24100219

Click here to access the repository: Git Repository Link

Enhancing Vision Transformers through Head-Specific Knowledge Distillation

Course Project for CS 437: Deep Learning

Rehan Ahmad

August 29, 2024

Contents

1 Introduction

In a TICC lecture delivered at Google Inc, Geofrey Hinton began by elaborating upon the title of his talk which was Dark Knowledge. He says that 95 percent of what the neural network learns is not what you were trying to get it to learn. The knowledge per parameter of deep learning models is extremely low and there is a lot of redundancy encoded in them. This redundancy is amplified in architectures like Vision Transformers (ViTs) due to their vast number of parameters.

Adopted for vision by the paper, "An Image is worth 16 x 16 words" (Dosovitskiy et al. [2020]), ViTs decompose an image into sequence of patches and treat each patch as a token in a sequence. Subsequently in the transformer block, attention is applied to compare each patch with other patches which enables the model to weigh the importance of one part of the image relative to others. After this, a classification token which interacts with all the other patches in the attention heads is used to perform classification in the MLP head. While Dosovitskiy, Alexey, et al were able to achieve State-of-the-art result on ImageNet using this methodology, they utilized a large private dataset JFT with 18k classes and 303 M high-resolution images in comparison to ImageNet's mere 1.3 Million labelled images. Training on such a large dataset requires large training times and highlights the expensive nature of attention.

The process of knowledge distillation mitigates this by compressing the knowledge of deep learning models into smaller more dense representations. It does this by training a small model to mimic a pre-trained, larger model or ensemble of models. This training setting is sometimes referred to as "teacher-student" where the large model is the teacher, and the small model is the student. While the objective introduced by the seminal work of Hinton et al. [2015] minimized the KL divergence between the teacher and student outputs, KL divergence is undefined when transferring knowledge about a representation. In the paper Contrastive Representation Distillation (Tian et al. [2019]), the authors derive an objective that maximizes a lower-bound to the mutual information between the teacher and the student representation, with their method outperforming 12 recent distillation methods that it was benchmarked against. Similarly, Malik et al. [2020] impart knowledge from a well-trained large network to smaller networks using the dense representation learned by the teacher network which resides in the layer before the logits. To do so, they split the dense representation of the teacher into mutually exclusive subspaces and apply reconstruction loss to each of the student's representation and the representation of the teacher's subspaces.

In the context of ViTs, approaches such as "Training Data-Efficient Image Transformers" (Touvron et al. [2021]) and "DearKD" (Chen et al. [2022]) focus on knowledge distillation between CNN's representations and Vision Transformers. However, there is a noticeable gap in the literature concerning the precise replication of the dense representations found in attention mechanisms. In our view, this does a great disservice to the rich dynamics of attention and how the Query and Key matrices interact with each other to produce the final attention weights. This interaction is crucial as it forms the backbone of the Transformer architecture and influences how information is weighted and integrated across different parts of the input. The blogpost Explainability for Vision Transformers by Jacon Gildenblat integrates a collection of noteworthy attempts by deep learning community to visualize this rich information flow. Examples include attention rollout as introduced by Abnar and Zuidema [2020] and Class Specific Explainability introduced by Chefer et al. [2021]

Our aim is to transfer knowledge by using the attention head representation from a teacher model to guide the training of a student model. In order to do so, we transfer the representation of the attention head of an already trained teacher model to enhance the corresponding representation in attention head of a student model. The details of our approach are given under methods and the subsequent experiments we have carried out in this limited time are given under results. We discuss our plan for the future in suggested improvements and conclude by giving our progress reports.

2 Proposed Method

Consider a teacher model ViT and a student model ViT. Let n be the number of heads in the teacher model and m be the number of heads in the student model where we impose the constraint:

$$m \leq n \tag{2.1}$$

For the teacher model, for each attention head in $\{h_{t_1}, \ldots, h_{t_n}\}$, we have the Query, Key, and Value Projections:

$$K_{t_i} = X_t W_{t_i}^K, \quad Q_{t_i} = X_t W_{t_i}^Q, \quad V_{t_i} = X_t W_{t_i}^V \tag{2.2}$$

Where X_t is the input to the teacher's attention mechanism and $\{W_{t_i}^K, W_{t_i}^Q, W_{t_i}^V\}$ are the projection matrices for keys, queries, and values associated with each head h_i. The attention weights A_{t_i} and the outputs O_{t_i} associated with the heads are the following:

$$A_{t_i} = \text{Softmax}\left(\frac{Q_{t_i} K_{t_i}^T}{\sqrt{d_k}}\right) \tag{2.3}$$

$$O_{t_i} = A_{t_i} V_{t_i} \tag{2.4}$$

These outputs $\{O_{t_1}, \ldots, O_{t_n}\}$ are then concatenated and multiplied with the matrix W^Q to produce the output of the layer. Similarly, the student model will have its own attention mechanism and the associated outputs:

$$K_s = X_s W_{s_i}^K, \quad Q_s = X_s W_{s_i}^Q, \quad V_s = X_s W_{s_i}^V \tag{2.5}$$

$$A_{s_i} = \text{Softmax}\left(\frac{Q_{s_i} K_{s_i}^T}{\sqrt{d_k}}\right) \tag{2.6}$$

$$O_{s_i} = A_{s_i} V_{s_i} \tag{2.7}$$

For $\{h_{s_1}, \ldots, h_{s_m}\}$, During the forward pass, we extract the attention weights of the teacher $\{A_{t_i}, \ldots, A_{t_n}\}$ for each block and distill this representation into the student model by directly making the attention maps of the student $\{A_{s_i}, \ldots, A_{s_m}\}$ block mimic these. Our approach allows us the flexibility of being able to isolate particular attentions maps A_{s_i} and distilling the representation the teacher's representation only for them. We believe that this can avoid diluting the distinctive learning potential of other heads in the student model since each attention head should ideally learn unique aspects of the data; if every head were to mimic the same teacher head, it could compromise the model's capacity to capture diverse representations and reduce its effectiveness.

Hence in our experiments, we explore different attention-distillations. In particular, we explore head-to-head attention map distillation, block-to-block attention map distillation, and encoder-to-encoder distillation. In head-to-head distillation, we transfer the attention map of only a particular head h_{t_i} onto the student's head h_{s_i}. For block-to-block attention, we impose all the heads $\{A_{s_i}, \ldots, A_{s_m}\}$ in the student's block to follow patterns of teacher's heads $\{A_{t_i}, \ldots, A_{t_m}\}$ where we choose m blocks out of the n blocks for the teacher for distillation. Finally, for encoder-to-encoder representation, we make all the blocks in the student's head mimic those of the teacher. These different levels of granularity in distillation can allow us to analyze which scales of knowledge transfer are most beneficial for the student model's generalization capabilities.

For the knowledge transfer between the maps, we considered and experimented with three loss functions. We employed the mean squared error (MSE) to evaluate the discrepancy between the teacher's attention maps and student's attention map before the softmax was applied. That is, we extract the unnormalized attention scores of both the teacher and student:

$$a_{t_i} = \frac{Q_{t_i} K_{t_i}^T}{\sqrt{d_k}}, \quad a_{s_i} = \frac{Q_{s_i} K_{s_i}^T}{\sqrt{d_k}} \tag{2.8}$$

Subsequently, we compute the MSE loss between them:

$$L_{\text{att}}(a_{t_i}, a_{s_i}) = \frac{1}{m} \sum_{i=1}^{m} (a_{t_i} - a_{s_i})^2 \tag{2.9}$$

In other experiments, we take the normalized probabilities and compute cross-entropy and Kullback-Leibler Divergence:

$$L_{\text{Att}}(A_{t_i}, A_{s_i}) = -\sum_i A_{t_i} \log(A_{s_i}) \tag{2.10}$$

$$L_{\text{Att}}(A_{t_i}, A_{s_i}) = -\sum_i A_{t_i} \log\left(\frac{A_{t_i}}{A_{s_i}}\right) \tag{2.11}$$

We then follow the same procedure as in normal ViTs. The output vector corresponding to the [CLS] token $\vec{z}^{\text{CLS}}$ is passed through a dense layer to produce logits for each class:

$$\vec{z} = W \vec{z}^{\text{CLS}} + \vec{b} \tag{2.12}$$

Each element in the logit is transformed into probability p_{c_i} using softmax and subsequently the classification loss is computed:

$$L_{\text{class}} = \sum_{c=1}^{C} y_{c_i} \log(p_{c_i}) \tag{2.13}$$

For our model, the total loss therefore becomes the following:

$$L_{\text{Tot}} = \alpha L_{\text{class}} + (1 - \alpha) L_{\text{Att}} \tag{2.14}$$

Where α is a hyperparameter. A higher value of α causes our model to focus on minimizing L_{class} while a smaller value constrains it to mimic the attention maps of the teacher more closely. A section

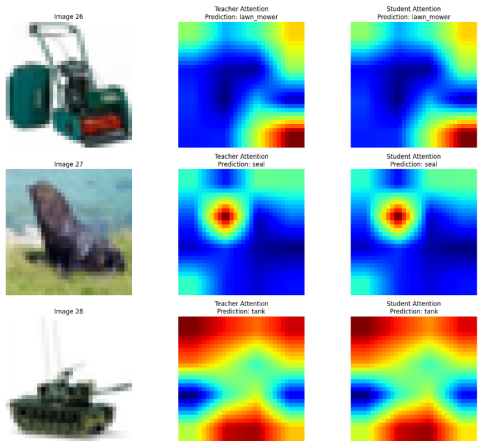
3 Experiments and Results

We initially conducted experiments using the MNIST dataset which comprises 28x28 grayscale images. We employed patches of size (4, 4), resulting in a total of 49 tokens (7 x 7). Positional embeddings and a CLS token were incorporated. We then investigated the performance across different embedding dimensions—64, 128, and 256. Due to the scalable nature of Vision Transformers (ViTs) even on simpler datasets like MNIST, almost all tested architectures consistently achieved accuracies above 95 %. This high performance precluded further attempts to enhance the accuracy of the smaller models through attention-based training. Consequently, we turned our attention to the CIFAR-10 dataset. However, we encountered similar challenges with all models achieving roughly equivalent accuracies.

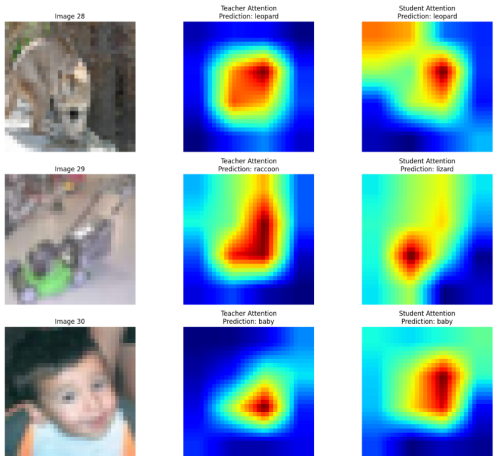
Eventually, we settled upon CIFAR-100 dataset for which we were able to find a suitable ViT student-teacher model with a noticeable accuracy difference. The CIFAR 100 data consists of (32 x 32 x 3) images which we converted into patches of size (4 x 4) resulting in a total of 16 token (not accounting for CLS token). For both our teacher and student model, we settled upon the embedding dimensions of 64 since decreasing or increasing them to 32, 128 or 256 were not effecting the accuracy of our model. For the teacher, the increased parameters were in the MLP which takes the output from encoder to feedforward layer of dimension 1024 in comparison to the student's 512. Furthermore, the teacher model consisted of 8 attention heads in comparison to the student's 4. The total parameter of the teacher model were 2.6 M and the total parameters of the student was 0.6 M. Dropout of 0.1 was applied while training both of the models.

Without our approach, the teacher model achieved a top accuracy of 46.18 % in 100 epochs while the student model achieved a top accuracy of 41.08 % in the same number of epochs. The Batch Size was set to 64, the learning rate to 0.001 and the optimizer utilized was AdamW for both student and teacher model. With our distillation approach, the student model consistently saw an accuracy boost of +1.00 % with different variations of attention distillation and loss functions. Under the submission "ts-cifar100-mse", you would find that using Mean Square Loss and encoder-to-encoder distillation approach, we were able to attain a top accuracy of 42.86 % (α=0.00015). Similarly, we utilized block-to-block distillation and attained an accuracy of 42.17 % ($\alpha = 0.015$) with Mean Square Loss. We also carried out the experiment with KL divergence using block-to-block distillation approach, attaining a top accuracy of 43.33 % ($\alpha = 0.5$).

In order to ensure that the accuracy boost we were achieving was due to the student mimicking the teacher pattern, we examined both the teacher's and the student's attention maps. Since both of them had (8 x 8) patches as their input, the attention map for the teacher and the student were tensors of dimensions [64, 8, 17, 17] and [64, 4, 17, 17] respectively. For head-to-head distillation, we extracted both these maps and averaged their representation across heads. We then reshaped them into 2D format so they correspond to the spatial dimensions of the input image. To make the attention maps the same size as the original input images for easier comparison, we scaled them up using interpolation. This in line with our understanding of the Blog Explainability in ViT as elaborated in our notebook **1.1) Explainability in ViTs**. For fair comparison, we considered a teacher-student block for which we carried out distillation 1a and a non-student block which learned its representation of the data independently of any teacher model 1b.



(a) Attention map for the teacher-student block used in distillation



(b) Attention map for a non-teacher-student block learning independently

Figure 1: Comparison of attention maps in teacher-student and non-teacher-student settings.

As can be observed from the figures, the attention maps for which we compute attention distillation follow the teacher's representation exactly while for the non-teacher-student block, both the labels and representations diverge.

4 Future Implementations

Although our preliminary results show a great deal of promise in our attention-based distillation approach, there is a lot of room for improvement. Experimentally, we would like to carry out our attention-based distillation approach for a large collection of models with different number of parameters as well as approaches (such as SWIN and Transformer in Transformer). Furthermore, we would like to find similar small size datasets such as CIFAR-100 on which we can train to ensure that our methods are reproducible across datasets. Potential datasets that we aim to explore are Fashion MNIST and tiny ImageNet.

More importantly, there are a lot of conceptual improvements from which our methodology can benefit from. For instance, consider the image and the associated attention maps distilled by the teacher and student model shown in the figure 2. We observe that although the student and teacher model both converge to the same attention representation, the teacher's prediction and student prediction is not the same with the student misclassifying the image.

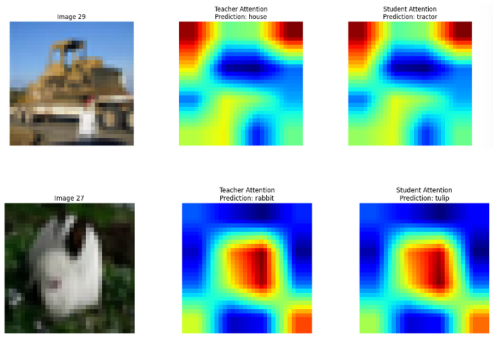


Figure 2: Student Misclassification

This phenomenon seems to indicate that the student might end up with similar attention but insufficient guidance on how to use that attention for accurate predictions. Potential reasons could be due to the lower number of parameters

in the MLP of the student which does not support as robust a decision-making process as the teacher. Interestingly, this occurs more during encoder-to-encoder distillation in which we don't give the student model as much of a free reign to integrate information across the image on its own. As such, it could be arising from lack of diversity or flexibility in allowing the student to integrate distinct representations across its head.

Besides this, we need to account for the phenomenon when the teacher whose distillation map we are transferring misclassifies an image 3.

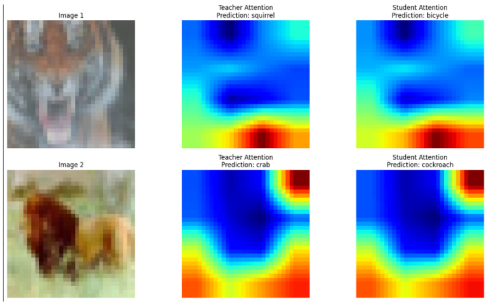


Figure 3: Teacher Misclassification

For this, we aim to explore adaptive distillation weighting in which we would the weight of the attention distillation loss based on the student's performance on the classification task. For example, we would reduce the influence of the distillation loss when the student's predictions are incorrect, encouraging the student to explore other patterns that might lead to correct predictions. Interestingly, our attention maps corroborate the insights from the paper, "Understanding Why ViT Trains Badly on Small Datasets" (Zhu et al. [2023]). The authors of the paper argue that lower layers of ViT can not learn the local relations well with a small amount of data and are more apt at capturing the "high-level overview" of the picture. This is exactly the patterns that our attentions map display, with often a clear demarcation of attention activation on the basis of scenes. Seeing this, we want to explore clustering of attention regions by the activity of attention activity there. For example, we might be able to take smaller [2 x 2] patches and consider the attention activity in these regions. We might then cluster them on the basis of attention activity with blue indicating less activity and red indicating maximum activity. For each of these categories, we want to incorporate a smaller ViT to focus on them individually. In this way, we might be able capture closer granular details that the representations of ViTs often miss out on for small datasets.

5 Progress Report

First Improvement: We implemented a baseline ViT in which we carried out patch level augmentation as introduced by the paper "Vision Transformer for Small-Size Datasets" (Lee et al. [2021]). In particular, we employed Shifted Patch Tokenization which spatially shifts images alongside the original image and then create combined patches consisting of the original patches and their accompanying shifts. This enhances the local inductive bias of the transformer, addressing one of the shortcomings of standard ViTs which is the handling of spatial relations. This is given in the notebook 1.05) ViT with Shifted Patch Tokenization.

Second Improvement: We carried out knowledge distillation for training ViTs on smaller datasets using Keras, focusing on the $tf_flowers$ dataset. With its modest size of 3,700 images across 5 classes , the $tf_flowers$ dataset offered a more manageable context for this experiment compared to larger datasets like Tiny ImageNet. We carried out the implementation of DEIT (Data-efficient Image Transformer) by Touvron et al. [2021]

Third Improvement: These are the entire contents of our final report. All the associated code can be found in the following repository.

6 Conclusions

In this project, we explored attention-based knowledge transfer from a teacher to a student model. Our methods have shown potential in improving ViTs model performance on the CIFAR-100 dataset where the student model consistently

achieved an accuracy boost. The process of knowledge distillation involved strategies such as head-to-head, block-to-block, and encoder-to-encoder attention map transfers. These strategies allowed us to tailor the distillation process according to the specific needs of the student model. Looking ahead, our focus will be on extending this methodology to a wider array of datasets and model configurations. Moreover, our observations from experiments where the teacher model misclassified images suggest that attention fidelity alone may not always correspond to improved accuracy. We plan to explore adaptive distillation techniques that dynamically adjust the influence of distillation based on real-time performance metrics to mitigate this. We also plan to explore clustering of attention regions based on their activity levels, using smaller [2 x 2] patches to potentially enhance the granularity of attention and improve model responsiveness to local features. Our future work will continue to refine these techniques, aiming to develop smaller ViTs with more robust attention representations.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: Data-efficient early knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12052–12062, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

Shaiq Munir Malik, Muhammad Umair Haider, Mohbat Tharani, Musab Rasheed, and Murtaza Taj. Teacher-class network: A neural network compression mechanism. *arXiv preprint arXiv:2004.03281*, 2020.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

Haoran Zhu, Boyuan Chen, and Carter Yang. Understanding why vit trains badly on small datasets: an intuitive perspective. *arXiv preprint arXiv:2302.03751*, 2023.