Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training

Filip Radenovic¹, Abhimanyu Dubey¹*, Abhishek Kadian¹*, Todor Mihaylov¹*, Simon Vandenhende¹*, Yash Patel², Yi Wen¹, Vignesh Ramanathan¹ and Dhruv Mahajan¹†
¹Meta AI ²CTU in Prague

Abstract

Vision-language models trained with contrastive learning on large-scale noisy data are becoming increasingly popular for zero-shot recognition problems. In this paper we improve the following three aspects of the contrastive pre-training pipeline: dataset noise, model initialization and the training objective. First, we propose a straightforward filtering strategy titled Complexity, Action, and Text-spotting (CAT) that significantly reduces dataset size, while achieving improved performance across zero-shot vision-language tasks. Next, we propose an approach titled Concept Distillation to leverage strong unimodal representations for contrastive training that does not increase training complexity while outperforming prior work. Finally, we modify the traditional contrastive alignment objective, and propose an importance-sampling approach to up-sample the importance of hard-negatives without adding additional complexity. On an extensive zero-shot benchmark of 29 tasks, our Distilled and Hard-Negative Training (DiHT) approach improves on 20 tasks compared to the baseline. Furthermore, for few-shot linear probing, we propose a novel approach that bridges the gap between zero-shot and few-shot performance, substantially improving over prior work. Models are available at github.com/facebookresearch/diht.

1. Introduction

An increasingly popular paradigm in multimodal learning is contrastive pre-training [11, 28, 41, 43, 62, 76, 85, 87], which involves training multimodal models on very large-scale noisy datasets of image-text pairs sourced from the web. It has been shown to be incredibly effective for a variety of vision-language tasks without any task-specific fine-tuning (i.e., zero-shot), such as image classification [65], text and image retrieval [45, 59], visual question answering [21], among several others. In this paper, we study the problem of contrastive pre-training for dual-encoder architectures [62] with the objective of improving image-text alignment for zero-shot tasks. We revisit three important aspects of the contrastive pre-training pipeline – noise in datasets, model initialization, and contrastive training, and present strategies that significantly improve model performance on a variety of zero-shot benchmarks, see Figure 1.

Most image-text datasets are noisy and poorly-aligned. Few recent efforts [27] have tried to clean the noise by filtering samples based on alignment scores from an existing model like CLIP [62]. However, this approach is limited by the biases and flaws of the model itself. On the other hand, momentum-based approaches [41] to reduce noise are infeasible for large-scale training due to their increased compute and memory requirements. To this end, we provide a scalable and effective approach titled Complexity, Action and Text-spotting (CAT) filtering. CAT is a filtering strategy to select only informative text-image pairs from noisy web-scale datasets. We show that training on a CAT-filtered version of large-scale noisy datasets such as LAION [66] can provide up to 12% relative improvements across vision-language tasks despite removing almost 80% of the training data, see Section 4.2 and Table 1 for more details.

A common strategy [58, 89] to further improve multimodal training is to warm-start it with image and text models pre-trained at large scale on their respective modalities. However, due to the increased noise in image-text data, fine-tuning the entire model undermines the benefits of the warm-start. One can alternatively use model freezing strategies like locked-image tuning [89], but they are unable to adapt to the complex queries present in multimodal problems (e.g., cross-modal retrieval) and the models perform poorly on retrieval benchmarks (see Section 4.2). We
propose an entirely different approach, concept distillation (CD), to leverage strong pre-trained vision models. The key idea behind concept distillation is to train a linear classifier on the image encoder to predict the distilled concepts from a pre-trained teacher model, inspired by results in weakly-supervised large-scale classification [49, 71].

Finally, we revisit the training objective: almost all prior work has utilized noise-contrastive estimation via the InfoNCE loss [55], shortcomings have been identified in the standard InfoNCE formulation [12, 30]. We demonstrate that by using a model-based importance sampling technique to emphasize harder negatives, one can obtain substantial improvements in performance.

A summary of our pipeline is available in Figure 2. Our combined approach obtains significant improvements over the baseline for dual-encoder architectures on an elaborate benchmark of 29 tasks. Specifically, with the ViT-B/16 [17] architecture, we improve zero-shot performance on 20 out of 29 tasks, over CLIP training on the LAION-2B dataset [27, 66], despite training on a subset that is 80% smaller, see Figure 4. Furthermore, we demonstrate that even when trained with the smaller (but relatively less noisy) pretraining dataset PMD, our performance is better on 28 out of 29 tasks than CLIP trained on the same data, often with a large margin, see Figure 5.

Additionally, we present a simple yet effective approach to maintain the performance continuum as one moves from zero-shot to k-shot learning in the low data regime. Prior work [62] has shown a substantial drop in performance as one moves from zero-shot to k-shot learning, which is undesirable for practical scenarios. We propose an alternate linear probing approach that initializes the linear classifier with zero-shot text prompts and ensures that final weights do not drift away too much via projected gradient descent [5]. On ImageNet1K, we show huge improvements over prior work for small k values. For example, our approach improves 5-shot top-1 accuracy by an absolute margin of 7% (see Figure 6) compared to the baseline strategy of linear probing with a random initialization.

2. Related work

Dataset curation for contrastive pretraining. Large-scale contrastive pretraining [11, 28, 41, 43, 62, 76, 85, 87] typically requires dataset sizes of the order of hundreds of millions to billions. Seminal works in this area, e.g., CLIP [62] and ALIGN [28], have largely relied on image-text pairs crawled from the web. Subsequently, versions of large-scale image-text datasets have been created but not released publicly, including WIT-400M [62], ALIGN-1.8B [28], FILIP-340M [85], FLD-900M [87], BASIC-6.6B [58], PaLI-10B [10]. These datasets often use unclear or primitive cleaning strategies, e.g., removing samples with short or non-English captions. Recently, LAION-400M [67] used CLIP-based scores to filter down a large dataset. The authors later released an English-only LAION-2B and a LAION-5B unfiltered dataset sourced from Common Crawl\(^1\). Apart from LAION-400M and BLIP [40] which uses the bootstrapped image-grounded text encoder to filter out noisy captions, there has not been a significant investment in systematic curation strategies to improve zero-shot alignment performance on large-scale pretraining. In contrast to the previous work, we use quality-motivated filters that retain images whose captions are sufficiently complex, contain semantic concepts (actions), and do not contain text that can be spotted in the image [38].

Distillation from pre-trained visual models. Knowledge distillation [25] aims to transfer knowledge from one model to another and has been used in many contexts ranging from improving performance and efficiency [6, 7, 42, 64, 68, 74, 81] to improving generalization capabilities [16, 43, 44]. Several approaches use self-distillation to improve performance with lower computational overhead [23, 82, 88]. For vision and language pre-training, [2, 31, 41] use soft-labels computed using embeddings from a moving average momentum model with the goal to reduce the adverse effects of noisy image-text pairs in the training data. Our concept distillation approach is a cheaper and more effective alternative, since it does not require us to run the expensive teacher model throughout the training\(^2\) while retaining the most useful information from the visual concepts.

Another approach to take advantage of pre-trained visual models is to use them to initialize the image encoder, and continue pre-training either by locking the image encoder [58, 89] or fine-tuning [58]. However, these approaches lack the ability to align complex text to a fully-trained image encoder, and thus perform poorly on multimodal tasks, e.g. cross-modal retrieval (see Section 4.3).

\(^1\)commoncrawal.org
\(^2\)Distillation targets can be pre-computed and stored.
Contrastive training with hard negatives. Noise-contrastive estimation (NCE) [22] is the typical objective for vision-text learning, with applications across large-scale multimodal alignment [11, 28, 43, 62] and unsupervised visual representation learning [24, 53]. Several lines of work have studied the shortcomings of the original InfoNCE objective [55], specifically, the selection and importance of negative samples. Chuang et al. [12] present a debiasing approach to account for false negatives at very large batch sizes, typical in large-scale pretraining. Kalantidis et al. [30] present a MixUp approach to improve the quality of hard negative samples for unsupervised alignment. Using model-specific hard negatives in the training objective is proven to reduce the estimation bias of the model as well [30]. Contrary to prior semi-supervised work, we extend the model-based hard negative objective, first proposed in Robinson et al. [63] to multimodal alignment.

3. Method

Background. We consider the task of contrastive image-text pretraining. Given a dataset \( D = \{(I_i, T_i)\}_{i=1}^N \) of image-text pairs, we want to learn a dual encoder model \( \Phi = \{\phi_{\text{image}}, \phi_{\text{text}}\} \), where \( \phi_{\text{image}} \) represents the image encoder, and \( \phi_{\text{text}} \) denotes the text encoder. We use the shorthand \( x = \phi_{\text{image}}(I) \) and \( t = \phi_{\text{text}}(T) \) to denote the encoded images and texts, respectively, for an image-text pair \( (I, T) \). We will now describe the three crucial components of our approach followed by the final training objective.

3.1. Complexity, Action, and Text (CAT) filtering

Our complexity, action, and text spotting (CAT) filtering is a combination of two filters: a caption complexity filter that removes image-caption pairs if a caption is not sufficiently complex, and an image filter that removes pairs if the image contains text matching the caption to prevent polysemy during alignment. We use the LAION-2B pre-cleaned obtained after using filters\(^3\) in [69] as the base dataset.

Filtering captions via complexity & actions. Noisy web-scale datasets do not have any semantic-based curation, and hence captions can be irrelevant, ungrammatical and unaligned. Our motivation is to decrease such noise by simply selecting captions that possess sufficient complexity, so that the training distribution matches the target tasks. To this end, we build a fast rule-based parser that extracts objects, attributes and action relations (see Figure 3 for an example) from text and we use the resulting semantic graph to estimate the complexity of the image captions. Specifically, we define the complexity of a caption as the maximum number of relations to any object present in the parse graph. For example, the caption “A black cat is chasing a small brown bird,” the object “bird” has the attributes “small”, “brown” and “A black cat is chasing”, and hence, the complexity of the caption is C3. We only retain samples that at least have a C1 caption complexity. To further remove pairs likely containing products, we filter out captions if they do not contain at least one action (as obtained from the parser).

Filtering images via text-spotting. Image-caption pairs in web-scale datasets often display part of the caption as text in the image (on visual inspection, we found up to \( \sim 30\% \) such examples for LAION [66]). Minimizing the objective, in these cases, can correspond to spotting text (e.g., optical character recognition) rather than the high-level visual semantics (e.g., objects, attributes) we would like the model to align to. This will reduce performance on object-centric and scene-centric downstream zero-shot tasks, and hence we remove such images from the training set using an off-the-shelf text spotter [38]. We remove image-text pairs with a text spotting confidence of at least 0.8 and at least 5 predicted characters matching the caption in a sliding window. We observe (by inspection) that this approach is efficient at identifying images with text, and failure cases are primarily in non-English text. Filtering with multilingual text spotters trained can fix this issue, however, we leave this as future work. Filtering statistics can be found in the supplement.

3.2. Concept distillation

Recognizing visual concepts in images that correspond to objects and attributes in corresponding captions is crucial for alignment. We therefore propose to distill these concepts from a pre-trained teacher model to our image encoder. Specifically, we add two auxiliary linear classifiers on top of the encoded image embeddings \( x \) to predict (i) objects and (ii) visual attributes and use the teacher model to generate the pseudo-labels for training them. These classifiers are trained jointly with the contrastive loss.

We parse image captions using a semantic parser that extracts objects and attributes from text (Section 3.1) and use these as pseudo-labels. We then train the linear classifiers on the teacher model embeddings with a soft-target cross-entropy loss [20], after square-root upsampling low-frequency concepts [49]. It is important to freeze the backbone of the teacher model to make sure we retain the advantages of using a stronger model for distillation. For

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{An example caption and its parse. The caption has C3 complexity (due to \textit{bird}) and has 1 action (\textit{chasing}).}
\end{figure}

\(^3\)Not-suitable-for-view images and toxic captions.
each image, we then use these trained linear classifiers to generate two softmax probability vectors $p^{\text{obj}}$ for objects, and $p^{\text{attr}}$ for attributes, respectively. To minimize the storage overhead, we further sparsify them by retaining only the top-$k$ predicted class values and re-normalizing them to generate the final pseudo-labels. During multimodal training, we use the cross-entropy loss with these pseudo-label vectors as targets. Unless specified otherwise, we use the ViT-H/14 [17] architecture pretrained from SWAG [71] as the teacher model. See Section 4.2 and the supplementary material for ablations on the effect of different backbones and retaining top-$k$ predictions, and further details.

There are several advantages of our concept distillation approach. First, the teacher predictions capture correlations from the strong vision encoding, making them more informative as labels compared to the captions themselves. The captions are limited to a few objects and attributes, while the teacher predictions yield a more exhaustive list. Moreover, our approach reaps the benefits of the recently proposed and publicly-available strong unimodal vision models more effectively than other distillation approaches, as training linear classifiers on a frozen teacher model is inexpensive. After predictions are stored, we discard the teacher model and thus bypass the memory and compute limitations of simultaneously running the student and teacher model in standard distillation approaches [25, 74], which is critical for large teacher models. We demonstrate empirically (see Section 4.2) that our strategy works better than distilling teacher embeddings directly. Additionally, compared to approaches that warm-start the image encoder with pre-trained models, our method can leverage higher capacity teacher models without difficulty and unlike locked-image trained models, our method can leverage higher capacity to approaches that warm-start the image encoder with pre-trained teacher embeddings directly. Additionally, compared to approaches that warm-start the image encoder with pre-trained models, our method can leverage higher capacity teacher models without difficulty and unlike locked-image tuning [58, 89], our approach gives the flexibility of training the image encoder for better alignment, while retaining the strength of the pre-trained visual features.

3.3. Multimodal alignment with hard negatives

Contrastive learning [55] has quickly become the de-facto approach for multimodal alignment, where most prior work focuses on the multimodal InfoNCE [55] objective, given for any batch $X = \{(x_i, t_i)\}_{i=1}^n$ of featurized image-text pairs as (for some learnable temperature $\tau > 0$),

$$
\mathcal{L}_{\text{NCE}}(X) = - \sum_{i=1}^{n} \log \frac{e^{x_i^T t_i / \tau}}{\sum_{j} e^{x_j^T t_i / \tau}} + \log \frac{e^{x_i^T t_i / \tau}}{\sum_{j} e^{x_j^T t_i / \tau}}.
$$

While this approach has enjoyed immense success in multimodal alignment [28, 62], when learning from large-scale noisy datasets, uniform sampling as applied in noise-contrastive estimation can often provide negative samples that are not necessarily discriminative, necessitating very large batch sizes. For the problem of contrastive self-supervised learning, Robinson et al. [63] propose an importance-sampling approach to reweight negative samples within a batch so that “harder” negatives are up-sampled in proportion to their difficulty. We present a similar strategy for multimodal alignment. Specifically, for some $0 < \alpha \leq 1$, $\beta \geq 0$, we propose the following hard-negative noise contrastive multimodal alignment objective:

$$
\mathcal{L}_{\text{HN-NCE}}(X) = - \sum_{i=1}^{n} \log \left[ \frac{e^{x_i^T t_i / \tau}}{\alpha \cdot e^{x_i^T t_i / \tau} + \sum_{j \neq i} e^{x_j^T t_i / \tau} w_{i,j}^{\text{attr}} t_i t_j \right]
\quad - \sum_{i=1}^{n} \log \left[ \frac{e^{x_i^T t_i / \tau}}{\alpha \cdot e^{x_i^T t_i / \tau} + \sum_{j \neq i} e^{x_j^T t_i / \tau} w_{i,j}^{\text{obj}} t_i t_j \right].
$$

Where the weighing functions are given as:

$$
w_{i,j}^{\text{attr}} = \frac{(n-1) \cdot e^{\beta x_i^T t_j / \tau}}{\sum_{k \neq i} e^{\beta x_k^T t_j / \tau}}, \quad w_{i,j}^{\text{obj}} = \frac{(n-1) \cdot e^{\beta x_i^T t_j / \tau}}{\sum_{k \neq i} e^{\beta x_k^T t_j / \tau}}.
$$

The weights $w_\beta$ are designed such that difficult negative pairs (with higher similarity) are emphasized, and easier pairs are ignored. Furthermore, $\alpha$ rescales the normalization with the positive terms to account for the case when false negatives are present within the data. The form of weights $w_\beta$ is an unnormalized von Mises-Fisher distribution [50] with concentration parameter $\beta$. Observe that we obtain the original objective when setting $\alpha = 1$ and $\beta = 0$. There are several key differences with the original formulation of [63] and the HN-NCE objective presented above. First, we utilize only cross-modal alignment terms, instead of the unimodal objective presented in [63]. Next, we employ separate penalties for text-to-image and image-to-text alignment. Finally, we incorporate a learnable temperature parameter $\tau$ to assist in the learning process. We discuss our design choices in more detail with additional theoretical and experimental justifications in the supplementary material.

3.4. Training objective

For any batch $X = \{(x_i, t_i)\}_{i=1}^n$ of $n$ image-text pairs, we minimize the following objective:

$$
\mathcal{L}_{\text{HN-NCE}}(X) + \mathcal{L}_{\text{CE-O}}(X) + \mathcal{L}_{\text{CE-A}}(X),
$$

where,

$$
\mathcal{L}_{\text{CE-O}}(X) = \sum_{i=1}^{n} \text{Cross-Entropy}(p_{i}^{\text{obj}}, f_{\text{obj}}(x_i)), \quad \text{and},
$$

$$
\mathcal{L}_{\text{CE-A}}(X) = \sum_{i=1}^{n} \text{Cross-Entropy}(p_{i}^{\text{attr}}, f_{\text{attr}}(x_i)).
$$

Here, both $f_{\text{obj}}$ and $f_{\text{attr}}$ are linear classifiers, the vectors $p_{i}^{\text{obj}}$, $p_{i}^{\text{attr}}$ are the top-$k$ predicted objects and attributes from the teacher model (Section 3.2), and $\mathcal{L}_{\text{HN-NCE}}$ is the hard-negative contrastive alignment loss (Section 3.3).

\footnote{We normalize by $n-1$ as this is the number of negatives.}
4. Experiments

Here we evaluate our approach across a broad range of vision and vision-language tasks. We provide extensive ablations on 29 tasks over the design choices in Section 4.2, and compare with state-of-the-art approaches on popular zero-shot benchmarks in Section 4.3. Finally, we present an alternate approach to do few-shot classification with prompt-based initialization in Section 4.4.

4.1. Experimental setup

Training datasets. We use a 2.1B English caption subset of the LAION-5B dataset [66]. Prior to training, we filter out sample pairs with NSFW images, toxic words in the text, or images with a watermark probability larger than 0.5, following [69]. This leaves us with 1.98B images, which we refer to throughout the paper as the LAION-2B dataset. Additionally, we explore training our models on a collection of Public Multimodal Datasets (PMD) from [70]. PMD contains training splits of various public datasets. After downloading\(^3\) the data we are left with 63M (vs. 70M reported in [70]) image-text pairs due to missing samples and SBU Captions [56] (originally in PMD) going offline.

Training details. For our model architecture, we closely follow CLIP by Radford et al. [62]. We utilize Vision Transformers (ViT) [17] for images and Text Transformers [75] for captions. We experiment with 3 different architectures, denoted as B/32, B/16, and L/14, where 32, 16, and 14 denote the input image patch size. See the supplementary for architecture details. For distillation and fine-tuning experiments, we utilize the public SWAG-ViT models [71], pre-trained with weak supervision from hashtags.

We use the Adam [33] optimizer with a decoupled weight decay [48] and a cosine learning rate schedule [47]. The input image size is 224×224 pixels. To accelerate training and save memory, we use mixed-precision training [51]. All hyperparameters are presented in the supplementary. They are selected by training B/32 on a small scale setup, and reused for all architectures. For objects and attributes classifiers, we found that scaling the learning rate by 10.0 and weight decay by 0.01 gave better results. We train our models on 4B, 8B, 16B, and 32B total samples. For ViT-L/14, we further train the model at a higher 336px resolution for 400M samples, denoting this model as L/14@336. We trained L/14 for 6 days on 512 A100 GPUs with 16B processed samples for a total of 7.4 × 10^4 GPU hours.

Evaluation benchmarks. We evaluate our models on a zero-shot benchmark of 29 tasks: (i) 17 image classification, (ii) 10 cross-modal retrieval, (iii) 2 visual question answering. Dataset details are presented in the supplement.

\(^3\)Downloaded following huggingface.co/datasets/facebook/pmd.

4.2. Ablations on zero-shot benchmarks

In this section, we ablate our three pretraining contributions: dataset filtering, distillation from objects and attributes predictions, and, hard negative contrastive objective. Ablations are performed over zero-shot Accuracy@1 on the ImageNet1K [65] (IN) validation set, text-to-image (T2I) and image-to-text (I2T) zero-shot Recall@1 on the COCO [60] and Flickr [59] test sets. We also report the change in accuracy (%) over 29 zero-shot tasks between our model and baselines. For a fair comparison, we train all approaches presented in this section (including baselines).

Effect of dataset filtering. We apply our filters, as well as filtering based on CLIP [62] alignment score (<0.35), and ablate the baseline performance, without distillation or hard negative contrastive training, in Table 1 for ViT-B/32 model architecture. All models see 4B total samples during training, while the number of unique samples drops after each filtering step. Complexity filter (C) in row (3) reduces the dataset size by around 270M, while slightly increasing image-text alignment as observed on T2I task. Next, action filter (A) in row (4) reduces the size by more than 1B, while it has a large benefit in aligning complex text. However, as expected, it hurts performance on object-centric ImageNet. Finally, text-spotting (T) filter in row (5) boosts alignment across the board, due to the fact that it removes the need to learn a bimodal visual representation of the text. We also compare with filtering based on CLIP score in row (2), which was selected such that the dataset size is comparable to ours, and show that it is too strict and removes plenty of useful training pairs, thus hurting the performance. Finally, LAION-CAT, with only 22% of the original dataset size, significantly boosts image-text zero-shot performance. We also observed that gains hold as we train for longer training schedules. See the supplementary for details.

Effect of distillation approach. To understand the effect of direct distillation from a pre-trained SWAG-ViT visual encoder [71], we investigate two baseline approaches: (1) Embedding distillation (ED) borrows from SimSiam [9] and uses an auxiliary negative cosine similarity loss between the image representation from the student visual encoder and the pre-trained SWAG model.
(2) **Distribution distillation (DD)** borrows ideas from momentum distillation in ALBEF [41] and computes the cross-modal similarities between the SWAG image representation and the student text representation and uses them as soft-labels for student image representation and text alignment. The soft-labels are linearly combined with the hard 0 − 1 labels before applying the InfoNCE [55] loss.

A comparison of our distillation from predicted concepts (CD) with the aforementioned distillation approaches is presented in Table 2 (upper section). Note that for a fair comparison, we do not use our hard-negative contrastive loss for these experiments. Our distillation approach performs the best, even though it has virtually no training overhead as the predicted concepts are pre-computed, while, e.g., ED is 60% slower with an 8% increase in GPU memory due to the need of running an additional copy of the vision tower. One could pre-compute embeddings for ED and DD as well, but that increases dataset size by 1.2TB and creates a data loading bottleneck, while our pre-computed predictions take only 32.6GB additional storage space when saving the top-10 predictions (see supplementary). We additionally show that our approach is robust to the number of top-k predictions used, details in the supplementary.

One could also use an external unimodal image model and fine-tune it on the image-text alignment task instead of using distillation. We follow [89] and explore three fine-tuning options as baselines: (i) locked-image tuning (LiT) where the image encoder is locked, and only the text encoder is trained, (ii) fine-tuning (FT) where the image encoder is trained with a learning rate scaled by 0.01 compared to the text encoder, (iii) fine-tuning with delay (FT-delay) where the image encoder is trained with a learning rate scaled by 0.01 compared to the text encoder, (iii) fine-tuning with delay (FT-delay) where the image encoder is locked for half of the pre-training epochs following (i), and then fine-tuned for the rest following (ii). Results of these setups are ablated in Table 2 (lower section). LiT vs. FT is a trade-off between strong performance on image recognition tasks (as measured with ImageNet1K) and better image-text alignment (as measured by COCO and Flickr). Locking the image encoder makes the alignment very hard to achieve, but fine-tuning it hurts its original image recognition power. On the other hand, we show that our concept distillation is the best of both worlds, it surpasses LiT or FT in 4 out of 5 metrics. Another drawback of FT is that it requires the same architecture in the final setup, while CD can be effortlessly combined with any architecture or training setup, by using stored predictions as metadata. To conclude, unlike related approaches, our proposed distillation: (i) has almost no cost at training, (ii) is architecture agnostic, (iii) improves both image recognition and complex image-text alignment.

**Effect of hard negative contrastive training.** We present the ablation when using hard negative contrastive objective (HN-NCE) in Table 3. Performance suggests that using the newly proposed loss is beneficial compared to the vanilla InfoNCE, and that its positive effects are complementary to the gains from the proposed distillation from objects and attributes predictions. Please see the supplementary for ablations on the effect of the hyperparameters $\alpha$ and $\beta$.

**Effect when pre-training on PMD.** Finally, we analyze our proposed recipes when training visual-language models on a much smaller dataset, i.e. PMD with 63M training samples. Results are shown in Table 4. All contributions improve the performance over baseline significantly, hence we conclude that using the proposed pipeline is very beneficial in low-resource training regimes. Note that, the PMD dataset contains COCO and Flickr training samples, hence, it is not strictly zero-shot evaluation. For that reason, we do not compare our models trained on PMD dataset with state-of-the-art models in the following section. However, we believe these strong findings will motivate usage of our approach on smaller and cleaner datasets, as well.

**Zero-shot benchmarks.** We denote model trained with our proposed concept distillation and hard-negative loss as DiHT. To showcase our model’s performance in more detail, we report our DiHT-B/16 trained on LAION-CAT with 438M samples vs. CLIP-B/16 baseline trained by us on LAION-2B with 2B samples in Figure 4. Additionally, we report DiHT-B/16 vs. CLIP-B/16 baseline, where both

<table>
<thead>
<tr>
<th>Method</th>
<th>SWAG (teacher)</th>
<th>IN</th>
<th>COCO</th>
<th>Flickr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>68.7</td>
<td>42.8</td>
<td>60.5</td>
<td>72.8</td>
</tr>
<tr>
<td>ED</td>
<td>69.2</td>
<td>42.6</td>
<td>59.4</td>
<td>72.8</td>
</tr>
<tr>
<td>DD</td>
<td>68.6</td>
<td>41.8</td>
<td>57.4</td>
<td>71.7</td>
</tr>
<tr>
<td>CD (ours) B/16</td>
<td>71.0</td>
<td>42.8</td>
<td>59.5</td>
<td>72.3</td>
</tr>
<tr>
<td>CD (ours) H/14</td>
<td>72.3</td>
<td>43.4</td>
<td>60.4</td>
<td>73.8</td>
</tr>
</tbody>
</table>

Table 3. Evaluating effect of using hard negative contrastive loss.

<table>
<thead>
<tr>
<th>#</th>
<th>Method</th>
<th>SWAG (teacher)</th>
<th>IN</th>
<th>COCO</th>
<th>Flickr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CD</td>
<td>68.7</td>
<td>42.8</td>
<td>60.5</td>
<td>72.8</td>
</tr>
<tr>
<td>2</td>
<td>CD</td>
<td>72.3</td>
<td>43.4</td>
<td>60.4</td>
<td>73.8</td>
</tr>
<tr>
<td>3</td>
<td>CD</td>
<td>72.0</td>
<td>43.7</td>
<td>62.0</td>
<td>73.2</td>
</tr>
</tbody>
</table>

Table 4. Evaluating effect of using different initialization or distillation approaches. Evaluation performed on ViT-B/16 model architecture trained for 16B processed samples on LAION-CAT. Init: Initialization with random or SWAG-B/16 weights. ED: Embedding distillation. DD: Distribution distillation. LiT: Locked image tuning. FT: Fine-tuning. FT-delay: Locked image tuning for 50% followed by fine-tuning for the rest. CD: Our concept distillation using teacher-predicted objects and attributes.

Table 2. Evaluating effect of using different initialization or distillation approaches. Evaluation performed on ViT-B/16 model architecture trained for 16B processed samples on LAION-CAT. Init: Initialization with random or SWAG-B/16 weights. ED: Embedding distillation. DD: Distribution distillation. LiT: Locked image tuning. FT: Fine-tuning. FT-delay: Locked image tuning for 50% followed by fine-tuning for the rest. CD: Our concept distillation using teacher-predicted objects and attributes.
We outline those details in the table, for easier comparison. Our proposed hard negative contrastive loss.

In dual-encoder models in Table 4, CLIP-B/16 wins on 28 out of 28 benchmarks, even when our training dataset, ImageNet1K [65] and image classification benchmarks, has severe downsides on multi-modal tasks such as cross-modal retrieval. Our ablation in Section 4.2 also confirms this issue. On the other hand, our approach does not suffer from such negative effects.

**4.4. Few-shot linear probing**

The ideal scenario for leveraging zero-shot recognition models is to warm start the task without training data and then improve the performance (via training a linear probe) via few-shot learning as more and more data is seen. However, in practice, few-shot models perform significantly worse than zero-shot models in the low-data regime.

We present an alternate approach to do few-shot classification with prompt-based initialization. The key idea of our approach is to initialize the classifier with the zero-shot text prompts for each class, but to also ensure that the final weights do not drift much from the prompt using projected gradient descent (PGD) [5]. While few-shot models
have been initialized with prompt priors in the past with naive $L_2$ penalties for weight to prevent catastrophic forgetting [34], these approaches do not improve performance and the model simply ignores the supervision. In contrast, for any target dataset $D_{\text{target}} = \{(x_i, y_i)\}_{i=1}^n$, where $x_i = \phi_{\text{image}}(I_i)$ denotes the image features from the trained image tower, we solve the following optimization problem, for some $\delta, \delta_b > 0$:

$$\min_{\|W\|_2 \leq \delta, \|b\|_2 \leq \delta_b} \sum_{i=1}^n \mathcal{L}_{CE}(y_i, x_i^T (W + W_0) + b).$$

Here $W_0 \in \mathbb{R}^{d \times n_c}$ denotes the prompt initialization from the text encoder. To optimize the objective, one can use projected gradient descent [5]. We observe that our approach is able to bridge the gap between zero-shot and 1-shot classification, a common issue in prior linear probe evaluations.

Table 6. Comparison with zero-shot state-of-the-art dual-encoder models. px: input image size; #P: model size; #D: training dataset size; #S: total samples processed at training. We evaluate CLIP [62] and OpenCLIP [27] using our codebase, other numbers are copied from respective papers. Grouped models (e.g., ViT-B/32) share same vision and language architecture as our model, following CLIP [62], others have different architectures and we outline the vision one. *FILIP uses token-wise similarity, which is more expensive than global-token similarity and requires adapting the architecture, hence we put it in “Other”.

<table>
<thead>
<tr>
<th>Method</th>
<th>px</th>
<th>#P</th>
<th>#D</th>
<th>#S</th>
<th>IN</th>
<th>COCO</th>
<th>Flickr</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B/32</td>
<td>224</td>
<td>151M</td>
<td>400M</td>
<td>12.8B</td>
<td>63.4</td>
<td>31.4</td>
<td>49.0</td>
</tr>
<tr>
<td>CLIP [62]</td>
<td>224</td>
<td>151M</td>
<td>400M</td>
<td>12.8B</td>
<td>62.9</td>
<td>34.8</td>
<td>52.3</td>
</tr>
<tr>
<td>OpenCLIP [27]</td>
<td>224</td>
<td>151M</td>
<td>400M</td>
<td>12.8B</td>
<td>67.5</td>
<td>40.3</td>
<td>56.3</td>
</tr>
<tr>
<td>DHT</td>
<td>224</td>
<td>151M</td>
<td>438M</td>
<td>16B</td>
<td>68.0</td>
<td>40.6</td>
<td>59.3</td>
</tr>
<tr>
<td>ViT-B/16</td>
<td>224</td>
<td>150M</td>
<td>400M</td>
<td>12.8B</td>
<td>68.4</td>
<td>33.7</td>
<td>51.3</td>
</tr>
<tr>
<td>CLIP [62]</td>
<td>224</td>
<td>150M</td>
<td>400M</td>
<td>12.8B</td>
<td>67.1</td>
<td>37.8</td>
<td>55.4</td>
</tr>
<tr>
<td>OpenCLIP [27]</td>
<td>224</td>
<td>150M</td>
<td>400M</td>
<td>12.8B</td>
<td>69.2</td>
<td>40.5</td>
<td>57.8</td>
</tr>
<tr>
<td>DHT</td>
<td>224</td>
<td>150M</td>
<td>438M</td>
<td>32B</td>
<td>72.2</td>
<td>43.3</td>
<td>60.3</td>
</tr>
<tr>
<td>ViT-L/14</td>
<td>224</td>
<td>428M</td>
<td>400M</td>
<td>12.8B</td>
<td>75.6</td>
<td>36.5</td>
<td>54.9</td>
</tr>
<tr>
<td>CLIP [62]</td>
<td>224</td>
<td>428M</td>
<td>400M</td>
<td>13.2B</td>
<td>76.6</td>
<td>37.7</td>
<td>57.1</td>
</tr>
<tr>
<td>OpenCLIP [27]</td>
<td>224</td>
<td>428M</td>
<td>400M</td>
<td>12.8B</td>
<td>78.8</td>
<td>42.1</td>
<td>60.1</td>
</tr>
<tr>
<td>DHT</td>
<td>224</td>
<td>428M</td>
<td>438M</td>
<td>32B</td>
<td>75.2</td>
<td>46.2</td>
<td>64.3</td>
</tr>
<tr>
<td>Other</td>
<td>289</td>
<td>820M</td>
<td>1.8B</td>
<td>19.7B</td>
<td>76.4</td>
<td>45.6</td>
<td>68.6</td>
</tr>
<tr>
<td>ALIGN [28]</td>
<td>224</td>
<td>340M</td>
<td>10.2B</td>
<td>77.1</td>
<td>51.0</td>
<td>68.4</td>
<td></td>
</tr>
<tr>
<td>OpenCLIP [27]</td>
<td>224</td>
<td>968M</td>
<td>2.3B</td>
<td>32B</td>
<td>77.9</td>
<td>49.3</td>
<td>65.3</td>
</tr>
</tbody>
</table>
| FiLIP uses token-wise similarity, which is more expensive than global-token similarity and requires adapting the architecture, hence we put it in “Other”.

**5. Conclusion and future work**

In this paper, we demonstrate that with careful dataset filtering and simple but effective modeling changes, it is possible to achieve substantial improvements in zero-shot performance on retrieval and classification tasks through large-scale pre-training. Our CAT filtering approach can be applied generically to any large-scale dataset for improved performance with smaller training schedules. Moreover, our concept distillation approach presents a compute and storage efficient way of leveraging very large capacity pre-trained image models for multimodal training. Finally, our simple projected gradient approach covers the crucial performance gap between zero-shot and few-shot learning.

In the future, we would like to extend our approach to multi-modal encoder/decoder architectures that although expensive, have better zero-shot performance compared to dual encoders. We also observe that benefits of our hard-negatives loss are less on noisier LAION dataset compared to PMD. It would be interesting to explore how to make it more effective in these very noisy settings. We hope that our improvements and extensive large-scale ablations will further advance the vision-language research.
References


[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML, 2021. 1, 2, 3, 4, 7, 8

[31] Zaid Khan, BG Vijay Kumar, Xiang Yu, Samuel Schulter, Mannoham Chandraker, and Yun Fu. Single-stream multilevel alignment for vision-language pretraining. In ECCV, 2022. 2
[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In ICML, 2022. 2
[41] Junnan Li, Ramprasaath Selvaraju, Akhilesht Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In NeurIPS, 2021. 1, 2, 6, 8
[44] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In ICCV, 2017. 2
[54] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In ICVGIP, 2008. 7, 13


