Briefing Document: Vision Transformer (ViT) for Image Recognition

Source: Excerpts from "an image is worth 16X16 words: transformers for image recognition at scale" by Dosovitskiy et al. (Google Research, Brain Team)
Date: October 22, 2020
Subject: Introduction and evaluation of the Vision Transformer (ViT) architecture for image recognition.
Summary: This paper introduces the Vision Transformer (ViT), a pure Transformer model applied directly to sequences of image patches for image classification. The authors demonstrate that, contrary to the prevailing dominance of convolutional neural networks (CNNs) in computer vision, a Transformer architecture, without significant image-specific inductive biases, can achieve state-of-the-art results on various image recognition benchmarks when pre-trained on large-scale datasets. The key finding is that large-scale training data mitigates the need for strong inductive biases inherent in CNNs, allowing the highly scalable Transformer architecture to excel.
Key Themes and Ideas:

1. **Applying Standard Transformers to Images:** The core concept of the paper is to adapt the Transformer architecture, which has been highly successful in Natural Language Processing (NLP), for image recognition. The approach is remarkably simple: "we split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application."
2. **ViT Architecture:**

- Images are divided into fixed-size patches.
- Each patch is flattened and linearly projected into a lower-dimensional embedding space.
- Position embeddings are added to the patch embeddings to retain spatial information. The authors used standard learnable 1D position embeddings and found them sufficient, with "no significant performance gains from using more advanced 2D-aware position embeddings."
- An extra learnable "classification token" is prepended to the sequence of embedded patches, similar to BERT's [class] token. The state of this token at the output of the Transformer encoder is used as the image representation for classification.
- The resulting sequence of vectors is fed into a standard Transformer encoder, which consists of alternating layers of multiheaded self-attention (MSA) and MLP blocks with Layernorm and residual connections.
- A classification head (MLP during pre-training, single linear layer during fine-tuning) is attached to the output state of the classification token.

3. **Reduced Inductive Bias Compared to CNNs:**

   ViT has significantly fewer image-specific inductive biases than CNNs. While CNNs inherently incorporate "locality, two-dimensional neighborhood structure, and translation equivariance," ViT only utilizes these biases minimally: "in the beginning of the model by cutting the image into patches and at fine-tuning time for adjusting the position embeddings for images of different resolution." The self-attention layers in ViT are global, meaning they can attend to information across the entire image from early layers.

4. **The Crucial Role of Large-Scale Pre-training:**

   A critical finding is that ViT performs poorly when trained on mid-sized datasets like ImageNet without strong regularization. The lack of inductive bias leads to poor generalization on insufficient data. However, "the picture changes if the models are trained on larger datasets (14M-300M images). We find that large scale training trumps inductive bias." When pre-trained on datasets like ImageNet-21k or JFT-300M, ViT "attains excellent results when pre-trained at sufficient scale and transferred to tasks with fewer datapoints."

5. **Competitive or Superior Performance to State-of-the-Art CNNs:**

   When pre-trained on sufficient data, ViT models "approach or beat state of the art on multiple image recognition benchmarks." For example, the best ViT model reached 88.55% accuracy on ImageNet and 94.55% on CIFAR-100. Importantly, ViT models achieved this performance while "requiring substantially fewer computational resources to train" compared to state-of-the-art ResNet-based models like BiT.

6. **Scaling Study Findings:**

   The authors conducted a controlled scaling study and found that Vision Transformers "dominate ResNets on the performance/compute trade-off," using significantly less compute to achieve similar performance. Hybrids (CNN backbone followed by a Transformer) showed slight improvements over pure Transformers at smaller computational budgets, but this advantage disappeared for larger models. Scaling the depth of the Transformer encoder resulted in the biggest performance improvements.

7. **Analysis of ViT's Internal Mechanisms:**

- The learned embedding filters in the first layer resemble "plausible basis functions for a low-dimensional representation of the fine structure within each patch."
- The position embeddings learn to encode spatial relationships, with "closer patches tend to have more similar position embeddings."
- Self-attention allows ViT to integrate global information from early layers. Some attention heads "attend to most of the image already in the lowest layers," while others are more localized. The "attention distance" generally increases with network depth.

8. **Promising Results with Self-Supervision:**

   The authors explored masked patch prediction as a self-supervision task, similar to masked language modeling in BERT. While the initial results showed a significant improvement over training from scratch (79.9% accuracy on ImageNet for ViT-B/16), there was still a notable gap compared to large-scale supervised pre-training. This suggests that further exploration of self-supervised methods for ViT is a key future direction.

9. **Future Directions:**

   The authors identify several challenges and opportunities, including applying ViT to other computer vision tasks (detection and segmentation), further exploring self-supervised pre-training, and continued scaling of ViT models.

Key Facts and Figures:

- **Model Variants:** ViT-Base (86M params), ViT-Large (307M params), ViT-Huge (632M params).
- **Patch Sizes:** Experiments explored different patch sizes, e.g., ViT-L/16 (16x16 patches), ViT-B/32 (32x32 patches). Smaller patch sizes result in longer input sequences and are computationally more expensive.
- **Pre-training Datasets:** ImageNet (1.3M images), ImageNet-21k (14M images), and JFT-300M (303M images). Large-scale pre-training on JFT-300M was crucial for ViT's state-of-the-art performance.
- **Benchmark Performance (Highlights from Table 2):**ViT-H/14 (JFT-300M pre-trained) achieved 88.55% on ImageNet, 94.55% on CIFAR-100, and 77.63% on the VTAB suite.
- ViT-L/16 (JFT-300M pre-trained) achieved 87.76% on ImageNet and 93.90% on CIFAR-100.
- **Computational Efficiency (Highlights from Table 2):** ViT models pre-trained on JFT-300M required "substantially less computational resources to train" than ResNet-based baselines pre-trained on the same dataset (e.g., ViT-L/16: 0.68k TPUv3-core-days vs. BiT-L (ResNet152x4): 9.9k TPUv3-core-days).
- **VTAB Performance (Figure 2 & Table 9):** ViT-H/14 outperformed BiT and other SOTA methods on Natural and Structured tasks in the VTAB benchmark.

Conclusion:
The paper successfully demonstrates that a pure Transformer architecture, when scaled and pre-trained on sufficiently large datasets, can be a competitive and computationally efficient alternative to convolutional networks for image recognition. The Vision Transformer (ViT) represents a significant step in applying general-purpose, highly scalable architectures from NLP to the domain of computer vision, highlighting the increasing importance of data scale over strong domain-specific inductive biases. This work opens up new avenues for research in applying Transformer-based models to a wider range of computer vision tasks.

What is the main idea behind the Vision Transformer (ViT) model?

The core idea of the Vision Transformer (ViT) is to apply a standard Transformer architecture, which has been highly successful in natural language processing (NLP), directly to image classification tasks. Instead of using convolutional neural networks (CNNs) as the primary building block for processing visual information, ViT treats an image as a sequence of fixed-size image patches. These patches are then linearly embedded, positional information is added, and this sequence of patch embeddings is fed into a standard Transformer encoder, similar to how sequences of words (tokens) are processed in NLP models. The reliance on CNNs for initial feature extraction is shown to be unnecessary when the Transformer is trained on large datasets.

How does ViT process images without relying on traditional CNN architectures?

ViT transforms a 2D image into a 1D sequence of vectors suitable for a standard Transformer encoder. This is achieved by first splitting the image into fixed-size, non-overlapping patches. Each patch is then flattened into a 1D vector. These flattened patch vectors are then projected linearly into a lower-dimensional embedding space. To preserve the spatial information lost by flattening, learnable position embeddings are added to the patch embeddings. This resulting sequence of embedded patches, along with an additional learnable "classification token", is then fed into a standard Transformer encoder which consists of layers of multi-headed self-attention and MLP blocks. The state of the classification token at the output of the encoder is used as the image representation for classification.

What are the key differences in inductive biases between ViT and traditional CNNs for image tasks?

Traditional CNNs inherently possess inductive biases that are beneficial for image data, such as locality, two-dimensional neighborhood structure, and translation equivariance. Locality refers to the fact that convolutions process information in small, local regions of the image. The two-dimensional neighborhood structure is encoded by the convolutional filters. Translation equivariance means that shifting the input image results in a corresponding shift in the output feature maps.
In contrast, the Vision Transformer has significantly fewer image-specific inductive biases. While the initial patch extraction step introduces a minimal 2D structure, and MLP layers within the Transformer are local and translationally equivariant, the self-attention layers are global, allowing information to be integrated across the entire image from the very first layers. The learned position embeddings capture spatial relationships between patches, but this information is learned from scratch, rather than being built into the architecture's structure like in CNNs.

When does the Vision Transformer perform particularly well compared to state-of-the-art CNNs?

The Vision Transformer demonstrates excellent performance, matching or exceeding state-of-the-art CNNs, when it is pre-trained on very large datasets (on the order of millions to hundreds of millions of images). On mid-sized datasets like ImageNet, without strong regularization, ViT models underperform compared to similarly sized ResNets. This suggests that the lack of strong built-in inductive biases in ViT makes it require more data to learn visual patterns effectively. However, when scaled to large datasets like ImageNet-21k or JFT-300M, ViT's performance significantly improves, often surpassing CNN baselines and requiring substantially fewer computational resources for pre-training.

How does pre-training data size affect the performance of Vision Transformers compared to ResNets?

Pre-training data size is a critical factor in the performance of Vision Transformers. On smaller datasets (like ImageNet), ViT models, especially larger ones, tend to overfit and perform worse than ResNets of comparable size. This is attributed to the weaker inductive biases in ViT. However, as the pre-training dataset size increases (to ImageNet-21k and particularly JFT-300M), the performance of ViT models improves dramatically. With sufficiently large datasets, the ability of ViT to learn relevant patterns directly from the data outweighs the benefits of the inductive biases in CNNs, leading ViT to outperform ResNets and show less sign of performance saturation.

What is the role of the "classification token" and position embeddings in the ViT architecture?

The "classification token" ([class] embedding) is a learnable embedding prepended to the sequence of embedded image patches. Its state at the output of the Transformer encoder serves as the aggregated image representation for classification. This is a standard technique borrowed from NLP Transformer models like BERT.

Position embeddings are added to the patch embeddings to encode the spatial location of each patch within the original image. Since the Transformer processes the patches as a sequence, the original 2D spatial information is lost. The learnable 1D position embeddings allow the model to understand the spatial relationships between the patches. While other forms like 2D-aware position embeddings or relative positional embeddings were explored, standard learnable 1D position embeddings were found to be sufficient for competitive performance, suggesting that the model effectively learns the 2D image topology from these embeddings on patch-level inputs.

How does ViT handle image resolution differences during fine-tuning?

When fine-tuning ViT on datasets with different image resolutions than the pre-training dataset, the patch size is kept the same. This results in a different effective sequence length for the Transformer. While the Transformer can handle variable sequence lengths, the pre-trained position embeddings might no longer be accurate for the new resolution. To address this, the pre-trained position embeddings are adjusted by performing 2D interpolation based on their location in the original image resolution. This resolution adjustment and the initial patch extraction are the primary points where a manual inductive bias about the 2D structure of images is introduced into the ViT.

What insights can be gained from analyzing the internal workings of the Vision Transformer, such as attention patterns?

Analyzing the internal representations of ViT provides insights into how it processes visual information. The initial linear projection learns filters that resemble basis functions for the fine structure within patches. The learned position embeddings capture spatial relationships, with closer patches having more similar embeddings and revealing the row-column structure. Analyzing the self-attention layers reveals that some attention heads integrate information across the entire image even in the lowest layers, while others focus on local regions, similar to early convolutional layers in CNNs. The "attention distance" (analogous to receptive field size) generally increases with network depth, indicating a progression from local to global information integration. Furthermore, the model appears to attend to semantically relevant image regions for classification. Preliminary work on self-supervised pre-training, mimicking masked language modeling, shows promise, though a gap remains compared to large-scale supervised pre-training.

Vision Transformer (ViT): A Study Guide

Quiz

1. What is the primary difference in architectural approach between Vision Transformer (ViT) and traditional Convolutional Neural Networks (CNNs) for image recognition tasks?
2. How does the Vision Transformer process a 2D image to make it compatible with a standard Transformer encoder?
3. What is the purpose of adding position embeddings to the sequence of embedded image patches in ViT?
4. When is the Vision Transformer's reliance on inductive biases less crucial for achieving high performance?
5. What is a "hybrid architecture" in the context of this paper, and how does it combine CNNs and Vision Transformers?
6. How does the ViT model handle fine-tuning on images with higher resolution than the pre-training resolution?
7. According to the paper, how does the pre-training computational cost of ViT compare to state-of-the-art convolutional networks on similar datasets?
8. What does the analysis of "attention distance" in ViT suggest about how the model integrates information across an image?
9. What self-supervised pre-training method was explored in the paper for Vision Transformer?
10. What are some of the remaining challenges mentioned by the authors for the application of Vision Transformer?

Quiz Answer Key

1. ViT applies a pure Transformer directly to sequences of image patches, treating them like tokens in NLP, whereas traditional CNNs rely heavily on convolutional layers with built-in inductive biases like locality and translation equivariance.
2. The Vision Transformer reshapes a 2D image into a sequence of flattened 2D patches, which are then linearly embedded to a constant latent vector size.
3. Position embeddings are added to the patch embeddings to provide the Transformer encoder with information about the spatial location of each patch within the original image.
4. The reliance on inductive biases is less crucial when ViT is pre-trained on large amounts of data, as large scale training "trumps inductive bias."
5. In a hybrid architecture, the input sequence to the Vision Transformer is formed from feature maps extracted from a Convolutional Neural Network, effectively combining aspects of both architectures.
6. When fine-tuning on higher resolution images, ViT keeps the patch size the same, resulting in a longer sequence. Pre-trained position embeddings are then adjusted using 2D interpolation based on their original image location.

7. The paper suggests that ViT requires substantially fewer computational resources to pre-train compared to state-of-the-art convolutional networks like BiT and Noisy Student while achieving comparable or better performance.
8. The analysis suggests that ViT's self-attention allows it to integrate information globally across the image from the lowest layers, analogous to the receptive field size in CNNs, with some heads having consistently small attention distances.
9. The paper explored masked patch prediction for self-supervised pre-training, mimicking the masked language modeling task used in BERT.
10. Remaining challenges include applying ViT to other computer vision tasks (detection, segmentation), further exploring self-supervised pre-training methods to close the gap with large-scale supervised pre-training, and further scaling of ViT.

Essay Questions

1. Discuss the significance of large-scale pre-training for the Vision Transformer, contrasting its performance on smaller datasets with its performance on larger datasets. Analyze why this difference exists in comparison to CNNs.
2. Compare and contrast the inductive biases present in Convolutional Neural Networks and the Vision Transformer. Explain how the architectural differences lead to these distinct biases and their implications for generalization.
3. Elaborate on the hybrid architecture proposed in the paper. How does it attempt to leverage the strengths of both CNNs and Vision Transformers? Discuss its performance relative to pure ViT and ResNet models at different computational budgets.
4. Analyze the findings from the "Inspecting Vision Transformer" section. Describe what the visualization of embedding filters, position embedding similarity, and attention distance reveal about how ViT processes visual information.
5. Evaluate the experimental results presented in the paper regarding computational cost and performance across different model sizes and pre-training datasets. What conclusions can be drawn about the efficiency and scalability of Vision Transformers compared to their CNN counterparts?

Glossary of Key Terms

- **Transformer Architecture:** A neural network architecture based on self-attention mechanisms, originally developed for natural language processing tasks.
- **Convolutional Neural Networks (CNNs):** A class of neural networks commonly used for image processing, characterized by convolutional layers that exploit spatial hierarchies.
- **Vision Transformer (ViT):** A model that applies a standard Transformer architecture directly to sequences of image patches for image recognition.
- **Image Patches:** Fixed-size subdivisions of an image used as input tokens for the Vision Transformer.
- **Linear Embedding:** A process of projecting flattened image patches into a higher-dimensional space using a trainable linear transformation.
- **Token:** In the context of ViT, an image patch after it has been flattened and linearly embedded, analogous to a word token in NLP.
- **Position Embeddings:** Learnable vectors added to patch embeddings to provide spatial information about the location of each patch.
- **Transformer Encoder:** The main component of the Vision Transformer, consisting of alternating layers of multi-headed self-attention and MLP blocks.

- **Multi-Headed Self-Attention (MSA):** A mechanism that allows the model to weigh the importance of different input tokens when processing a given token, performed in parallel by multiple "heads."
- **MLP Blocks:** Multi-layer perceptron blocks within the Transformer encoder, typically consisting of two layers with a GELU non-linearity.
- **Layernorm (LN):** Layer Normalization, a technique applied before each block in the Transformer encoder to stabilize training.
- **Residual Connections:** Connections that add the input of a block to its output, helping to mitigate vanishing gradients.
- **Classification Token ([class] token):** An extra learnable embedding prepended to the sequence of patch embeddings, whose final state is used as the image representation for classification.
- **Classification Head:** A layer or set of layers attached to the Transformer encoder output (specifically, the classification token's state) that produces the final class predictions.
- **Fine-tuning:** The process of adapting a pre-trained model to a specific downstream task by training it on a smaller task-specific dataset.
- **Pre-training:** The initial training of a model on a large dataset, typically unsupervised or self-supervised, to learn general representations.
- **Inductive Bias:** Assumptions made in the design of a model that influence its learning process and generalization, such as locality and translation equivariance in CNNs.
- **Hybrid Model:** An architecture that combines components of both CNNs and Vision Transformers.
- **Self-supervised Pre-training:** Pre-training where the model learns from the data itself by solving a pretext task, such as masked patch prediction.
- **Few-shot Accuracy:** Performance metric obtained by training a linear classifier on a small subset of training examples using the frozen representations from a pre-trained model.
- **VTAB:** Visual Task Adaptation Benchmark, a suite of 19 diverse classification tasks used to evaluate the transfer learning capabilities of vision models.
- **Attention Distance:** A metric used to analyze how far information is integrated across an image by the self-attention mechanism, analogous to receptive field size.
- **Attention Rollout:** A technique used to visualize attention maps by recursively multiplying the attention weight matrices across all layers.
- **Axial Attention:** A technique for applying self-attention along individual axes of a multi-dimensional tensor, often used to reduce computational cost for large inputs.
- **ImageNet:** A large-scale dataset for visual object recognition.
- **JFT-300M:** An in-house dataset with 303 million high-resolution images used for large-scale pre-training.