DeiT: Training Data-Efficient Image Transformers & Distillation Through Attention - Briefing Document

Key Findings and Themes:
This paper addresses the challenge of training high-performing vision transformers (ViT) without requiring massive, privately curated datasets and extensive computational resources, which has been a limiting factor in their adoption. The authors introduce **Data-efficient image Transformers (DeiT)**, a method that achieves competitive results by focusing on improved training strategies and a novel distillation technique when trained on ImageNet-1k.

1. Training Vision Transformers on Smaller Datasets:

- **Problem:** Previous work (Dosovitskiy et al.) showed that vision transformers "do not generalize well when trained on insufficient amounts of data" and required pre-training on datasets like JFT-300M (300 million images).
- **DeiT Solution:** The authors demonstrate that it is possible to train convolution-free transformers on ImageNet-1k (1.28 million images) alone and achieve competitive results.
- **Efficiency:** DeiT models can be trained on a single computer (with 4-8 GPUs) in less than 3 days, making them significantly more accessible than previous ViT models.
- **Key Ingredients for Data-Efficient Training:** While not explicitly detailed in the provided excerpts, the abstract and Section 6 mention an "extensive ablation of our data-efficient training choices," highlighting factors like:
- Repeated Augmentation
- Rand-Augment or Auto-Augment
- Mixup
- Cutmix
- Stochastic Depth
- Appropriate initialization and hyper-parameters (AdamW optimizer with a smaller weight decay than ViT)

2. Novel Distillation Strategy: Distillation Token:

- **Problem:** Traditional knowledge distillation (KD) methods (soft and hard) can be applied to transformers, but the authors introduce a transformer-specific approach.
- **DeiT Distillation:** A new **"distillation token"** is introduced and added to the initial embeddings (along with patch and class tokens). This token interacts with others through self-attention.
- **Mechanism:** The distillation token's objective is to reproduce the **(hard) label predicted by the teacher model**, while the class token aims to reproduce the true label. Both tokens are learned through back-propagation.
- **Benefits:** This token-based distillation strategy significantly outperforms vanilla distillation methods.

- **Complementary Information:** The class and distillation tokens converge towards different vectors during training but become more similar at the final layer. This suggests they learn complementary information.
- **Comparison to Additional Class Token:** Adding a second class token aiming at the same target label does not provide a performance boost, indicating the unique contribution of the distillation token.

3. The Importance of the Teacher Model in Distillation:

- **Unexpected Finding:** The authors observe that using a **convnet as a teacher model gives better performance** than using another transformer with comparable performance.
- **Reasoning:** This is attributed to the transfer of inductive biases from the convnet to the transformer through distillation. Convnets possess strong inductive biases related to image processing due to their architecture (e.g., locality, translation invariance), which can be beneficial for transformers, especially when trained on limited data.

4. Competitive Performance and Efficiency:

- **ImageNet:** DeiT models, particularly when distilled, achieve results competitive with state-of-the-art convnets like EfficientNet on ImageNet-1k.
- **Distillation Advantage:** The distilled DeiT models (DeiT⚗) can even outperform their teachers and achieve a better trade-off between accuracy and throughput compared to EfficientNets.
- **Outperforming Large-Scale Pre-training:** The best DeiT model on ImageNet-1k (85.2% top-1 accuracy) outperforms a ViT-B model pre-trained on the much larger JFT-300M dataset (84.15%), while being significantly faster to train.
- **Transfer Learning:** DeiT models pre-learned on ImageNet are competitive when transferred to various downstream tasks, such as fine-grained classification on CIFAR-10, CIFAR-100, Flowers-102, Stanford Cars, and iNaturalist.
- **Fine-tuning at Higher Resolution:** The standard practice of fine-tuning at a higher resolution (e.g., 384x384) further improves accuracy.

5. Ablation Studies and Training Details:

- The paper includes extensive ablation studies to analyze the impact of different training techniques, data augmentation strategies, and regularization methods. (While not fully detailed in the excerpts provided, the tables and text indicate the importance of these factors).
- The authors provide specific hyper-parameters used for training.

Important Facts and Statistics:

- **ImageNet-1k Dataset Size:** 1,281,167 images, 1000 classes.
- **DeiT Model Sizes:**DeiT-Ti: 5M parameters, ~2536 im/sec throughput (224x224)
- DeiT-S: 22M parameters, ~940 im/sec throughput (224x224)
- DeiT-B: 86M parameters, ~292 im/sec throughput (224x224)

- **DeiT-B Accuracy on ImageNet-1k:**Without distillation: 81.8% (224x224), 83.1% (384x384)
- With distillation (DeiT⚗): 83.4% (224x224), 84.5% (384x384)
- With distillation and extended training (1000 epochs): up to 85.2% (384x384)
- **Training Time:** DeiT-B takes 53 hours on a single node (8 GPUs) for pre-training, and 20 hours for fine-tuning.
- **Distillation Token and Class Token Similarity:** Initially low cosine similarity (~0.06), increasing to high similarity (~0.93) at the last layer.

Conclusion:
The DeiT paper demonstrates a significant step forward in making vision transformers more accessible by enabling competitive performance with significantly less training data and computational resources compared to prior methods. The introduction of the novel distillation token, and the finding that convnets serve as effective teachers, are key contributions. The results suggest that vision transformers are rapidly becoming a viable alternative to convnets for image understanding tasks, even without massive proprietary datasets. The authors provide their code and models, promoting further research and adoption.

What is the primary challenge addressed by the Data-efficient image Transformers (DeiT)?

The paper introducing DeiT addresses the challenge of the high computational cost and large data requirements typically needed to train high-performing vision transformers (ViT). Traditional ViTs, like the one proposed by Dosovitskiy et al., required pre-training on massive datasets with hundreds of millions of images using substantial infrastructure, limiting their accessibility and adoption. DeiT aims to make training competitive convolution-free transformers more accessible by training them on ImageNet only, using significantly less computing power and time (e.g., on a single computer in less than 3 days).

How does DeiT achieve competitive performance without relying on massive external datasets?

DeiT achieves competitive performance on ImageNet using only the standard ImageNet-1k dataset through several key strategies. These include building upon the ViT architecture with specific improvements from libraries like timm and, crucially, employing a novel training scheme adapted for a "data-starving regime." This training incorporates techniques like extensive data augmentation (including Rand-Augment, Mixup, Cutmix, and Repeated Augmentation) and regularization (like stochastic depth). Furthermore, the introduction of a teacher-student distillation strategy, especially when using a convnet as a teacher, plays a significant role in transferring inductive biases and improving performance.

What is the novel distillation strategy introduced in the DeiT paper?

The DeiT paper introduces a novel teacher-student distillation strategy specifically designed for transformers. This strategy relies on a "distillation token" which is added to the initial embeddings alongside the patch tokens and the class token. This distillation token interacts with other tokens

through self-attention layers. At the output of the network, this distillation token's objective is to reproduce the label estimated by a teacher model, unlike the class token which aims to reproduce the true label. This token-based approach allows the student transformer to learn from the teacher's output in a manner complementary to learning from the ground truth labels.

How does the distillation token differ from the class token and why is it beneficial?

While both the class token and the distillation token interact with other tokens through self-attention and are used for classification at the output, their target objectives differ. The class token is trained to predict the true ground truth label, while the distillation token is trained to reproduce the label predicted by a teacher model. The paper observes that these two tokens converge towards different vectors during training, suggesting they capture distinct yet complementary information. The distillation token is particularly beneficial as it allows the transformer student to effectively learn from the teacher's predictions, including soft labels or hard labels, and to potentially inherit useful inductive biases from the teacher, especially when the teacher is a convnet. Using both tokens in a joint classifier at test time provides a significant improvement over using either token independently or traditional distillation methods.

What is "hard-label distillation" and how does it compare to "soft distillation" in the context of DeiT?

Hard-label distillation, as introduced in the DeiT paper, is a variant of knowledge distillation where the student model is trained using the "hard" decision of the teacher as if it were a true label. This means taking the argmax of the teacher's logits to get a single predicted class. This is in contrast to "soft distillation," which minimizes the Kullback-Leibler divergence between the softmax outputs (the "soft" probabilities) of the teacher and the student. The paper found that hard distillation significantly outperforms soft distillation for training transformers, offering better accuracy while being parameter-free and conceptually simpler.

What kind of teacher models are most effective for distilling knowledge into DeiT transformers?

The paper found that using a convnet (convolutional neural network) as a teacher model results in better performance when distilling knowledge into DeiT transformers compared to using another transformer with comparable performance. This is likely due to the convnet teacher transferring its inductive biases, which are beneficial for image understanding tasks, to the transformer student through the distillation process. The authors primarily used a RegNetY-16GF, a state-of-the-art convnet, as their default teacher in distillation experiments.

Can DeiT models be effectively used for transfer learning on other image classification tasks?

Yes, DeiT models pre-learned on ImageNet are competitive when transferred to different downstream tasks such as fine-grained classification on several popular public benchmarks. The paper demonstrates

that DeiT models perform on par with competitive convnet models in transfer learning scenarios, showing their ability to generalize effectively to new datasets and tasks after initial training on ImageNet.

What are some of the key training techniques and hyperparameters used to make DeiT data-efficient?

Several key training techniques and hyperparameters contribute to the data-efficient training of DeiT. These include using the AdamW optimizer with carefully tuned learning rates and a smaller weight decay compared to previous ViT implementations. Extensive data augmentation methods such as Rand-Augment, Mixup, Cutmix, and importantly, Repeated Augmentation, are crucial for improving performance in a data-scarce setting. Regularization techniques like stochastic depth are also employed. The paper provides a detailed ablation study highlighting the impact of each of these choices on the final performance.

DeiT Study Guide

Quiz

1. What is the primary limitation of high-performing vision transformers mentioned in the abstract of "Training data-efficient image transformers..."?
2. What dataset was primarily used to train the DeiT models in this paper?
3. What is the key innovation introduced in the paper to improve distillation for transformers?
4. How does hard-label distillation differ from soft distillation?
5. What are the three types of tokens used in the proposed transformer architecture with distillation?
6. Why is a convnet often a better teacher for a transformer student than another transformer, according to the paper?
7. How does the paper address the challenge of adapting positional embeddings when fine-tuning at a higher image resolution?
8. What is "late fusion" in the context of classification with DeiT?
9. What is the advantage of using the AdamW optimizer with a smaller weight decay compared to the setting used for ViT?
10. What is one training technique that significantly boosts performance in DeiT and is considered a key ingredient?

Quiz Answer Key

1. They require pre-training with hundreds of millions of images using large infrastructure, limiting their adoption.
2. The ImageNet dataset was used as the sole training set.
3. They introduce a teacher-student strategy specific to transformers using a distillation token.

4. Hard-label distillation uses the teacher's single highest probability prediction (argmax) as a "true" label, while soft distillation uses the full probability distribution (softmax output) of the teacher.
5. The architecture uses patch tokens, a class token, and a distillation token.
6. It is suggested that convnets are better teachers because they can transfer their inductive biases to the transformer student through distillation.
7. They interpolate the positional embeddings, often using bicubic interpolation to preserve the norm of the vectors before fine-tuning.
8. Late fusion refers to combining the softmax outputs from separate linear classifiers associated with both the class and distillation embeddings to make the final prediction at test time.
9. A smaller weight decay with AdamW helps improve convergence in their DeiT training setting.
10. Repeated augmentation is identified as a key ingredient that provides a significant boost in performance.

Essay Questions

1. Compare and contrast the Vision Transformer (ViT) and the Data-efficient image Transformer (DeiT) architectures and training methodologies, focusing on how DeiT addresses the data dependency challenge of ViT.
2. Explain in detail the novel distillation strategy proposed in the paper, including the role of the distillation token and how it interacts with other components of the transformer architecture. Discuss the benefits observed when using this method.
3. Analyze the experimental results presented in the paper regarding the efficiency and accuracy of DeiT compared to state-of-the-art convolutional neural networks (convnets) and other vision transformers. What conclusions can be drawn about the trade-off between accuracy and throughput?
4. Discuss the findings on transfer learning with DeiT models to different downstream tasks. How does the performance of DeiT models pre-trained on ImageNet compare to training from scratch on smaller datasets and to other architectures?
5. Describe the various data augmentation and regularization techniques employed in the training of DeiT models. Explain the rationale behind using these techniques and discuss the impact of different methods based on the ablation study results.

Glossary of Key Terms

- **Attention:** A mechanism in neural networks that allows the model to focus on different parts of the input data when processing it. In self-attention, it allows the model to weigh the importance of different elements within the same input sequence.
- **Convolutional Neural Networks (ConvNets):** A class of neural networks widely used for image understanding tasks, characterized by the use of convolutional layers to extract features from images.
- **Transformer:** A neural network architecture originally developed for natural language processing that relies entirely on attention mechanisms, without using recurrent or convolutional layers.

- **Vision Transformer (ViT):** A transformer architecture adapted for image classification, which treats images as sequences of patches.
- **Data-efficient image Transformer (DeiT):** The vision transformer architecture and training methodology proposed in the paper, designed to achieve competitive performance with less training data than traditional ViT models.
- **Distillation:** A training technique where a smaller "student" model is trained to mimic the behavior of a larger, higher-performing "teacher" model, often by learning from the teacher's output probabilities ("soft labels") or predictions ("hard labels").
- **Distillation Token:** A new, learned token added to the transformer input sequence during training, specifically designed to learn from the teacher model's output through attention.
- **Class Token:** A trainable vector appended to the input sequence of a transformer, used to represent the entire input (e.g., an image) for the purpose of classification.
- **Patch Tokens:** Vectors representing individual image patches after being projected into a higher-dimensional space, serving as the sequence input to the transformer.
- **Multi-head Self Attention (MSA):** An extension of the self-attention mechanism that allows the model to jointly attend to information from different representation subspaces at different positions.
- **Feed-Forward Network (FFN):** A standard two-layer neural network with a non-linear activation function (like GeLu) applied independently to each position in the transformer's output.
- **Positional Embeddings:** Learned or fixed vectors added to the input tokens to incorporate information about their position in the sequence, as transformers are otherwise invariant to order.
- **Softmax Function:** A function that converts a vector of numbers into a probability distribution, where the sum of the elements is 1.
- **Kullback-Leibler (KL) Divergence:** A measure of how one probability distribution is different from a second, reference probability distribution. Used as a loss function in soft distillation.
- **Cross-Entropy (LCE):** A measure of the difference between two probability distributions. Used as a loss function for classification tasks, typically comparing the predicted distribution to the ground truth labels.
- **Logits:** The raw, unnormalized outputs from the last layer of a neural network before applying a softmax function.
- **Hard Labels:** The discrete class labels (e.g., the index of the class with the highest probability).
- **Soft Labels:** The probability distribution over all possible classes, typically the output of a softmax function.
- **Repeated Augmentation:** A data augmentation technique that increases the effective batch size by applying multiple different augmentations to the same image within a single batch.
- **Throughput:** A measure of the number of images a model can process per unit of time (e.g., images per second), often used to evaluate the efficiency of a model.
- **Top-1 Accuracy:** A common metric for classification performance, representing the percentage of test samples for which the model's top prediction matches the ground truth label.
- **Fine-tuning:** The process of adapting a pre-trained model to a new dataset or task by continuing training on the new data, often with a lower learning rate.

- **Inductive Bias:** Assumptions made by a learning algorithm to generalize from training data to unseen examples. For example, convolutional layers have an inductive bias towards local spatial relationships.