Evaluating Distilled Hard Negative Training on Google's Conceptual Captions Dataset

Andrew Fang
Department of Computer Science
Virginia Tech (VT) University
fafaaf61@vt.edu

Kanad Naleshwarkar
Department of Computer Science
Virginia Tech (VT) University
kanadn@vt.edu

Sanket Bhujbal
Department of Computer Science
Virginia Tech (VT) University
sanketb@vt.edu

Abstract

Contrastive pre-training has repeatedly shown to be an effective way to solve zero-shot learning problems for computer vision. However, as it frequently involves large scale datasets, the benefits of zero-shot learning can be severely reduced. DiHT is a model created to resolve many of these issues through the introduction of CAT filtering, concept distillation and multimodal alignment with hard negatives. However, as a recent development, it's efficacy on a totally unique dataset is not yet determined. We aim to determine the efficacy of this model by testing it on Google's Conceptual Captions dataset, thereby porving weather or not this will be a new and effective development for the future of zero-shot learning.

1 Project topic

Zero-shot learning is an emerging area of research in machine learning that aims to enable machines to learn from new concepts and tasks without the need for explicit training data. One of the most challenging and promising applications of zero-shot learning is cross-modal retrieval, such as text-image and image-text. In this project, we plan to evaluate the performance of a newly proposed model called Distilled and Hard-negative Training (DiHT)[11] on Google's Conceptual Captions dataset[9] and hope to discover any possible limitations as well as show future developments that can be made to the model for more accurate results.

Traditional machine learning models require large amounts of labeled data to learn, which can be a bottleneck in many real-world applications where data is scarce or expensive to obtain. Zero-shot learning aims to overcome this limitation by leveraging prior knowledge about the relationships between concepts and tasks and using it to generalize to new tasks and concepts. Thus, we wanted to explore recent advances in zero-shot learning that are outperforming benchmarks like CLIP[1].

The Distilled and Hard-negative Training (DiHT) approach has been recently proposed by Radenovic et al.[11] at Meta AI that puts forth three novel methods for improving contrastive pre-training: i) Complexity, Action, and Text-spotting (CAT) that is a filtering strategy to select only informative text-image pairs from noisy datasets ii) Concept Distillation that leverages pre-trained vision models and iii) Multimodal alignment with hard negatives. This approach is based on the existing pre-training paradigm for multimodal models, modifying the pre-training process by adding a filter to reduce noise, improving the "warm-start" of large-scale pre-training by training a linear classifier over an image encoder to predict distilled concepts, and changing from InfoNCS loss to model-based importance

Preprint. Under review.

sampling techniques. This model has been evaluated on datasets like ImageNet1K, COCO and Flickr for T2I and I2T retrieval tasks and found to be outperforming benchmark models like CLIP and Open-CLIP.

Hence, we wanted to perform an evaluation on yet another dataset, Google's Conceptual Captions[10], which provides 3.3M images annotated with captions and represents a wider variety of styles. We are planning to divide this project into two phases, the first goal will be to apply the DiHT model to the GCC dataset and evaluate the performance and the second, we will explore possible approaches that can increase the performance of this implementation.

We are hoping to explore and understand the GCC dataset and complete the first phase by March end. This will give us enough time to perform a brief exploratory study to search for approaches that can help us improve the performance of our implementation. We are aiming to complete this model improvement phase by mid-April so that we have some time in hand as a buffer and also for documenting all of the things that we learned along the way which we believe will be very useful.

2 Related work

Within computer vision, zero-shot learning problems wherein we must classify images into categories that we had previously not trained with have become increasingly common. One notable approach to zero-shot learning is a multimodal approach training both image data and text and semantic data to make predictions regarding image labels not observed in data. For instance, Fomme et al.[4] utilizes both images with labels as well as semantic data in order to improve zero-shot learning performance for data with a large number of object classifiers.

Chen et al.[2] aims to expand upon existing models through the methodology of contrastive pre-training. Contrastive pre-training utilizes large, noise datasets of image and caption pairs to train. This model has shown to be effective for computer vision tasks without fine-tuning. Jia et al.[6] had done further work on this model, adding a simple dual encoder architecture which aims to learn visual and language representation from image-text pairs using contrastive loss. The problem is that both of these approaches is that the use of large noisy datasets offsets the benefits of zero-shot learning which is the reduction of training size.

Radenovic et al.[11] aims to resolve this issues through their DiHT model, expanding on the model made by Jia. Our work will primarily be focused on this. The model, which utilizes filtering, concept distillation and a new contrastive loss is based on the work of several others. The CAT filtering used in the model is an expansion of several other, more basic filtering strateges used in models such as LAION-400M and BLIP[8]. Knowledge distillation in Neural Networks is already a common practice as shown by the work of Hinton et al.[5] Finally, the use of a loss function that focuses on hard negatives is directly lifted from the work of Robinson et al.[12] In which they show that utilizing hard negative samples can lead to better accuracy on models. Combining all of these methods is the main goal of Radenovic of which we hope to then expand by testing the model in a novel way.

3 Methodology

The model we will utilize will have already been trained by by Radenovic et al.[11] Our main goal will primarily to test this model using Google's conceptual captions dataset. We will however detail the training methodology as we feel that it is an important to understanding the unique efficacy of this model that we will later prove. As such the following will detail the work done by Radenovic et al. For training they will be using a dataset of $D = \{(I_i, T_i)\}_{i=1}^{N}$ which represent image text pairs. Specifically, the model is trained using the LAION-2B dataset after it has already been pre-cleaned using filters[13]. Our goal is to obtain image text encoders $\phi = \{\phi_{image}, \phi_{text}\}$ where $\phi_{image}(I)$ gives us an encoding for images that allows us to find the probability that certain captions apply to that image and $\phi_{text}(T)$ gives us an encoding for text that allows us to find the probability that certain images are related to the given text. Using this, we have a multimodal model for training as described in the work of Jia et al.[6] The further improvements made by the DiHT model will be detailed below.

3.1 Filtering

DiHT's first improvement is to pre-train the model by filtering out any samples from the dataset that removes non-informative image text pairs. There are two layers of filtering that the data will undergo. Firstly, a complexity step that removes image text pairs that are insufficently complex and a text to image that removes pairs wherein objects detailed in the text are displayed as text in the image.

3.1.1 Complexity filtering

The first step towards creating a filter for this information is to remove any text that is not informative. When training with very noisy datasets that are sourced from the web, often captions will be ungrammatical or unaligned which can create problems for model accuracy. DiHT seeks to remedy this by implementing a rule based parser in order to determine if a caption has sufficent complexity and as such is more likely to be aligned to the image. The parser defines the complexity as the highest number of relations for a single object. For instance, the sentence "The quarterback kicked the small brown football" has three tiers of complexity as the word "football" has the relations "small", "brown" and "the quarterback kicked" As "small" and "brown" are both attributes of the object while "a quaterback kicked" is the subject performing an action on the object. The DiHT parser is able to identify the complexity tier of the caption and filters out any captions that do not have a complexity tier of at least one. In other words, captions that lack more than a single relation in a sentence are not considered for training.

3.1.2 Text detection filtering

For this filter DiHT will utilize the MMOCR[7] open source toolbox for text detection. The goal is to remove any image text pairs wherein the given text can be found within the image itself (for example the word "cat" is spelled out in the image). We wish to remove these examples as it could make the model trained to spot text in images rather than to recognize object in images which could reduce the accuracy of the model. Since contrastive learning is sourcing our dataset from large noisy web based data, it is inevitable that such examples will exist in training data. To prevent this, the MMOCR is utilized to identify text within images. Any text within an images that matches a caption with at least .8 accuracy is removed from the dataset.

3.2 Concept distillation

One major problem with contrastive learning is that the captions we utilize to train our data only contains sparse information, often limited to a few attributes or object. This means that the amount of actual information that can be obtained from these image-text pairs is equally spares. However, this can be remedied by utilizing an additional linear classifier that's role is to identify objects and attributes in images. These identified objects and attributes are thereby added to the captions, enriching the originally sparse text data and often making them more informative than the original pairs.

The specific architecture that DiHT uses as a linear classifier is a strong, publicly available unimodal vision model ViT-H/14[3] which is pre-trained using the SWAG dataset[14]. DiHT then is able to utilize these trained linear models to generate probability vectors p^{obj} and p^{attr} which correspond to the objects and attributes predicted by the classifier for each image. The top k objects and attributes (k is manually chosen). Then we can add these objects and attributes to the captions to create more informative texts. It is important to note that this linear classifier will be trained alongside the training of the DiHT model so that predicted features can utilize a fully trained linear classifier to distil concepts from images. This will be reflected in the final training objective functions.

3.3 Loss function

In contrastive learning, the InfoNCE[10] has become the most popular loss function to be optimized. The function given a batch of images and texts $X = \{(x_i, t_i)\}$ will learn the temperature $Temp > 0$ by optimizing the following function:

$$L_{NCE}(X) = -\sum_{i=1}^{n}[log\frac{e^{x_i^T t_i/Temp}}{\sum_j e^{x_i^T t_j/Temp}} + log\frac{e^{x_i^T t_i/Temp}}{\sum_j e^{x_j^T t_i/Temp}}]$$

While popular however, there exists a major problem with this formation of loss: large scale noisy datasets like the ones we employ can provide negative samples that are not discriminative. In other words, certain datasets that are negative (wherein the image does not match the text) can not contribute to learning a separator and in fact actually decrease model accuracy. To combat this, larger batch sizes are often required.

Many of these cases often come from samples that are "easier". This is because "harder" pairs of images and text are often similar other pairs that are not negative examples. These pairs are often more discriminative and thus often better for training as demonstrated by Robinson et al.[12] Thus an alternative solution presents itself: we can add an weight factor to this loss function such that "hard" pairs, or rather pairs with higher similarity to other pairs are emphasized while other "easier" pairs are not.

The new loss function will require some given $\alpha \in (0, 1]$ and $\beta >= 0$ which will manually be determined beforehand. Thus the loss function in DiHT is the following alternative to InfoNCE:

$$L_{HN\_NCE} = -\sum_{i=1}^{n} log[\frac{e^{x_i^T t_i/Temp}}{\alpha * e^{x_i^T t_i/Temp} + \sum_{j \neq i} e^{x_i^T t_j/Temp} * w_{x_i,t_j}^{i \to t}}] - \sum_{i=1}^{n} log[\frac{e^{x_i^T t_i/Temp}}{\alpha * e^{x_i^T t_i/Temp} + \sum_{j \neq i} e^{x_j^T t_i/Temp} * w_{x_j,t_i}^{t \to i}}]$$

Wherein the weight functions are:

$$w_{x_i,t_j}^{i \to t} = \frac{(n-1) * e^{\beta x_i^T t_j/Temp}}{\sum_{k \neq i} e^{\beta x_i^T t_k/Temp}}, w_{x_j,t_i}^{t \to i} = \frac{(n-1) * e^{\beta x_j^T t_i/Temp}}{\sum_{k \neq i} e^{\beta x_k^T t_i/Temp}}$$

We can notice that this loss function is identical to InfoNCE if $\alpha = 1$ and $\beta = 0$. We are also still trying to learn $Temp > 0$ as the temperature. The difference is the existence of a new weight which is dependent on β that is used to penalize negative pairs which are "easy" and emphasize those that are "hard" which is determined by their similarity to other, positive pairs. α is necessariy to rescale normalization with positive terms when false negatives are present in the data. All of this accounted for, the DiHT model is able to learn the temperature $Temp$ more effectively by discounting negative samples that give little information.

3.4 Training objectives

Our final training objective will leverage both the adjusted InfoNCE loss algorithm and the training of the concept distillation for both objects and attributes. As such DiHT will be minimizing:

$$L_{HN\_NCE}(X) + L_{CE\_O}(X) + L_{CE\_A}(X)$$

Where L_{HN_NCE} is detailed above. L_{CE_O} and L_{CE_A} is the loss function for the linear classifier for concept distillation and is defined as:

$$L_{CE\_O} = \sum_{i=1}^{n} crossentropy(p_i^{obj}, f_{obj}(x_i)), L_{CE\_A} = \sum_{i=1}^{n} crossentropy(p_i^{attr}, f_{attr}(x_i))$$

Where f_{obj} and f_{attr} are the linear classifiers for objects and attributes while p^{obj} and p^{attr} are the top k predicted objects and attributes (these are the true values). As such the final training objective takes into account both concept distillation and hard negative training improvements made by DiHT.

3.5 Conceptual captions

Most of the above detail the training step of the model which has already been performed beforehand on DiHT by Radenovic et al.[11] The following is our own contribution to this work. We will be testing the model against Google's conceptual captions dataset[9]. The conceptual captions is a dataset of more than 3 million image-text pairs. Images and descriptors are harvested from the web while the final labels are generated using Google's machine learning identifier.

For our purposes, we utilized the provided validation set containing 14000 samples. Importantly, not only were none of these images used in training but none of the captions were provided during training either. This will thus allow us to test DiHT's efficacy as a zero-shot solution.

4 Experiment

For our experiment, we chose to use Google Colab and Kaggle notebooks as they provided a convenient collaborative environment without the need to set up our own hardware. Due to limited computing resources, we divided our validation set into batches of 2000 images and executed our code in parallel on these two platforms.

To import the validation set, we utilized the Hugging Face dataset library. However, since the dataset only contained URLs of images and not the images themselves, we had to download the images individually from their respective URLs. This process was hindered by slow response times from some of the image servers, which significantly slowed down our evaluation pipeline. Additionally, we found that roughly 20% of the image URLs were invalid and did not correspond to any actual images, further complicating the process.

Once we had the data, we passed it through the retrieval evaluation module that was modified based on the one provided in the DiHT repository. The original method was developed for datasets like MS-COCO and Flickr30K where each image is associated with multiple captions. However, in Google's Conceptual Caption dataset, each image has only one caption.

The evaluation module computed the similarity between the encoded image and text features provided by the *diht_vitl14_336px* model. These similarity scores were then subjected to Recall@1 eval metric that calculates the fraction of the top most relevant items out of the sample size.

5 Results

Following are the results for text-to-image(T2I) and image-to-text(I2T) retrieval for Google's Conceptual Captions dataset using the *diht_vitl14_336px* model from DiHT[10]:

Table 1: DiHT Evaluation on Google's Conceptual Captions Dataset

Image batch	GCC T2I (Recall@1)	GCC I2T (Recall@1)
0 to 2000	62.8	64.1
2000 to 4000	64.6	65.0
4000 to 6000	63.3	65.9
6000 to 8000	62.8	65.1
8000 to 10000	63.4	66.3
10000 to 12000	62.5	65.0
12000 to 14000	61.7	64.3

Therefore, the average recall for text-to-image(T2I) retrieval is **63.01%** and for image-to-text(I2T) it is **65.1%**.

6 Conclusion

The results of our evaluation indicate that the performance of the DiHT model for the conceptual captions dataset is almost similar to the evaluation on COCO(Recall T2I: 49.3% I2T: 65.3%) and Flickr30K(Recall T2I: 78.2% I2T: 91.1%) datasets carried out by the authors. This indicates that the model is quite robust even for the dataset it is not trained on.

In this evaluation study, we had the chance to gain insights on how to effectively import and preprocess a dataset. Additionally, we learned how to evaluate a zero-shot learning model on retrieval tasks. In the future, we would like to improve the data preprocessing methodology so that the evaluation pipeline is not slowed down.

7 Contributions

Andrew Fang: Presentation and Paper
Kanad Naleshwarkar (Coordinator): Coding, Experiment and Paper
Sanket Bhujbal: Paper

8 Code repository

The Jupyter notebook for our code can be found here: `https://github.com/kanadn/DiHT-GCC`

References

[1] Alec, R. & Jong, K. & Chris, H. & Aditya, R. & Gabriel, G. & Sandhini, A. & Girish, S. & Amanda, A. & Pamela, M. & Jack, C. & Gretchen, K. & Ilya, S. (2021) Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning. arXiv:2103.00020.

[2] Chen, Y. & Li, L. & Yu, L. & Kholy, A.E. & Ahmed, F. & Gan, Z. & Cheng, Y. & Liu, J. (2020) UNITER: UNiversal Image-TExt Representation Learning. Microsoft Dynamics 365 AI Research. arXiv:1909.11740.

[3] Dosovitskiy, A. & Beyer, L. & Kolesnikov, A. & Weissenborn, D. & Zhai, X. & Unterthiner, T. & Dehghani, M. & Minderer, M. & Heigold, G. & Gelly, S. & Uszkoreit, J. & Houlsby, N. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. ICLR. Google Research, Brain Team. arXiv:2010.11929

[4] Frome, A. & Corrado, G. & Shlens, J. & Bengio, S. & Dean, J. & Ranzato, M. & Mikolov, T (2013) DeViSE: A Deep Visual-Semantic Embedding Model. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26*. Dutchess County, NY: Curran Associates, Inc.

[5] Hinton, G. & Vinyals, O. & Dean, J. (2015) Distilling the Knowledge in a Neural Network. Google Inc. arXiv:1503.02531

[6] Jia, C. & Yang, Y. & Xia, Y. & Chen, Y.T. & Parekh, Z. & Pham, H. & Le, Q.V. & Sung, Y. & Li, Z. & Duerig, T (2021) Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. arXiv:2102.05918

[7] Kuang, Z. & Sun, H. & Li, Z. & Yue, X. & Lin, T.H. & Chen, J. & Wei, H. & Zhu, Y. & Gao, T. & Zhang, W. & Chen, K. & Zhang, W. & Lin, D. (2021) MMOCR: A Comprehensive Toolbox for Text Detection, Recognition and Understanding. arXiv:2108.06543

[8] Li, J. & Li, D. & Xiong, C. & Hoi, S. (2022) BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. ICML. arXiv:2201.12086

[9] Ng, E. & Pang, B. & Sharma, P. & Soricut, R. (2020) Understanding Guided Image Captioning Performance across Domains. Google Research. arXiv:2012.02339

[10] Oord, A.V.D. & Li, Y. & Vinyals, O. (2019) Representation Learning with Contrastive Predictive Coding. Google Deepmind. arXiv:1807.03748

[11] Radenovic, F. & Dubey, A. & Kadian, A. & Mihaylov, T. & Vandenhende, S. & Patel, Y. & Wen, Y. & Ramanathan, V. & Mahajan, D. (2023) Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training. Meta AI. arXiv:2301.02280.

[12] Robinson, J. & Chuang, C.Y. & Sra, S. & Jegelka, S. (2021) Contrastive Learning with Hard Negative Samples. ICLR. arXiv:2010.04592

[13] Singer, U. & Polyak, A. & Hayes, T. & Yin, X. & An, J. & Zhang, S. & Hu, Q. & Yang, H. & Ashual, O. & Gafni, O. & Parikh, D. & Gupta, S. & Taigman, Y. (2022) Make-A-Video: Text-to-Video Generation without Text-Video Data. Meta AI. arXiv:2209.14792

[14] Singh, M. & Gustafson, L. & Adcock, A. & Reis, V.D.F. & Gedik, B. & Kosaraju, R.P. & Mahajan, D. & Girshick, R. & Dollar, P. & Maaten, L.V.D. (2022) Revisiting weakly supervised pre-training of visual perception models. Meta AI. CVPR, arXiv:2201.08371.