CLIP-Lite: Information Efficient Visual Representation Learning with Language Supervision

Aman Shrivastava
University of Virginia

Ramprasaath R. Selvaraju
Salesforce Research

Nikhil Naik
Salesforce Research

Vicente Ordonez
Rice University

Abstract

We propose CLIP-Lite, an information efficient method for visual representation learning by feature alignment with textual annotations. Compared to the previously proposed CLIP model, CLIP-Lite requires only one negative image-text sample pair for every positive image-text sample during the optimization of its contrastive learning objective. We accomplish this by taking advantage of an information efficient lower-bound to maximize the mutual information between the two input modalities. This allows CLIP-Lite to be trained with significantly reduced amounts of data and batch sizes while obtaining better performance than CLIP at the same scale. We evaluate CLIP-Lite by pretraining on the COCO-Captions dataset and testing transfer learning to other datasets. CLIP-Lite obtains a +14.0% mAP absolute gain in performance on Pascal VOC classification, and a +22.1% top-1 accuracy gain on ImageNet, while being comparable or superior to other, more complex, text-supervised models. CLIP-Lite is also superior to CLIP on image and text retrieval, zero-shot classification, and visual grounding. Finally, we show that CLIP-Lite can leverage language semantics to encourage bias-free visual representations that can be used in downstream tasks. Implementation: https://github.com/4m4n5/CLIP-Lite

1 Introduction

Pretraining image classification networks on the Imagenet dataset has led to visual representations that transfer to other tasks (Girshick et al., 2014; Long et al., 2015; Vinyals et al., 2015; Antol et al., 2015; Zhu et al., 2016). However, such classification based pretraining requires a large amount of human-annotated data which is hard to obtain at scale. In contrast, captioned image data is an information-dense source of supervision that is relatively cheap to collect and plentiful on the internet. Therefore, recent methods have used joint vision-language pretraining to learn representations from image-caption pairs (Desai and Johnson, 2021; Sariyildiz et al., 2020). However, methods such as VirTex (Desai and Johnson, 2021) which train on complex language modeling tasks such as masked language modeling, token classification, and captioning fail to align features in a common latent space.

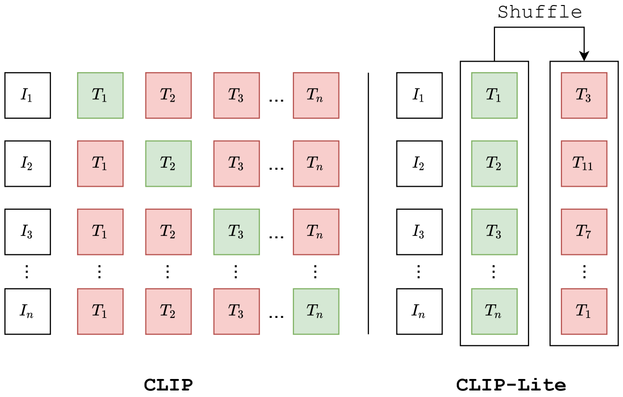

Figure 1: Given a batch of n image-caption pairs $\{(I_i, T_i)\}$, CLIP requires a large number of negative pairs $\{(I_i, T_j) \mid i \neq j\}$ due to the need to pair every image in the batch with captions from other images. Whereas, CLIP-Lite can learn representations using a single negative pair (in red) for every positive pair (in green).

Recently, CLIP (Radford et al., 2021), a vision-language pretraining model, was developed using contrastive learning between the two modalities on an Internet-sized dataset of 400 million image-caption pairs. Contrastive learning methods work by pulling closer the representations of independent views of the same datum *i.e.* a positive or matching image-caption pair and pushing apart the representations of independent views of different data *i.e.* negative or non-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

matching image-caption pairs. However, contrastive learning in vision-language pretraining still has some limitations as it seems to be most effective only with large scale data, and it requires a large number of negative image-caption pairs during training. Our work aims to address and explore these two limitations by proposing CLIP-Lite, an information efficient variation of CLIP that is useful even in smaller data regimes, does not rely in as many negative sample pairs during training, and provides comparable or superior performance on standard benchmarks against other methods trained at the same scale. Our work is motivated by the observation that multiple contrastive objectives maximize a lower-bound on the mutual information between two or more views of the same datum (Wu et al., 2020). CLIP particularly maximizes the mutual information between the image and its caption by using a mutual information lower bound based on InfoNCE (Oord et al., 2018). The InfoNCE bound has seen wide adoption due to its favorable properties such as stability and low variance. However, the the bound is theoretically loose in cases when the true mutual information is larger than $\log K$ where $(K-1)$ is the number of negative samples used for training. The negative pairs can be randomly sampled but usually a large amount of negative pairs are required to have a good estimate of the mutual information between the two input streams, and hence the need for rather large batch sizes (Bachman et al., 2019; Chen et al., 2015) or memory-banks (Chen et al., 2020b; Tian et al., 2019; He et al., 2020).

We instead adopt a lower bound based on Jenssen Shannon Divergence to maximize the mutual information (Hjelm et al., 2018; Nowozin et al., 2016), thus requiring no more than one negative example pair for each positive example pair. This reduces the number of negative examples in a training batch to $O(n)$, where n is the batch size. In contrast, CLIP uses $O(n^2)$ negative example pairs per batch. Figure 2 (right) illustrates this difference. We implement this strategy and demonstrate thoroughly the efficacy of CLIP-Lite through experiments on several tasks and datasets at various scales. Our method demonstrates impressive data efficiency and is able to outperform CLIP trained on the entire COCO-Captions dataset while only training on 20% of the same dataset. We also demonstrate that CLIP-Lite can be used as a good source of pretrained features by showing good generalization on Pascal VOC and Imagenet classification. We also show that the visual feature backbone of CLIP-Lite can be finetuned in the iNaturalist dataset to match top performances on this benchmark with caption supervision pretraining. Furthermore, we show that CLIP-Lite leads to good visual features for image retrieval compared to regular CLIP trained on COCO Captions. We also demonstrate that CLIP-Lite enables the removal of concepts from visual representations which we show can be applied in bias mitigation. Our work extends and complements the work using contrastive learning, especially addressing the computational requirements of the original CLIP model in terms of memory overhead through minimizing the number of negative sample image-text pairs required during training and shows its effectiveness in smaller data regimes including for zero-shot learning on CIFAR-10, image-text retrieval and unsupervised object localization.

2 Related Work

Our work is related to several strands of research on visual pretraining without full-supervision.

Vision-Language Pretraining: Research on learning visual representations by using textual labels or annotations has a long history. In (Quattoni et al., 2007), the authors learn data-efficient image representations using manifold learning in the weight space of classifiers trained to predict tokens in image captions. Following this work, (Joulin et al., 2016) used convolutional neural networks to predict words in image captions to learn image representations. This approach was later extended in (Lei Ba et al., 2015) where the model learns to predict phrase n-grams, which demonstrated impressive zero-shot performance on downstream classification tasks. Recently, VirTex (Desai and Johnson, 2021) used proxy language modeling tasks, such as image-captioning to train a visual encoder and a transformer based language decoder which generates captions. ICMLM (Sariyildiz et al., 2020) demonstrated a similar masked language modeling approach but relied on pretrained textual encoders for generating textual features. In (Stroud et al., 2020), video representations are learned using paired textual metadata, however the method does not extend to visual pretraining for images. In general, these methods distill the rich semantic information from a caption into the visual representation by learning to predict each token in the caption given the corresponding image. More recent work, such as CLIP (Radford et al., 2021), has shown that a simpler contrastive objective for aligning image and caption pairs is also able to learn a powerful visual representation. Our work extends CLIP using a more information-efficient approach.

Contrastive Representation Learning and Mutual Information Estimation: As demonstrated in (Wu et al., 2020), we observe that contrastive frameworks learn by maximizing the mutual information (MI) between different views of a given data point. For images, this is achieved by maximizing the MI between different augmentations of the data as in SimCLR (Chen et al., 2020a; Bachman et al., 2019). While for sequential data such as conversational text, consecutive utterances can be considered as different views (Stratos, 2018). Similarly, several other contrastive frameworks have been proposed that learn representations in domains such as images (Grill et al., 2020; Caron et al., 2020), text (Mikolov et al., 2013; Stratos, 2018), graphs (Veličković et al., 2018), and videos (Jabri et al., 2020). The value of mutual information is extremely challenging to estimate, especially for the high-dimensional continuous representations used in deep learning. To this end, various tractable lower-bounds on mutual information are used for optimization. Recently, MINE (Belghazi et al., 2018) proposed a general-purpose parameterized neural estimator of mutual information. It uses a Donsker-Varadhan (Donsker and Varadhan, 1983) representation of KL-divergence as the lower-bound on mutual information. MINE (Belghazi et al., 2018) used a neural network critic to distinguish positive and negative pairs of samples. Another popular bound on mutual information that has seen wide adoption due to its low variance is the InfoNCE (Oord et al., 2018) bound. In (Hjelm et al., 2018), the infoNCE bound on the mutual information is used for unsupervised representation learning. While it is used by several other methods for self-supervised (Chen et al., 2020a) representation learning for images. The capacity of the bound is limited by the number of contrastive samples used (McAllester and Stratos, 2020). Additionally, InfoNCE can underestimate large amounts of true MI which is generally the case with high-dimensional representations of natural images. To this end, DeepInfoMax (Hjelm et al., 2018) proposed using a lower-bound on mutual information that is based on the Jensen-Shannon Divergence (JSD) instead of the traditional KL-divergence (KLD). The authors show that the JSD based lower bound is stable, differentiable, and can be optimized with just one negative sample. Inspired by this, we extend the use of this bound for vision-language pretraining and demonstrate its effectiveness through extensive experimental evaluations.

3 CLIP-Lite

Given a dataset of image-caption pairs, the goal of our pretraining framework is to train an image encoder and a text encoder such that representations learned from the visual and the textual streams share maximum information (Figure 2 shows an overview). Consider an image encoder network, f_i with parameters θ_i and a textual encoder, f_t with parameters θ_t. Let (x_i, x_t) be a sampled image-caption pair from the dataset and $f_i(x_i)$ and $f_t(x_t)$ denote the representations extracted from the networks. Based on the information bottleneck principle (Tishby and Zaslavsky, 2015), the maximum mutual information (MI) predictive coding framework (Oord et al., 2018; Hjelm et al., 2018; McAllester and Stratos, 2020) aims to learn representations that maximize the MI between inputs and representations. In recent years, several methods (Chen et al., 2020a; He et al., 2020; Bachman et al., 2019) have used this principle to maximize MI between representations extracted from multiple views of a shared context. In the case of visual self-supervised learning, this is achieved by creating two independently-augmented copies of the same input and maximizing the MI between the respective features produced by an encoder. This framework can be extended further by considering an image x_i and its caption x_t as distinct views of the same input. This setup is motivated by the observation that image captions contain rich semantic information about images, for instance, presence of objects, location of objects, their relative spatial configurations, etc. Distilling this information into our visual representation is useful for robust representation learning (Radford et al., 2021). To this end, we formulate our objective as follows:

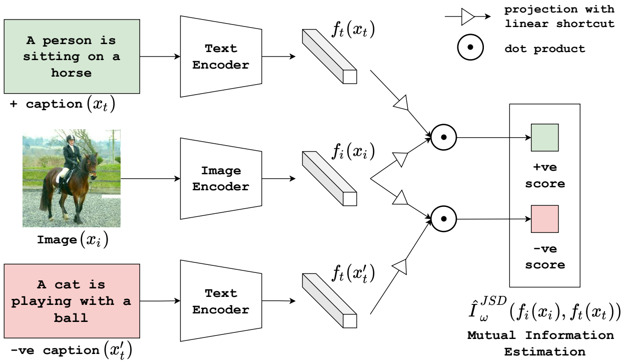
A person is sitting on a horse
+ caption (x_t)
Image (x_i)
A cat is playing with a ball
-ve caption (x'_t)
Text Encoder
Image Encoder
Text Encoder
$f_t(x_t)$
$f_i(x_i)$
$f_t(x'_t)$
projection with linear shortcut
dot product
+ve score
-ve score
$\hat{I}_{\omega}^{JSD}(f_i(x_i), f_t(x_t))$
Mutual Information Estimation

Figure 2: **CLIP-Lite:** We extract representations for an image, its positive caption, and one negative caption. Image-caption pairs are then fed into the mutual information discriminator function which outputs a score for each pair. These scores are then used to estimate and maximize mutual information using Jensen-Shannon Divergence (JSD) to optimize the parameters of the encoders and the mutual information discriminator end-to-end. The projection and dot function represents the MI discriminator function T_ω.

$$(\hat{\theta}_i, \hat{\theta}_t) = \underset{\theta_i, \theta_t}{\arg\max}\ I(f_i(x_i), f_t(x_t)), \tag{1}$$

where $I(f_i(x_i), f_t(x_t)) \leq I(x_i; x_t)$; due to the data processing inequality between visual and textual streams.

3.1 Mutual Information Maximization

For given random variables y and z, their mutual information is defined as a Kullback-Leibler (KL) divergence between their joint distribution $p(y, z)$ and the product of their marginal distributions, $p(y)p(z)$ as,

$$I(y; z) = D_{\text{KL}}(p(y, z) \,||\, p(y)p(z)). \tag{2}$$

However, mutual information is notoriously hard to estimate for high-dimensional continuous variables, especially when the distributions $p(y, z)$, $p(x)$, or $p(z)$ are not explicitly known. As a result, recent approaches use various tractable lower bounds on the mutual information which are differentiable and hence can be maximized with gradient-descent based optimization. For contrastive learning, a commonly used bound is infoNCE (Oord et al.,

2018) based on Noise-Contrastive Estimation (Gutmann and Hyvärinen, 2010). This bound is relatively more stable and has been shown to work in a wide variety of tasks (Chen et al., 2020a; Bachman et al., 2019; Chen et al., 2020b) including CLIP (Radford et al., 2021) which, similar to our method, aims to learn visual representations from textual annotations. The infoNCE bound has seen wider adoption as it demonstrates lower variance compared to the Donsker-Varadhan bound (Donsker and Varadhan, 1983). However, both of these bounds require a large number of negative samples and as a result, recent methods either train with extremely large batch-sizes (Radford et al., 2021; Chen et al., 2020a); or an additional memory-bank of negative samples (Chen et al., 2020b; Tian et al., 2020).

Unlike these works, we estimate mutual information using a Jensen-Shannon Divergence (JSD) bound, similar to formulations used for generative modeling (Nowozin et al., 2016); and source separation (Brakel and Bengio, 2017). This bound on mutual information is derived by replacing the KL-divergence in equation 2 with the Jensen-Shannon divergence (ref. appendix for further discussion). Interestingly, the lower bound derived as such is stable, differentiable, monotonically related to the mutual information $I(y;z)$, and most importantly, not dependent on the number of negative samples. Hence we have, $I(Y;Z) \geq \hat{I}_{\omega}^{JSD}(Y;Z)$ where,

$$\begin{aligned} \hat{I}_{\omega}^{JSD}(Y;Z) :=& \mathbb{E}_{P(Y,Z)}[-\log(1+e^{-T_{\omega}})] \\ &- \mathbb{E}_{P(Y)P(Z)}[\log(1+e^{T_{\omega}})], \end{aligned} \tag{3}$$

and $T_{\omega} : \mathcal{Y} \times \mathcal{Z} \rightarrow \mathbb{R}$ is a discriminator neural network with trainable parameters ω which are jointly optimized to distinguish between a paired-sample from a joint distribution (positive image-caption pair) and one pair from the product of marginals (negative image-caption pair). Therefore we are able to optimize our overall objective with just one negative sample as follows:

$$(\hat{\omega}, \hat{\theta}_i, \hat{\theta}_t) = \underset{\omega,\theta_i,\theta_t}{\operatorname{argmax}}\ \hat{I}_{\omega}^{JSD}(f_i(x_i), f_t(x_t)), \tag{4}$$

where the visual encoder is a convolution neural network, and features are extracted from the pre-classification layer of the network. The textual encoder is parameterized by a neural network that takes the caption as a string of textual-tokens and generates a one-dimensional representation.

4 Experiments

In this section, we describe the experiments that demonstrate the value of using textual captions for learning visual representations using CLIP-Lite. In our experiments, the CLIP-Lite architecture consists of a ResNet-50 image encoder and the BERT-base textual encoder and is trained on the COCO Captions (Chen et al., 2015) dataset. We evaluate the robustness of our visual encoder through the following downstream tasks which use the visual encoder (1) as a frozen feature extractor, or (2) as source of weight initialization for finetuning (ref. appendix). In addition, we also demonstrate the data efficiency of our method by evaluating performance on fractional datasets.

```
# image_encoder - CNN (eg. ResNet50)
# text_encoder - Transformer (eg. BERT)
# mi_discriminator - Project, Normalize and Dot
# I[n, h, w, c] - Batch of images
# T[n, l] - Batch of texts

# Extract image and text features
image_feats = image_encoder(I)
text_feats = text_encoder(T)

# Shuffle text features to get negative samples
text_feats_neg = shuffle(text_feats)

# Compute alignment scores using project, normalize and dot
positive_scores = mi_discriminator(image_feats, text_feats)
negative_scores = mi_discriminator(image_feats, text_feats_neg)

# MI Estimation / Loss function
loss = softplus(-1.0 * positive_score) + softplus(negative_score)
```

Figure 3: **CLIP-Lite:** Pytorch style pseudo-code for our pretraining framework.

4.1 Architecture and Training Details

In all experiments, we use a standard ResNet-50 (He et al., 2016) that takes in a 224 × 224 image and generates 2048-dimensional features at the pre-logit layer. For textual encoding, we use a transformer (Vaswani et al., 2017) model initialized using BERT$_{\text{base}}$ (Devlin et al., 2018) and use the output [CLS] token as the text representation. We use the COCO Captions dataset (Chen et al., 2015) which has 118K images with five captions per image. During training time we apply (1) random cropping, (2) color jittering, (3) random horizontal flips while interchanging the words 'left' and 'right' in the caption, and (4) normalization using the ImageNet image mean. We use SGD with momentum 0.9 (Sutskever et al., 2013; Polyak, 1964) and weight decay 10^{-4} wrapped in LookAhead (Zhang et al., 2019) with $\alpha = 0.5$, and 5 steps. We perform distributed training across 8 GPUs with batch normalization (Ioffe and Szegedy, 2015) per GPU with an overall batch size of 1024 images for 250K iterations. We use linear learning rate warmup (Goyal et al., 2019) for the first 10K iterations followed by cosine decay (Loshchilov and Hutter, 2016) to zero. Additionally, we train CLIP (Radford et al., 2021) on the COCO-dataset using an open-source implementation[1] with the originally recommended (Radford et al., 2021) training schedule that suit smaller datasets, reasonable batch-sizes, and compute resources. Specifically, we train using the Adam Optimizer (Kingma and Ba, 2014) with decoupled weight decay regularization (Loshchilov and Hutter, 2016) for all weights except gains or biases.

[1] https://github.com/mlfoundations/open_clip

We train with a batch-size of 1024 and warm-up to an initial learning rate of 10^{-4} in 10K steps and decay to zero with the cosine schedule. We found that the performance slightly improves with longer training therefore we train for 250K iterations, similar to ours. All other training details and hyper-parameters were kept the same as the original work (Radford et al., 2021). Please note that our ResNet-50 based CLIP-COCO model outperforms (+1.2% Zero-shot Acc. on CIFAR10) publicly available weights[2], refer to appendix for further details on CLIP-COCO training.

4.2 Mutual Information Discriminator

As described in main paper, our JSD-based lower-bound on mutual information relies on a discriminator function, $T_\omega : \mathcal{Y} \times \mathcal{Z} \rightarrow \mathbb{R}$, which distinguishes between samples extracted from the joint distribution, $P(Y, Z)$ i.e. a positive image-caption pair and the product of marginals, $P(Y)P(Z)$ i.e. a negative image-caption pair. This discriminator function can be modelled as an arbitrary neural network with parameters ω that can be jointly optimized with the encoders during training (Belghazi et al., 2018). In this work, we use a projection and alignment based architecture similar to the one presented in Deep InfoMax (Hjelm et al., 2018).

Given a pair of input one-dimensional representations, both vectors are first projected using a projection module with two linear layers separated by a ReLU and a linear shortcut. A dot-product of these projections is then computed to get alignment scores. The projection function maps these representations to an aligned cross-modal latent space. Separate projection functions are used for image and text representations. Positive and negative pairs of image-text representations are passed through the discriminator to get respective scores which are then used to estimate and maximize mutual information using our objective. This architecture, in addition to being simple and computationally inexpensive, also offers alignment of the representations into a common cross-modal latent space which uses cosine similarity as the distance metric.

4.3 Transfer Learning with Frozen Backbone

In these experiments, we train linear models on frozen visual backbones pretrained using CLIP-Lite and compare with other pretraining methods on PASCAL VOC (Everingham et al., 2010) and ImageNet-1k (Russakovsky et al., 2015) classification problems.

PASCAL VOC linear classification: For this experiment, our setup is identical to VirTex (Desai and Johnson, 2021). We train on VOC07 trainval split (9K images, 20 classes) and report mAP on the test split. For classification, we train per-class SVMs on 2048-dimensional global average pooled features extracted from the last layer of our trained visual encoder. For each class, we train SVMs for cost values $C \in \{0.01, 0.1, 1, 10\}$ and select best C by 3-fold cross-validation.

Table 1: **Frozen Backbone Results:** On Pascal VOC07 and Imagenet-1k classification, CLIP-Lite outperforms baseline CLIP when evaluated using linear classifiers trained on top of frozen backbone networks pretrained on the COCO Dataset. CLIP-Lite's performance is competitive with more complex vision-language models. CLIP-Lite also performs better than supervised and self-supervised models trained on COCO images, without captions (ref. supplemental materials for additional results).

Method	**# images**	**Annotations**	**VOC07**	**IN-1k**
COCO-Sup.	118K	labels	86.2	46.4
MoCo-COCO	118K	self-sup.	67.5	46.5
ICMLM	118K	captions	87.5	47.9
VirTex	118K	captions	**88.7**	<u>53.8</u>
CLIP-COCO	118K	captions	74.2	33.2
CLIP-Lite	118K	captions	<u>88.2</u>	**55.3**

Imagenet-1k linear classification: For this experiment, our setup is identical to VirTex (Desai and Johnson, 2021). We train on the ILSVRC 2012 train split and report top-1 accuracy on val split. We train a linear classifier (fully connected layer + softmax) on 2048-dimensional global average pooled features extracted from the last layer of the visual backbone. For training, we use a batch-size of 256 for 100 epochs. We use SGD with momentum 0.9 and weight decay 0. The learning rate schedule is decayed by 0.1 after 60 & 80 epochs with an initial LR of 30.

Results: We compare CLIP-Lite to supervised, self-supervised and textually-supervised models in Table 1. CLIP-Lite significantly outperforms baseline CLIP when trained with the same amount of data on both tasks. When compared to other image-caption pretraining methods, CLIP-Lite performs competitively with VirTex (Desai and Johnson, 2021) on VOC2007 and outperforms both VirTex (Desai and Johnson, 2021) and ICMLM (Sariyildiz et al., 2020), which are trained on relatively complex language modeling tasks, on Imagenet classification. In addition, different from them, our method also generates a shared latent space that encodes both image and text modalities and enables cheap computation of cross-modal alignment, which enables additional downstream tasks such as zero-shot retrieval, and zero-shot transfer. It also allows us to find subspaces associated with abstract concepts that are better expressed with language than with visual examples, which allows for applications in bias mitigation through the synthesis of gender-neutral image representa-

[2] https://github.com/revantteotia/clip-training/blob/main/zero_shot_eval_output/coco_trained_clip_observations.md

Table 2: **Data Efficiency:** CLIP-Lite is more data efficient than CLIP, as shown in this experiment where we pretrain on $\{25, 50, 75, 100\}\%$ of the COCO Captions dataset and evaluate the models on VOC and ImageNet classification tasks with a frozen backbone. CLIP-Lite trained with just 25% of COCO already surpasses CLIP trained on the whole dataset.

	# images	**VOC07**	**IN-1k**
CLIP COCO-100%	118K	74.2	33.2
CLIP-Lite COCO-25%	29.5K	77.7$_{+3.5}$	45.1$_{+11.9}$
CLIP-Lite COCO-50%	59K	84.4$_{+10.2}$	51.3$_{+18.1}$
CLIP-Lite COCO-75%	88.5K	86.8$_{+12.6}$	53.2$_{+20.0}$
CLIP-Lite COCO-100%	118K	88.2$_{+14.0}$	55.3$_{+22.1}$

tions. CLIP-Lite also outperforms a fully-supervised model trained with COCO image labels, showing that it learns a better visual representation from information-dense captions as compared to training with labels alone. Additional results in the supplement show that CLIP-Lite is comparable or better than image-only SSL learning models trained on ImageNet, even though it is trained on much fewer images, albeit with textual supervision.

Data Efficiency: Due to our information-efficient approach for mutual information maximization, CLIP-Lite should be able to learn effective feature representations without requiring as much pretraining data as CLIP. To evaluate this claim, we train ResNet-50 backbones with our pretraining setup on multiple fractional subsets of the COCO Captions dataset and measure their downstream performance on both VOC and ImageNet classification tasks. As demonstrated in Table 2, CLIP-Lite outperforms the original CLIP training objective on VOC with 20% and on Imagenet with just 10% of the data, while obtaining a substantial improvement when both are trained with 100% data. Additionally, when compared with Virtex, CLIP-Lite performs competitively on VOC while being consistently better on Imagenet-1k.

4.4 Transfer Learning with Backbone Finetuning

Next, we evaluate the performance of of our visual backbone when the entire network is finetuned for the downstream task. For this purpose, we perform fine-grained classification on the iNaturalist 2018 (Van Horn et al., 2018) dataset, which contains images from $8,142$ fine-grained categories, with a long-tailed distribution. We train with the 'train2018' split and evaluate in the 'val2018' split. We finetune pretrained ResNet-50 models with a linear layer, using SGD with momentum 0.9 and weight decay 10^{-4} for 100 epochs. Initial learning rate is set to 0.025, which is reduced by $10\times$ at epochs 70 and 90. We use a batch size of 256 distributed across 8 GPUs.

Results: We summarize our results in Table 3. CLIP-Lite is competitive with supervised and self-supervised learning models trained with images alone even those trained with 5-10x more images. Its performance matches closely a model trained with full-supervision on 50% of the ImageNet (Krizhevsky et al., 2012) dataset, equal to $5.4\times$ the number of images as our pretraining dataset. Finally, CLIP-Lite obtains a 1.3% improvement over CLIP-COCO, while being competitive with VirTex.

Table 3: **Backbone Finetuning Results:** CLIP-Lite outperforms CLIP-COCO on iNaturalist, and performs comparably to VirTex. (IN-Sup. = ImageNet-supervised.)

Method	**# images**	**Annotations**	**iNat 18**
Random Init	-	-	61.4
IN-sup	1.28M	labels	65.2
IN-sup-50%	640K	labels	63.2
IN-sup-10%	128K	labels	60.2
MoCo-COCO	118K	self-sup.	60.5
MoCo-IN	1.28M	self-sup.	63.2
VirTex	118K	captions	63.4
CLIP-COCO	118K	captions	61.8
CLIP-Lite	118K	captions	63.1

4.5 Image-Text and Text-Image Retrieval

Our method is expected to produce effective representations for the task of image-text retrieval as it is trained by aligning text and image representations. We evaluate the image-text retrieval capabilities of CLIP-Lite on the validation set of COCO and the test split of Flickr30k (Young et al., 2014) datasets, following CLIP. We perform zero-shot image-text and text-image retrieval by ranking image-text pairs by their alignment score, which is the dot product of the normalized representations in the shared latent space. This ability to perform zero-shot retrieval is a salient feature of our and CLIP-like methods over previously proposed works that rely on language modeling tasks.

Results: Table 4 shows that CLIP-Lite substantially outperforms CLIP-COCO on all metrics for both text and image retrieval. The performance improvement is large both when evaluated on the COCO validation set, which is similar to the the COCO-Captions training split used for CLIP-Lite training; and when testing zero-shot on unseen text vocabulary and object categories of Flickr30K. Taken together, these results show that CLIP-Lite learns a superior representation for retrieval tasks as compared to CLIP, when trained on same amounts of data.

4.6 Zero-Shot Transfer

We use the cross-modal alignment capability of CLIP-Lite to perform zero-shot classification on unseen datasets

Table 4: **Retrieval Results:** CLIP-Lite substantially outperforms CLIP-COCO and the baseline Visual N-grams (Li et al., 2017) approach. CLIP-Lite is superior when evaluated on the COCO test split, which is similar to the CLIP-Lite training set and on Flickr30K, generalizing to unseen images and text in a zero-shot manner.

	Text Retrieval						**Image Retrieval**					
	Flickr30k			**MSCOCO**			**Flickr30k**			**MSCOCO**		
Method	*R@1*	*R@5*	*R@10*	*R@1*	*R@5*	*R@10*	*R@1*	*R@5*	*R@10*	*R@1*	*R@5*	*R@10*
Visual N-Grams	15.4	35.7	45.1	8.7	23.1	33.3	8.8	21.2	29.9	5.0	14.5	21.9
CLIP-COCO	19.9	41.9	54.9	18.9	42.9	54.6	13.9	33.0	43.8	13.9	33.5	44.2
CLIP-Lite	**28.8**	**55.8**	**67.4**	**26.0**	**54.6**	**68.0**	**23.1**	**51.1**	**62.9**	**20.2**	**48.1**	**62.2**

CIFAR-10, CIFAR100 (Krizhevsky et al., 2009), ImageNetV2 (Recht et al., 2019), and ImageNet-A (Hendrycks et al., 2021). Our model generates a shared latent space where we can readily compute the alignment between given (image, text) pairs as the cosine similarity of their representations. Therefore, we use the names of the classes to generate a textual description of each class label (class prompt). In this experiment, we use templates such as, "a photo of a {class name}" to generate such class prompts, following CLIP (Radford et al., 2021). Please refer to the appendix for comparison between different templates for generating the prompts. For a given image, we compute its alignment with each of the class prompts which are then normalized into a probability distribution via a softmax.

Results: Our results for the zero-shot transfer task on unseen datasets are compiled in table 5. Given the zero-shot nature of the task, CLIP-Lite obtains satisfactory performance on the complex ImageNet evaluations while clearly outperforming CLIP trained with the same amount of data in all settings.

4.7 Evaluating Visual Grounding

Next, we evaluate the capability of CLIP-Lite to localize a region in the image that corresponds to a given textual description. We compute the dot-product of the visual and textual embedding and compute its gradients with respect to the last convolutional layer of ResNet. We global average pool these gradients and perform a weighted sum with the last convolutional activations and clip the negative values to obtain Grad-CAM (Selvaraju et al., 2017). We then use the areas highlighted by Grad-CAM to approximate a predicted bounding box. We evaluate this experiment on the RefCOCO+ (Yu et al., 2016) dataset. We note that the images in the RefCOCO+ dataset are extracted from the training set of the COCO (Chen et al., 2015) dataset which our model uses for pretraining. Therefore, we view this evaluation as an explorative study to establish that our model is focusing on the relevant areas of the image while computing the alignment score with the caption.

RefCOCO+ results can be seen in the table to the right. CLIP-Lite significantly outperforms CLIP on all settings.

Table 5: **Zero Shot Transfer:** CLIP-Lite obtains satisfactory zero-shot transfer to unseen datasets.

	CLIP-COCO		**CLIP-Lite**	
Dataset	*Top1*	*Top5*	*Top1*	*Top5*
CIFAR10	16.3	68.9	**33.0**	**82.7**
CIFAR100	2.9	12.4	**6.8**	**33.1**
ImageNet-V2	4.4	11.1	**9.9**	**21.4**
ImageNet-A	1.7	7.3	**3.8**	**14.9**

Method	Val-acc	TestA-acc	TestB-acc
CLIP-COCO	29.1	28.5	28.5
CLIP-Lite (ours)	**36.1**	**41.4**	**32.0**

Qualitative results in Figure 4 demonstrate that even though the network has not been trained with any localization supervision, it is surprisingly good at localizing phrases in the image. For instance, in Figure 4 bottom left, for the phrase "blue", the network attends to all blue regions in the player's outfit. Interestingly, it is also able to localize abstract concepts as "blurry player".

4.8 Editing Concepts from Image Representations

One salient feature of CLIP-like methods, which other methods such as VirTex (Desai and Johnson, 2021) and ICMLM (Sariyildiz et al., 2020) lack, is that they are able to generate a shared latent space that encodes both image and text modalities. This enables us to find representations and subspaces associated with abstract concepts that are better expressed with language than with visual examples. Using this property, we demonstrate a methodology to remove concepts from visual representations. For instance, it is non trivial and even problematic to collect visual examples that capture the concept of gender, while it is relatively straightforward to express this concept in a sentence using language. Therefore, we can identify the gender subspace in our shared embedding space using text and use it to remove variance along this direction to smooth out the concept of gender from image representations. We

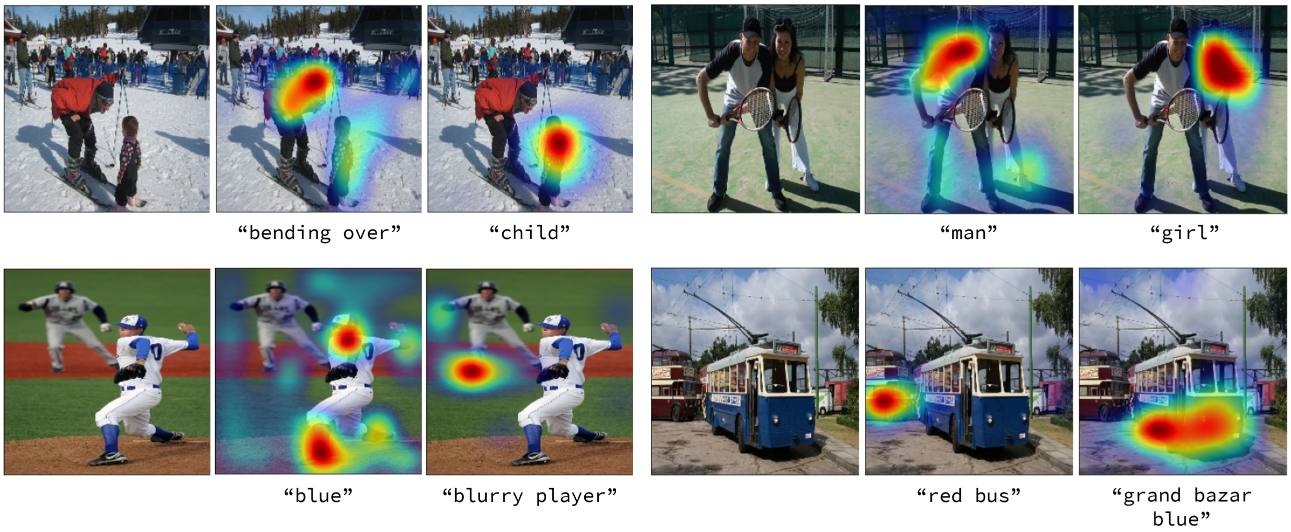

Figure 4: **Visual Grounding on RefCOCO+:** CLIP-Lite is able to localize textual descriptions to relevant areas in the image, shown here through Grad-CAM visualization using the alignment score with the mentioned textual description. *Top left:* CLIP-Lite is able to localize the action phrases such as "bending over". This demonstrates the value of learning from semantically rich textual captions.

motivate this experiment in the growing body of literature regarding bias mitigation, where the objective is to build invariant representations with respect to sensitive or protected attributes (Wang et al., 2019, 2020). In comparison to our work other methods require retraining the models to obtain invariant bias representations through adversarial learning (Wang et al., 2019) or effectively combining domain independent classifiers (Wang et al., 2020).

Identifying the Concept Subspace: The first step of our approach is to isolate the direction in the embedding space that captures maximum gender variance. For this purpose, we follow a strategy similar to Bolukbasi et al. (Bolukbasi et al., 2016) that deals with debiasing word representations. For characterizing features for male and female genders, we use word pairs (*man, woman*), (*son, daughter*) that indicate opposite genders. Now, consider a dataset $\mathcal{D} = \{(w_m, w_f)\}_{i=1}^{m}$ where each entry (w_m, w_f) is a tuple of opposite gendered words. Intuitively, each tuple should contain words that have the same meaning if not for the target attribute. To make the set $\mathcal{D}$ more robust, we used the sentence contextualization strategy presented in Liang et al. (Liang et al., 2020). In this step, the predefined sets of gendered tokens in the set, $\mathcal{D}$, are used to generate paired sentences which have the same meaning except for the gender attribute. We perform this contextualization by using simple sentence templates such as "I am a [word]" where [word] can be replaced with the word pairs in our dataset $\mathcal{D}$ to give, for instance, (*"I am a boy."*, *"I am a girl."*). Hence, we obtain a contextualized bias attribute dataset $\mathcal{S} = \{(s_m, s_f)\}_{i=1}^{n}$ where each entry is a tuple of semantically similar sentences with opposite genders. We extract the sentence representations for all entries in the set $\mathcal{S}$ by passing them through our pretrained text encoder and then projecting them to the shared latent space using the projector trained with our mutual information discriminator T_ω. We define sets $\mathcal{R}_m$ and $\mathcal{R}_f$ that contain sentence representations of the male and the female category, for example, $\mathcal{R}_m = \{F_t(s_m)\}_{i=1}^{n}$ where $F_t(.)$ is the sequential combination of our pretrained text-encoder and text-projection functions. Now we estimate the gender subspace $V = \{v_1, ..., v_k\}$ using the Principal Component Analysis corresponding mean shifted representation from both sets as described in (Liang et al., 2020).

Table 6: **Concept Editing Results:** We compute the mean alignment scores for the top 10 images queried using prompts that either contain male or female gendered tokens. The images are queried using gendered and neutralized representations. We observe that after gender-deletion the alignment score for images with men and women converge to similar values.

	Images with Men			Images with Women		
	gendered	neutral	delta	gendered	neutral	delta
Male queries	0.085	0.069	+0.016	0.057	0.067	-0.010
Female queries	0.042	0.068	-0.026	0.089	0.062	+0.027

Removing Concept from Image Representations: After estimating the gender subspace in our shared cross-modal latent space, we extend the hard debias algorithm (Bolukbasi et al., 2016) to edit visual representations. This is achieved by first projecting the representation onto the bias subspace, this projection is then subtracted from the original representation to give the de-gendered representation. Given an image, we first encode the image onto our multi-modal shared latent space to get, say, h. Now, consider the identified gender subspace V, we first com-

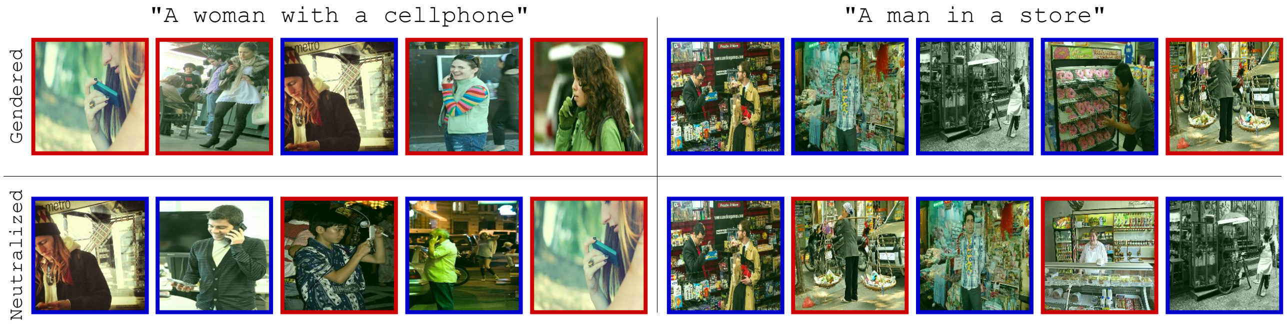

Figure 5: **Demonstrating Neutral Representations:** Qualitative demonstration of our concept editing method. For each text prompt, most aligned images are retrieved from male and female buckets of the gendered COCO subset before (top row) and after (bottom row) gender smoothing. Once representations are gender-neutralized the gendered references in the query become irrelevant and the image is only retrieved based on its remaining contents. Alignment score decreases from left to right for each set of queried images. Boundary color denotes perceived image gender; red for female, blue for male.

pute the projection of h onto this gender subspace V to get $h_V = \sum_{j=1}^{k} \langle h, v_j \rangle v_j$. We subtract this projection from the original representation to get a vector, $\hat{h} = h - h_V$ that is orthogonal to the bias subspace and therefore does not encode the target bias.

Analysis: To evaluate concept editing, we use the gendered subset of COCO-Captions (Wang et al., 2019; Zhao et al., 2017) for studying bias. The gender labels for images in the COCO dataset are derived from the captions. We obtain a subset from the COCO dataset with $16,225$ images with men and $6,601$ images with women. We use 10 sentences with male references and 10 sentences with female references from the set $\mathcal{S}$ and use them as prompts for this study. For each gendered prompt, we query the top 10 images independently from the male and the female image sets using both biased and debiased representations to compute alignment with the prompt. The mean alignment scores are then computed for each set given the prompt. Table 6 shows that the alignment scores roughly equalize for members of the two groups after removing the variance along the gender direction from the visual representations which indicates the invariance of the visual representations to gendered language tokens.

5 Limitations and Broader Impacts

CLIP-Lite trains the visual encoder by maximizing the mutual information between images and their captions. We observe that language supervision provides rich semantic density which can be distilled into visual representations. The visual encoder is encouraged to learn visual representations that encode maximum information from captions. As such, the visual encoder is only aware of concepts and objects that human-annotators have mentioned in the captions. Therefore the visual encoder lags behind task-specific models that are trained specifically for a given fine-grained task. For instance, visual encoders trained with CLIP-Lite struggle with relatively contextual downstream tasks that involve reading text or counting number of objects in an image. In this work, we train CLIP-Lite on the COCO-Captions (Chen et al., 2015) dataset which has high-quality curated captions for images. However, when trained on datasets with text paired with images from the internet, the textual captions can be significantly unfiltered and noisy. Our method essentially learns by aligning the caption and text representations. Therefore, the model is susceptible to learning harmful biases that are represented in the captions. Hence, deployment of visual backbones trained with CLIP-Lite and other pretraining methods which use natural language supervision need to be analyzed specifically for such biases. In this work, we present an approach to edit concepts from visual representations using the shared vision-language latent space learnt by our method. For instance, we demonstrate this capability by editing visual representations such that they are invariant to gendered tokens in language. However, further explorations are required to develop this concept editing mechanism further.

6 Conclusion

We introduced CLIP-Lite an image-text pretrained model using contrastive learning that leverages a different objective than the CLIP model that allows for it to be more data efficient. CLIP-Lite's objective is insensitive to the number of negative samples and hence can be trained with just one negative image-caption pair and shows superior results on lower data regimes while still demonstrating some of the most remarkable capabilities of the original CLIP model such as transferable features, zero-shot capabilities, and a shared latent space. Additionally, we present a concept editing methodology for neutralizing visual representations with respect to a chosen abstract concept. Please refer to the supplement for a detailed discussion on limitations and potential impact of our approach.

Acknowledgments This work was supported by NSF Awards IIS-2221943 and IIS-2201710, and through gift funding from a Salesforce AI Research Grant.

References

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*, pages 2425–2433. 1

Bachman, P., Hjelm, R. D., and Buchwalter, W. (2019). Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*. 2, 3, 4

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. (2018). Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR. 3, 5, 13

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357. 8

Brakel, P. and Bengio, Y. (2017). Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*. 4

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*. 2

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR. 2, 3, 4

Chen, X., Fan, H., Girshick, R., and He, K. (2020b). Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*. 2, 4

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*. 2, 4, 7, 9, 14

Desai, K. and Johnson, J. (2021). Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173. 1, 2, 5, 7

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*. 4, 14

Donsker, M. D. and Varadhan, S. S. (1983). Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212. 3, 4

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338. 5, 14

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587. 1

Goyal, P., Mahajan, D., Gupta, A., and Misra, I. (2019). Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6391–6400. 4

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*. 2

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304. JMLR Workshop and Conference Proceedings. 4

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738. 2, 3

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 4, 14

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*, pages 15262–15271. 7

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*. 2, 3, 5, 13

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR. 4, 15

Jabri, A., Owens, A., and Efros, A. A. (2020). Space-time correspondence as a contrastive random walk. *arXiv preprint arXiv:2006.14613*. 2

Joulin, A., Van Der Maaten, L., Jabri, A., and Vasilache, N. (2016). Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer. 2

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 4, 14

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. 7

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105. 6

Lei Ba, J., Swersky, K., Fidler, S., et al. (2015). Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4247–4255. 2

Li, A., Jabri, A., Joulin, A., and van der Maaten, L. (2017). Learning visual n-grams from web data. *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192. 7

Liang, P. P., Li, I. M., Zheng, E., Lim, Y. C., Salakhutdinov, R., and Morency, L.-P. (2020). Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*. 8

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440. 1

Loshchilov, I. and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*. 4, 14

McAllester, D. and Stratos, K. (2020). Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR. 3

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 2

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 271–279. 2, 4

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*. 2, 3, 13

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17. 4

Quattoni, A., Collins, M., and Darrell, T. (2007). Learning visual representations using images with captions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE. 2

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. *arXiv preprint arXiv:2103.00020*. 1, 2, 3, 4, 5, 7, 14, 15

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR. 7

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252. 5

Sariyildiz, M. B., Perez, J., and Larlus, D. (2020). Learning visual representations with caption annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 153–170. Springer. 1, 2, 5, 7

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626. 7

Stratos, K. (2018). Mutual information maximization for simple and accurate part-of-speech induction. *arXiv preprint arXiv:1804.07849*. 2

Stroud, J. C., Lu, Z., Sun, C., Deng, J., Sukthankar, R., Schmid, C., and Ross, D. A. (2020). Learning video representations from textual web supervision. *arXiv preprint arXiv:2007.14937*. 2

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR. 4

Tian, Y., Krishnan, D., and Isola, P. (2019). Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*. 2

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. (2020). What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*. 4

Tishby, N. and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE. 3

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2018). The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778. 6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).

Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008. 4, 14

Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., and Hjelm, R. D. (2018). Deep graph infomax. *arXiv preprint arXiv:1809.10341*. 2

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164. 1

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., and Ordonez, V. (2019). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319. 8, 9

Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., and Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928. 8

Wu, M., Zhuang, C., Mosse, M., Yamins, D., and Goodman, N. (2020). On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*. 2

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78. 6

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. (2016). Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer. 7

Zhang, M. R., Lucas, J., Hinton, G., and Ba, J. (2019). Lookahead optimizer: k steps forward, 1 step back. *arXiv preprint arXiv:1907.08610*. 4

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*. 9

Zhu, Y., Groth, O., Bernstein, M., and Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004. 1

A Appendix

This appendix is organized as follows:

- **A.1.** Discussion on the JSD-based lower bound on MI
- **A.2.** Comparison with Self-Supervised and other pre-training methods
- **A.3.** Mutual Information Discriminator
- **A.4.** Ablations on (1) batch sizes, (2) visual encoders, (3) textual encoders, (3) Zero-shot templates
- **A.5.** Training CLIP on the COCO-Captions dataset

A.1 Discussion on JSD-based lower bound on Mutual Information

Recall that for given random variables y and z, their mutual information is defined as a Kullback-Leibler (KL) divergence between their joint distribution $p(y, z)$ and the product of their marginal distributions, $p(y)p(z)$ as, $I(y; z) = D_{\text{KL}}(p(y, z) \,||\, p(y)p(z))$. The above formulation of MI gives rise to the commonly used contrastive objective InfoNCE (Oord et al., 2018). Alternatively, the KL-divergence can be replaced with the Jensen-Shannon divergence (JSD) between the joint and the product of marginals as an estimate of the Pointwise Mutual Information(PMI) between two views of the data i.e. $I^{JSD}(y; z) = D_{\text{JSD}}(p(y, z) \,||\, p(y)p(z))$. And as discussed in Hjelm et al. (2018), this formulation of MI leads to the following relation,

$$JSD(p(y,z)||p(y)p(z)) \propto \mathbb{E}_{y\sim p(y)}\left[\mathbb{E}_{z\sim p(z|y)}\left[\log\frac{p(z|y)}{p(z)} - (1+\frac{p(z)}{p(z|y)})\log\left(1+\frac{p(z|y)}{p(z)}\right)\right]\right] \tag{5}$$

Now, the quantity inside the expectation above is a concave, monotonically increasing function of the ratio $p(z|y)/p(z)$, which is exactly the exponential of the Pointwise Mutual Information, i.e. $e^{PMI(y,z)}$.

A.2 Comparison with SSL Pretraining Methods

In this section, we evaluate the performance of our method against other pre-training frameworks and image-only SSL methods. We observe that CLIP-Lite is comparable or better to image-only SSL learning models trained on downstream ImageNet classification with a frozen ResNet-50 backbone, even though our method is trained on much fewer images, albeit with textual supervision.

Table 7: CLIP-Lite outperforms CLIP-COCO on both VOC and ImageNet classification tasks, and performs comparably to VirTex. CLIP-Lite's performance is comparable or superior to both supervised and self-supervised learning models trained with images alone, even those trained with 10x more images. (IN-Sup. = ImageNet-supervised.)

Method	# images	Annotations	VOC07	IN-1k
COCO-Sup.	118K	labels	86.2	46.4
IN-Sup.	1.28M	labels	87.6	75.6
MoCo-COCO	118K	self-sup.	67.5	46.5
MoCo-IN v1	1.28M	self-sup.	79.4	60.8
PCL v1	1.28M	self-sup.	83.1	61.5
SwAV (200 ep.)	1.28M	self-sup.	87.9	72.7
ICMLM	118K	captions	87.5	47.9
VirTex	118K	captions	88.7	53.8
CLIP-COCO	118K	captions	74.2	33.2
CLIP-Lite	118K	captions	88.2	55.3

A.3 Mutual Information Discriminator

As described in main paper, our JSD-based lower-bound on mutual information relies on a discriminator function, $T_\omega : \mathcal{Y} \times \mathcal{Z} \rightarrow \mathbb{R}$, which distinguishes between samples extracted from the joint distribution, $P(Y, Z)$ i.e. a positive image-caption pair and the product of marginals, $P(Y)P(Z)$ i.e. a negative image-caption pair. This discriminator function can be modelled as an arbitrary neural network with parameters ω that can be jointly optimized with the encoders during training (Belghazi et al., 2018). In this work, we use a projection and alignment based architecture similar to the one presented in Deep InfoMax (Hjelm et al., 2018).

Given a pair of input one-dimensional representations, both vectors are first projected using a projection module with two linear layers separated by a ReLU and a linear shortcut. A dot-product of these projections is then computed to get alignment scores. The projection function maps these representations to an aligned cross-modal latent space. Separate projection functions are used for image and text representations. Positive and negative pairs of image-text representations are passed through the discriminator to get respective scores which are then used to estimate and maximize mutual information using our objective. This architecture, in addition to being simple and computationally inexpensive, also offers alignment of the representations into a common cross-modal latent space which uses cosine similarity as the distance metric.

A.4 Ablations

Batch-size Ablations: A salient feature of our pre-training framework is that we use a lower-bound on the mutual information that can be optimized with only one negative sample. This allows us to use much smaller batch-sizes compared to the original CLIP (Radford et al., 2021) model. In this section, we evaluate the PASCAL VOC classification performance of the visual backbones trained with a batch sizes 64, 128, 256, 512 and 1024. These ablations are performed with a 2-layered BERT model as the text-encoder and a ResNet-50 as the image encoder for 200K iterations.

Table 8: **Batch size Ablations:** We show the performance of a ResNet-50 trained with CLIP-Lite using varying batch-sizes. We observe that the performance drops marginally with the batch size 512. Additionally, we can see that the model is able to converge fairly well with the significantly lower batch size of 64.

Batch Size	VOC07
64	74.7
128	81.3
256	84.9
512	87.5
1024	87.9

Visual Encoder Ablations: In this section, we compare the performance of our pretraining method using a ResNet-18, ResNet-50, and ResNet-101 backbones using the downstream PASCAL VOC classification task. These ablations are performed with a 2-layered BERT model as the text-encoder with a batch-size of 512 for 200K iterations.

Table 9: **Visual Encoder Ablations:** We show the performance of CLIP-Lite using 3 visual backbones of varying sizes.

Visual Backbone	VOC07
ResNet-18	83.8
ResNet-50	87.5
ResNet-101	87.8

Text Encoder Ablations: In this section, we compare the downstream PASCAL VOC (Everingham et al., 2010) classification performance of a ResNet-50 visual backbone pretrained using a text encoder transformer with varying capacities. We train 4 transformer variants, (1) pre-trained BERT$_{\text{base}}$ (Devlin et al., 2018), (2) 2-layered, (3) 4-layered, (4) 6-layered, and a (5) 12-layered BERT-like transformer. These ablations are performed with a ResNet-50 as the image encoder with a batch-size of 512 for 200K iterations.

Table 10: **Text Encoder Ablations:** We show the performance of a ResNet-50 trained with CLIP-Lite using different text encoders. We observe that the performance drops marginally when training from scratch. Additionally, we also see that using a transformer with 2-layers works almost as well as a 12-layered transformer when trained from scratch.

Text Encoder	VOC07
BERT$_{\text{base}}$ init.	88.1
2-layers	87.5
4-layers	87.6
6-layers	87.6
12-layers	87.9

Zero-shot classification templates While performing zero-shot classification, we use the class names of target images to generate captions that the images should align with. The performance is compared when captions are generated using three different templates. We test three different class prompt templates and compare our performance against an equivalently trained CLIP model on the COCO dataset. As seen in Table 11, both CLIP and CLIP-Lite prefer more descriptive prompts.

Table 11: **Zero-Shot Templates on CIFAR-10:** We evaluate different prompts and find the CLIP-Lite prefers more descriptive prompts.

Class Prompt	CLIP-COCO	CLIP-Lite
"a {class name}"	13.3	30.8
"a picture of a {class name}"	14.5	32.6
"a photo of a {class name}"	16.3	33.0

A.5 Training CLIP on COCO-Captions Dataset

We use a CLIP model trained on the COCO dataset as a baseline for several demonstrated tasks. For this purpose, we use an open-source implementation[3] of CLIP. We train a standard ResNet-50 (He et al., 2016) based CLIP model that takes in a 224 × 224 image and generates 2048-dimensional features at the pre-logit layer. For textual encoding, we use a transformer (Vaswani et al., 2017) model and use the output `[CLS]` token as the text representation. We use the COCO Captions dataset (Chen et al., 2015) which has 118K images with five captions per image. During training time we apply (1) random cropping, (2) color jittering, (3) random horizontal flips while interchanging the words 'left' and 'right' in the caption, and (4) normalization using the ImageNet image mean. We train using the Adam Optimizer (Kingma and Ba, 2014) with decoupled weight decay regularization (Loshchilov and Hutter, 2016) for all weights except gains or biases. We perform

[3] https://github.com/mlfoundations/open_clip

distributed training across 8 GPUs with batch normalization (Ioffe and Szegedy, 2015) per GPU with an overall batch-size of 1024. We warm-up to the initial learning rate in 10K steps and decay to zero with the cosine schedule. We found that using the learning rate of 10^4 works slightly better ($+1.4\%$ on VOC07) than the originally recommended 5×10^5. We also found that the performance incrementally improves ($+1.9\%$ on VOC07) with longer training therefore we train for 250K iterations, similar to ours. All other training details and hyper-parameters were kept the same as the original work (Radford et al., 2021). Please note that the ResNet-50 backed CLIP model trained by us on the COCO dataset outperforms (+1.2% Zero-shot Acc. on CIFAR10) publicly available weights[4].

[4] https://github.com/revantteotia/clip-training/blob/main/zero_shot_eval_output/coco_trained_clip_observations.md