, 28092–28103.

[31] Loshchilov, I. and Hutter, F. [2017], 'Decoupled weight decay regularization', *arXiv preprint arXiv:1711.05101* .

[32] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. and Gebru, T. [2019], Model cards for model reporting, *in* 'Proceedings of the conference on fairness, accountability, and transparency', pp. 220–229.

[33] Mu, N., Kirillov, A., Wagner, D. and Xie, S. [2021], 'Slip: Self-supervision meets language-image pre-training', *arXiv preprint arXiv:2112.12750* .

[34] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I. and Chen, M. [2021], 'Glide: Towards photorealistic image generation and editing with text-guided diffusion models', *arXiv preprint arXiv:2112.10741* .

[35] Oord, A. v. d., Li, Y. and Vinyals, O. [2018], 'Representation learning with contrastive predictive coding', *arXiv preprint arXiv:1807.03748* .

[36] Parkhi, O. M., Vedaldi, A., Zisserman, A. and Jawahar, C. [2012], Cats and dogs, *in* '2012 IEEE conference on computer vision and pattern recognition', IEEE, pp. 3498–3505.

[37] Pham, H., Dai, Z., Ghiasi, G., Liu, H., Yu, A. W., Luong, M.-T., Tan, M. and Le, Q. V. [2021], 'Combined scaling for zero-shot transfer learning', *arXiv preprint arXiv:2111.10050* .

[38] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. [2021], Learning transferable visual models from natural language supervision, *in* 'International Conference on Machine Learning', PMLR, pp. 8748–8763.

[39] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M. [2022], 'Hierarchical text-conditional image generation with clip latents', *arXiv preprint arXiv:2204.06125* .

[40] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I. [2021], Zero-shot text-to-image generation, *in* 'International Conference on Machine Learning', PMLR, pp. 8821–8831.

[41] Recht, B., Roelofs, R., Schmidt, L. and Shankar, V. [2019], Do imagenet classifiers generalize to imagenet?, *in* 'International Conference on Machine Learning', PMLR, pp. 5389–5400.

[42] Ren, S. and Zhu, K. Q. [2022], 'Leaner and faster: Two-stage model compression for lightweight text-image retrieval', *arXiv preprint arXiv:2204.13913* .

[43] Sanh, V., Debut, L., Chaumond, J. and Wolf, T. [2019], 'Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter', *arXiv preprint arXiv:1910.01108* .

[44] Sharma, P., Ding, N., Goodman, S. and Soricut, R. [2018], Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 2556–2565.

[45] Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z. and Keutzer, K. [2021], 'How much can clip benefit vision-and-language tasks?', *arXiv preprint arXiv:2107.06383* .

[46] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M. and Kiela, D. [2022], Flava: A foundational language and vision alignment model, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 15638–15650.

[47] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J. and Beyer, L. [2021], 'How to train your vit? data, augmentation, and regularization in vision transformers', *arXiv preprint arXiv:2106.10270* .

[48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. [2017], 'Attention is all you need', *Advances in neural information processing systems* **30**.

[49] Wang, J., Wang, C., Wang, X., Huang, J. and Jin, L. [2023], 'Conaclip: Exploring distillation of fully-connected knowledge interaction graph for lightweight text-image retrieval', *arXiv preprint arXiv:2305.17652* .

[50] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N. and Zhou, M. [2020], 'Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers', *Advances in Neural Information Processing Systems* **33**, 5776–5788.

[51] Wang, Z., Codella, N., Chen, Y.-C., Zhou, L., Dai, X., Xiao, B., Yang, J., You, H., Chang, K.-W., Chang, S.-f. et al. [2022], 'Multimodal adaptive distillation for leveraging unimodal encoders for vision-language tasks', *arXiv preprint arXiv:2204.10496* .

[52] Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M. and Liu, T. [2022], Cris: Clip-driven referring image segmentation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 11686–11695.

[53] Wang, Z., Wang, W., Zhu, H., Liu, M., Qin, B. and Wei, F. [2021], 'Distilled dual-encoder model for vision-language understanding', *arXiv preprint arXiv:2112.08723* .

[54] Xie, N., Lai, F., Doran, D. and Kadav, A. [2019], 'Visual entailment: A novel task for fine-grained image understanding', *arXiv preprint arXiv:1901.06706* .

[55] Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C. et al. [2021], 'Florence: A new foundation model for computer vision', *arXiv preprint arXiv:2111.11432* .

[56] Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y. et al. [2022], Regionclip: Region-based language-image pretraining, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 16793–16803.

[57] Zhu, M., Tang, Y. and Han, K. [2021], 'Vision transformer pruning', *arXiv preprint arXiv:2104.08500* .

APPENDIX A
MODEL CARD

We present the model card [32] of DistillCLIP in Table A.1.

Model Details	
Organization developing model	ALIVE
Model date	June 2023
Model version	Version 1

<table>
<tr><td>Model type</td><td>Image encoder is a ViT-S/16. Text encoder is a 6-layer Transformer encoder. The final embedding size is 256.
A more detailed list of hyperparameters can be found below:
<table>
<tr><th>Hyperparameter</th><th>Value</th></tr>
<tr><td>batch size</td><td>84</td></tr>
<tr><td>optimizer</td><td>AdamW</td></tr>
<tr><td>learning rate</td><td>3e−5</td></tr>
<tr><td>weight decay</td><td>1e−1</td></tr>
<tr><td>AdamW β_1</td><td>0.9</td></tr>
<tr><td>AdamW β_2</td><td>0.98</td></tr>
<tr><td>AdamW e</td><td>1e−6</td></tr>
<tr><td>learning rate decay</td><td>cosine</td></tr>
<tr><td>warm-up steps</td><td>10,000</td></tr>
<tr><td>training steps</td><td>33,513</td></tr>
</table>
For model hyperparameters, please refer to Appendix B.</td></tr>
<tr><td>Paper or other resource for more information</td><td>DistillCLIP: Knowledge Distillation of Contrastive Language-Image Pretrained Models</td></tr>
<tr><td>Citation details</td><td>Ramos, P., Alampay, R., and Abu, P. [2023], ‘DistillCLIP: Knowledge Distillation of Contrastive Language-Image Pretrained Models’.</td></tr>
</table>

Where to send questions or comments about the model	patrick.john.ramos@obf.ateneo.edu
Intended Use	
Primary intended uses	Research on vision-language models e.g. natural language supervised image classification, visual question answering, text-to-image synthesis
Primary intended users	Researchers in the field of vision-language representation learning
Out-of-scope use cases	In-the-wild applications e.g. industrial deployment
Factors	
Relevant factors	The training data, Conceptual Captions 3M, may contain biases that may make performance different for members of different social groups.
Metrics	
Model performance measures	Classification accuracy
Variation approaches	We only distill CLIP once due to computational expenses.
Evaluation Data	

Datasets	CIFAR-10, CIFAR-100, STL-10, Oxford-IIIT Pet, ImageNetV2, ImageNet-A, Aadv
Training Data	
Training data	Conceptual Captions 3M
Quantitative Analyses	
Quantitative results	Please refer to Table 5.1.
Ethical Considerations	
Data	The training data, Conceptual Captions 3M, may contain biases that may make performance different for members of different social groups.

Table A.1: Model Card

APPENDIX B
MODEL ARCHITECTURE CONFIGURATIONS

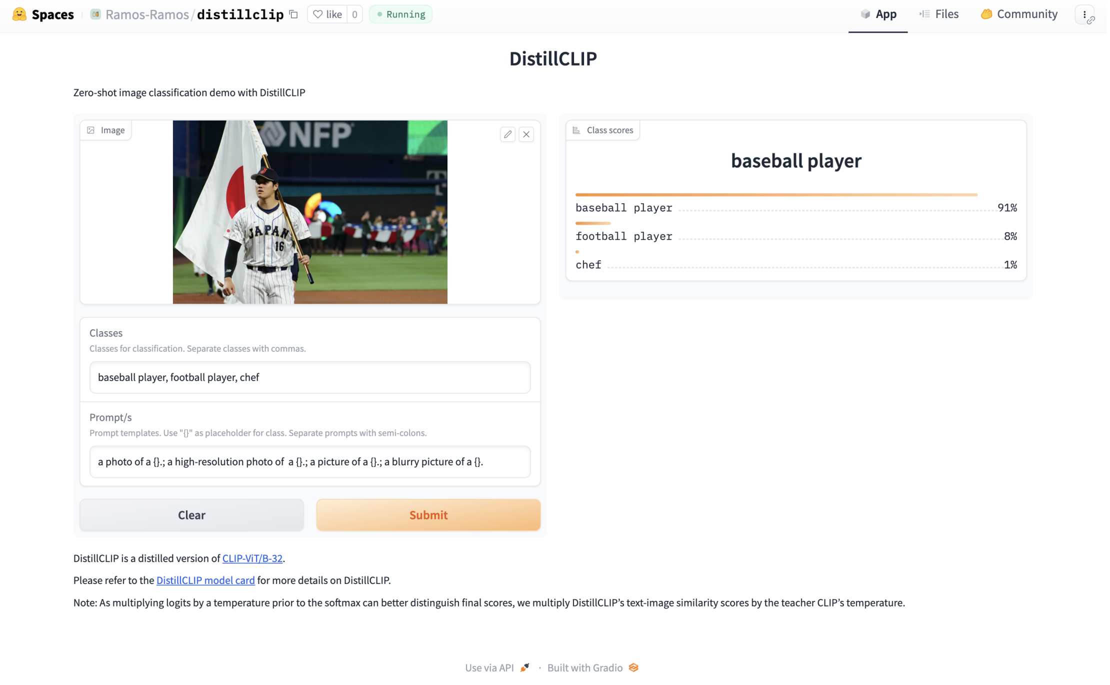
We compare the model architectures of the teacher and student models in Table B.1.

Hyperparameter		**CLIP-ViT-B/32 (teacher)**	**DistillCLIP (student)**
Vision encoder	layers	12	12
	hidden size	768	384
	intermediate size	3072	1536
	attention heads	12	6
	image size	224	224
	patch size	32	16
Text encoder	layers	12	6
	hidden size	512	512
	intermediate size	2048	2048
	attention heads	8	8
	sequence length	77	77
	vocabulary size	49408	49408
Projection size		512	256

Table B.1: Teacher and student model configurations.

APPENDIX C
WEB DEMO

We release a web demo for zero-shot image classification with DistillCLIP at https://huggingface.co/spaces/Ramos-Ramos/distillclip. An example of the demo interface in use is presented in Figure C.1. The user provides an image, classes separated by commas, and prompts separated by semi-colons. Given these inputs, the demo performs zero-shot image classification using the method discussed in Section 3.3.2. Image-text cosine similarity scores are softmaxed to create a probability distribution. As multiplying logits prior to softmax can create larger discrepancies between final scores without changing their order, we multiply the similarity scores by the temperature of the teacher CLIP ($e^{4.6052}$) before the softmax.



DistillCLIP

Zero-shot image classification demo with DistillCLIP

baseball player

baseball player	91%
football player	8%
chef	1%

Classes
Classes for classification. Separate classes with commas.
baseball player, football player, chef

Prompt/s
Prompt templates. Use "{}" as placeholder for class. Separate prompts with semi-colons.
a photo of a {}.; a high-resolution photo of a {}.; a picture of a {}.; a blurry picture of a {}.

DistillCLIP is a distilled version of CLIP-ViT/B-32.

Please refer to the DistillCLIP model card for more details on DistillCLIP.

Note: As multiplying logits by a temperature prior to the softmax can better distinguish final scores, we multiply DistillCLIP's text-image similarity scores by the teacher CLIP's temperature.

Figure C.1: DistillCLIP zero-shot image classification web demo.