Knowledge distillation Experiments

Fysikoudi,Eleni
University of Gothenburg
gusfysel@student.gu.se

Abstract

In recent years, the focus of machine learning has shifted from prioritizing neural networks that maximize performance to ones emphasizing the design of resource-efficient models that achieve comparable levels of accuracy while reducing computational and memory demands. Several methods have been developed to address this challenge, such as knowledge distillation (KD). This paper explores the application of KD to the multimodal task of binary image-text matching, aiming to evaluate its effectiveness. Comparing a fine-tuned state-of-the-art model to a smaller distilled model, we uncover that it is possible to bridge the gap in performance through student-teacher learning.

1 Introduction

Large-scale machine learning models excel in performing a vast variety of Natural Language Processing and Computer Vision tasks. However, those models come with a high cost. Specifically, they require a lot of training with big GPUs, which are often impractical in real-life applications due to the resources needed to deploy them. To tackle this issue, there has been increasing interest in developing well-performing models with less cost as well as methods that require less computational cost. Some instances of that are quantization of existing neural networks (Wu et al., 2016) and knowledge distillation (KD) (Hinton et al., 2015).

Quantization involves representing weights and activations with lower-bit data types (e.g., 8-bit instead of 32-bit), enabling faster computations and reduced memory usage with minimal loss in accuracy whereas KD involves training smaller and cheaper models(students) to mimic larger and cumbersome models(teachers). There have been different positive and negative assumptions for both methods. However, this paper is interested in leveraging knowledge distillation and examining how efficient it is in comparison to a fine-tuned model. Research has already illustrated that KD can be beneficial and provide good results. Nevertheless, it has been mostly used with very similar in complexity and parts teacher-student models. Furthermore, knowledge distillation has been used broadly for multi-classification tasks but not as much for binary.

One of the key questions my experiments want to answer is if two very different models can be used effectively and whether and how Kullback-Leibler(KL) Divergence loss can be utilized for a binary classification task. In addition, this work examines how these two models compare in performance by a series of steps. First, by fine-tuning CLIP, a powerful model trained on a vast amount of image-text pairs, then fine-tuning a resnet-miniLM smaller model as a baseline on the same data and finally performing knowledge distillation and checking how it affects metrics. The findings from this research will contribute to a deeper understanding of knowledge distillation in the context of binary classification tasks and provide insights into optimizing distillation techniques.

The paper is organized as follows. Section 2 describes the dataset creation process used for the experiments. Section 3 outlines the methodology and the details and intricacies of the models. Section 4 presents the results of the experiments conducted. Section 5 discusses the findings, and Section 6 concludes the paper and considers potential future work that should be done to confirm the results as well as testing different approaches.

2 Dataset

The task of image-text matching usually involves trying to match an image and a text within a batch. In my research, I am exploring matching and mismatching image and text relations where the model has to predict whether the two modalities correlate or not. As there was no availability of such a dataset, the only option was creating one. In or-

Figure 1: Example of a match and mismatch in the dataset

der to achieve that, the Microsoft COCO(MsCoco) captions dataset was utilized. There are "413,915 captions for 82,783 images in training, 202,520 captions for 40,504 images in validation and 379,249 captions for 40,775 images in testing" (Chen et al., 2015).

The data was processed as such that first there was only one caption kept for each image, then the dataset was shuffled and split in half so both labels were balanced. The next step was extracting and creating the matching and mismatching pairs which was done randomly. Figure 1 is a plot of two images in the dataset along with their labels and captions, one is a match and the other a mismatch. While this approach allows for the creation of a functional dataset, it does have limitations. Specifically, the dataset lacks challenging examples such as hard negatives or contrastive pairs, which could make the task more demanding for the model.

3 Methodology

The following three experiments, conducted using the created image-text matching dataset, aim to test the hypothesis that while a fine-tuned model may achieve the highest performance, knowledge distillation can enable a smaller model to achieve nearly comparable performance to the larger, more complex model.

3.1 Contrastive Language-Image Pre-training (CLIP)

CLIP is one of the state-of-the-art multimodal machine leaning models that was trained to associate images and texts in a shared embedding by using contrastive learning (Radford et al., 2021). In this research, it was fine-tuned to the dataset to reach the best performance and serve as a teacher to the baseline model whose objective would be reaching similar results.

The dataset was preprocessed for working with CLIP by using the pretrained processor to create the input ids, attention masks and pixel values which were used during training. The model was trained for 5 epochs using the Adam optimizer, Binary Cross entropy loss(BCE) with logits and a learning rate of 5e-5, ensuring stability and convergence during the fine-tuning process.

To evaluate the model, standard metrics were used such as accuracy, precision, recall and F1 score as well as a qualitative analysis of a sample from the errors. The objective was verifying the hypothesis that the model will perform well on the task so it can be used for teaching the smaller model.

3.2 Baseline Resnet-miniLM

The baseline and student model chosen to support the hypothesis of learning through knowledge distillation was a custom model that combines Resnet and MiniLM.

Particularly, the dataset was preprocessed to match the expected input of the models by using a custom processor that converted the images to colour, then resizing and normalization of the tensors as well as the AutoTokenizer required by MiniLM for creating input ids and attention masks for the captions. Resnet was used after removing the classification layer to extract image features while MiniLM was used for textual embeddings. Then, the image and text features were simply con-

catenated and passed through a linear layer. Choosing not to leverage attention mechanisms in this case was an architectural decision to test the hypothesis that with the correct hyperparameters, a complex teacher can successfully transfer knowledge to a simpler student.

The model was then trained by fine-tuning its components for 10 epochs at a learning rate of 1e-4 and again using the Adam optimizer and BCE loss with logits to create a baseline. The evaluation was conducted with the same metrics but without a qualitative analysis.

3.3 Distillation methods and experiments

Knowledge distillation can be implemented through various methods such as feature matching and logit matching. Feature matching involves transferring knowledge in intermediate layers using Mean Squared Error (MSE) Loss with the idea that the student model will learn to replicate the embeddings of the teacher model (Tran et al., 2021). Logit matching on the other hand, involves using Kullback-Leibler(KL) Divergence loss by "transfering the generalization ability of the cumbersome model to a small model using the class probabilities produced by the cumbersome model as “soft targets” for training the small model" (Hinton et al., 2015). This method also compromises of temperature scaling for applying the "softness" to the logits.

The experiments conducted were initially aimed at using the latter method with KL Divergence loss to align teacher and student distributions but due to not reaching the expected results was changed as it will be described below. As aforementioned, the teacher model was the fine-tuned CLIP mentioned in Section 3.1 and the student model was the baseline described in 3.2. The dataset was preprocessed so it can be used for both models which means that there are two different input ids and attention masks corresponding to each model as well as two different image transformations for the pixels. The teacher model was set to evaluation mode to freeze its components and not update its gradients. As before the Adam optimizer was used with BCE logits loss and a learning rate of 1e-5. All experiments were run on a GPU.

The first experiments, focused on using KL Divergence as described in (Hinton et al., 2015) and the tutorial in Link to tutorial from Pytorch, however, that approach is only appropriate for multi-classification. This method comprises of combining two losses, the classification loss and the KL loss with the hyperparameter alpha(α) which defines the weight of their contribution in the final total loss function. I tried numbers from 0.4-0.6 for the variable alpha. The approach also calls for using softmax to the logits of the teacher with a temperature which in my experiments I set from 1.5-3.0 and log softmax to the logits of the student model with the same temperature.

The second batch of experiments was inspired by this approach and I endeavoured to tailor it to binary classification. The key changes were applying sigmoid to the logis of the teacher and student instead of softmax to get the probabilities and then instead of using KL divergence loss for the distillation loss, I opted for using binary cross entropy loss to measure the difference between their probabilities. The same temperature and alpha ranges were tested. The most successful configuration combined a temperature of 3.0 and a balanced alpha of 0.5 for the combination of losses. This configuration also verified Hinton’s hypothesis that when the difference between models is bigger using intermediate temperatures work best which strongly suggests that ignoring the large negative logits can be helpful (Hinton et al., 2015).

The standard evaluation metrics were used for inference as before and compared to the teacher and baseline.

4 Results

In this section, the results of the experiments on knowledge distillation for binary classification are presented and analyzed. The primary aim is to evaluate the performance of the student model compared to the teacher model and a baseline model, highlighting the impact of key hyperparameters such as temperature and alpha on the distillation process. The three models that were evaluated as described before are the CLIP model, a fine-tuned version of the student model serving as the baseline and the same model trained through knowledge distillation. In table 1, there is a general overview of the metrics.

Metrics	Accuracy	Precision	Recall	F1 Score
CLIP	86.40%	84.06%	89.84%	86.85%
Baseline	49.22%	49.27%	52.88%	51.01%
Distillation	85.56%	80.26%	94.32%	86.72%

Table 1: Evaluation Metrics of experiments

The CLIP model as expected only with 5 epochs performs around 85% in all metrics and upon qualitative investigation seems to make mistakes on accidentally made hard negatives like having a woman in the image and the caption saying a woman sitting in a train.

The baseline fine-tuned model performs poorly despite being trained for 10 epochs and achieves less than 50% in most metrics. After checking its predictions, and in combination with the training loss I suspect that it did not generalize well. The lack of any attention mechanism is also a possible reason.

The student model after the distillation process as visible from the metrics almost reached the same results (about 85%) as CLIP which suggests that knowledge distillation significantly improves the student model's performance, especially when alpha is balanced and a temperature of 3.0 is applied. The hypothesis of being able to distill knowledge to a very different and less complex model is confirmed at least for this particular case study.

5 Discussion

Taking into consideration the results from the models, someone can conclude that knowledge distillation does perform well if the correct hyperparameters and methods are chosen. For instance, the first few experiments I conducted were performing only slightly better than the baseline so it is more than important to consider the method and parameters utilized. Based on the first experiments, we also conclude that Kullback-Leibler Divergence can not successfully be used for binary classification tasks.

This research contributes to the methodology of knowledge distillation as I have not seen a lot of use in binary tasks such as this one. Furthermore, it gives ground for supporting the claim that any model can learn from another, without particularly needing a lot of training. Some of the research has suggested that distillation requires a big amount of training(Beyer et al., 2022) but this experiment illustrates that it is possible to do it in less epochs similar to fine-tuning. More experiments should take place but knowledge distillation seems to be a valuable method to minimize computational cost.

6 Conclusion

To summarize, the purpose of this research was to examine knowledge distillation of logits in challenging settings such as different capacity of models, limited epochs training, and lack of complex student model.

The research focused on implementing KD in a multimodal binary classification task and proposed a practical approach tailored for such scenarios.

The results have shown that the performance gap between a fine-tuned teacher model and a simple student model can be shrunk using knowledge distillation even with restrictions.

Despite these results being fruitful, more work still needs to be conducted in order to determine the accuracy of these results on more datasets, tasks and models with varying architectures. Other methods such as feature alignment or feature and logits alignment simultaneously can be investigated for comparison and further improvement of the process of distillation.

As we can see, this work adds to the existing research on knowledge distillation by outlining some of the bounds and uses of the technique in binary classification and defines the directions for further exploration in the sector of multimodal learning. In this link, all the code is available as well as information on how to run it (Github Repository).

References

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. 2022. Knowledge distillation: A good teacher is patient and consistent.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Dat Thanh Tran, Moncef Gabbouj, and Alexandros Iosifidis. 2021. Knowledge distillation by sparse representation matching.

Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. 2016. Quantized convolutional neural networks for mobile devices.