Week 11 Tuesday Paper 1:[1503.02531] Distilling the Knowledge in a Neural Network

Explained by Distilling the Knowledge in a Neural Network

Supplemental: Distillation as a Defense to Adversarial Perturbations against…

1. What is knowledge distillation and how is it achieved in this paper? Don't forget to explain the "temperature" and "soft targets".
Ans: Knowledge distillation is a process where a smaller, simpler model, referred to as the "distilled" model, is trained to reproduce the behavior of a larger, more complex model, termed as the "cumbersome" model. This approach is grounded in the idea that the cumbersome model's learned mapping from input vectors to output vectors contains valuable information, not just in the form of the correct answers(hard targets), but also in the probabilities assigned to each possible output(soft targets). These soft targets, especially when derived at a high "temperature," provide nuanced insights into the model's learned representations.

To achieve distillation, the paper proposes using the soft targets from the cumbersome model to train the distilled model. These soft targets are obtained by applying a softmax function at a high temperature to the cumbersome model's outputs, which generates a softer probability distribution over classes. This distribution is then used as the target for training the smaller model, using the same high temperature in its softmax function. The technique allows the distilled model to learn not just from the correct classifications but also from the structure of the probabilities assigned to incorrect classifications, effectively transferring the cumbersome model's generalization capabilities.

2. Where does overfitting come into the distillation story?
Ans: Overfitting is a crucial concern in the distillation process, especially when dealing with very large datasets or a significant number of classes. The paper discusses overfitting in the context of training specialist models that focus on making fine-grained

distinctions among a subset of classes. These specialist models are prone to overfitting due to their focus on a narrow slice of the data. To mitigate this, the paper suggests using soft targets as a form of regularization. By training specialist models with both hard targets (for their specific subset of classes) and soft targets (for the rest), the models can retain knowledge about the broader dataset and prevent overfitting on their specialized subsets. Soft targets, by providing information about the relationship between different classes, help to encode more generalizable knowledge into the specialist models.

3. What applications (apart from the examples mentioned in the paper) benefit from knowledge distillation? When is it less likely to be useful?

Ans: Other applications of knowledge distillation include,

- Mobile and embedded systems, where computational resources are limited.
- Real-time applications requiring low latency.
- Enhancing privacy by distilling knowledge from a model trained on sensitive data onto a model that can be deployed without direct access to the original data.

Knowledge distillation might be less useful when the target model has sufficient capacity to directly learn from the original dataset without the need for a cumbersome model. Additionally, tasks that require interpreting the model's decisions like regulatory compliance may prefer more transparent models over distilled ones, as the distillation process can obscure the rationale behind the model's predictions.

Week 11 Tuesday Paper 2:TinyViT: Fast Pretraining Distillation for Small Vision Transformers

Explained by (Not an exact explanation) Swin Transformer paper animated and explained

Supplemental: https://github.com/microsoft/Cream/tree/main/TinyViT, EfficientViT: Memory Efficient Vision Transformer With Cascaded Group Attention

1. Explain how positional encoding is used in vision transformers.
Ans: Vision Transformers utilize positional encoding to retain spatial information lost during the transformation of image patches into a sequence of embeddings. Positional encodings are added to the input embeddings to provide spatial context, enabling the model to learn the relative or absolute position of the patches in the original image

2. Explain what is meant by teacher and student models. How are they used in deep learning?
Ans: Teacher and student models refer to the components in knowledge distillation, where a larger, more complex model (teacher) transfers knowledge to a smaller, simpler model (student). In TinyViT, distillation during pre training is used for knowledge transfer, where the logits of large teacher models are sparsified and stored to guide the training of the tiny student transformers.

3. What is meant by sparse soft labels.
Ans: Sparse soft labels in TinyViT are generated by selecting the top-K values of the teacher model's output logits for each input image, significantly reducing the memory and storage cost while retaining most of the informative signals for distillation

4. Elaborate on the model architecture- what types of layers are used and why?
Ans: TinyViT features a hierarchical transformer architecture with four stages and utilizes mechanisms like window attention and attention biases for efficiency and effectiveness. The architecture is designed for a good trade-off between computational

cost and accuracy, leveraging techniques like MBConvs in early stages for efficient low-level feature extraction

MBConv Blocks: At the earliest stage, MBConv blocks (Mobile Inverted Bottleneck Convolution) are used. These blocks are a type of efficient convolutional layers popularized by MobileNets, which utilize depthwise separable convolutions for reducing computational cost. In TinyViT, MBConv blocks help in efficiently learning low-level features with strong inductive biases inherent to convolutional operations, making the model more parameter-efficient and faster.

5. What different ablation studies were done? What is the conclusion from each? Explain intuitively why the conclusions obtained meet expectations.
Ans: The TinyViT paper conducts several ablation studies to evaluate the impact of different aspects of the model and training methodology, including the effectiveness of pretraining distillation, the influence of the scale of pretraining data, the role of the number of saved logits, and the impact of different teacher models on the performance of TinyViT.

Effectiveness of Pre Training Distillation:

Pretraining with distillation significantly improves the performance of TinyViT models on the ImageNet-1k dataset, demonstrating the effectiveness of leveraging knowledge from large teacher models.

Distillation allows the small model to learn from the rich, high-level representations captured by the larger model, effectively compressing and transferring this knowledge. The improvement validates the hypothesis that small models can benefit from the distilled information, overcoming their inherent limitations in capacity.

Impact of Pre Training Data Scale:

The representation quality of TinyViT improves as the total number of images seen during pre-training increases, but there is a diminishing return effect, especially for the smaller model variant.

Increasing the amount of pretraining data allows the model to learn from a more diverse set of examples, leading to better generalization. However, the capacity of smaller models limits their ability to continuously benefit from additional data, leading to performance saturation. This aligns with the expectation that there's a balance between model capacity and the amount of data it can effectively leverage.

Effect of the Number of Saved Logits:

Storing and using a greater number of top-K logits from teacher models' predictions during distillation leads to better performance up to a point, after which increasing K offers diminishing improvements.

The top-K logits contain the most relevant information for distillation, capturing the teacher model's confidence distribution over the most probable classes. Initially, increasing K provides more nuanced information for the student to learn from. However, beyond a certain point, additional logits contribute less to learning, possibly introducing noise or redundant information.

Impact of Teacher Models:

The choice of teacher model affects the performance of the distilled TinyViT model, with more powerful teacher models yielding better-performing student models.

More powerful teacher models have learned richer and more accurate representations of the data, thus providing more informative guidance during distillation. This results in student models that not only mimic the teacher's output more closely but also inherit its ability to generalize from the training data to unseen examples. The correlation between the teacher's capability and the student's performance emphasizes the importance of the quality of guidance in knowledge distillation.

6. What is meant by linear probing? How is it used in the paper?
Ans: Linear probing refers to evaluating the quality of learned representations by training a linear classifier on top of the frozen features extracted by the model. It was used to assess the transferability and richness of the representations learned by TinyViT.

7. What is meant by zero-shot and few-shot learning? Which component of the architecture helps with domain adaptation? Explain in detail.
Ans: Zero-Shot Learning (ZSL) refers to the ability of a model to correctly make predictions for tasks or classes it has never seen during training. In ZSL, the model uses its knowledge about seen classes and their relationships with unseen classes to make inferences, often leveraging auxiliary information like class attributes or textual descriptions.

Few-Shot Learning (FSL), on the other hand, involves training a model on a very limited number of examples (shots) for each class. Unlike traditional learning scenarios that require large datasets to learn effectively, FSL aims to adapt the model to new tasks or classes with as few as one to five examples per class.

The ability of a model to excel at zero-shot and few-shot learning, especially in the context of domain adaptation, relies heavily on the representational richness of its learned features. For TinyViT, several components and design choices contribute to this capability but the knowledge distillation plays a crucial role.

Pretraining and Distillation: Pretraining on large-scale datasets with subsequent distillation from a larger teacher model equips TinyViT with a broad and generalizable feature space. The distilled knowledge, especially when derived from a teacher model experienced in a wide array of classes, embeds an understanding of various visual concepts and their interrelations within the student model. This foundational knowledge base is crucial for zero-shot learning, where the model needs to generalize to unseen classes, and for few-shot learning, where it must quickly adapt to new tasks with minimal examples.

8. How would you address the future directions stated in the conclusion of this paper.
Ans: They mentioned -In future work, we will consider using more data to further unlock the representability of small models with the assistance of more powerful teacher models. Designing a more effective scaling down method to generate small models with better computation/accuracy is another interesting research direction. There is nothing to address here as they seem like a good direction to head to.

9. What questions do you have?
Ans: Where else have teacher-student models been famously used?