Literature Review for Ingur Thesis References

Reza

August 22, 2024

Contents

Chapter 1

BinaryViT [1]

1.1 Introduction

The paper addresses the challenge of improving the performance of binary Vision Transformers (ViTs), a class of deep learning models used in computer vision. While ViTs have shown great potential, particularly when trained on large datasets, they suffer significant performance loss when binarized — a technique that reduces computational costs by converting model weights and activations into binary values. This performance drop is especially notable compared to convolutional neural networks (CNNs), which handle binarization more effectively.

1.2 Proposed Model

The paper identifies that the architecture of standard ViTs lacks key features present in CNNs, which allows CNNs to maintain higher representational capability even after binarization. To address this, the authors propose BinaryViT, a model that incorporates several features inspired by CNNs into the ViT architecture, without using convolutions. These enhancements include:

1.2.1 Global Average Pooling Layer

Replacing the token pooling layer with a global average pooling layer, which helps gather more information from input patches.

Previous Approach: In standard ViTs, a token pooling layer is used before the classifier layer, which only takes into account information from the CLS token rather than considering all tokens in the input sequence.

Proposed Change: The authors replace the token pooling layer with a global average pooling layer. This ensures that the model incorporates information from all input tokens (or patches), not just the CLS token. By doing so,

the final classifier layer has more flexibility and can capture richer feature representations. This addition significantly increases the representational capability of the binary model by aggregating the information from all patches, which is crucial for improving accuracy in binary settings.

1.2.2 Multiple Pooling Branches

Introducing multiple pooling branches in each block to increase representational capability. Adding an affine transformation before each residual connection to balance the scales of different layers.

Previous Approach: Traditional ViTs, and earlier works in binary ViTs, use simple feed-forward layers (FFNs) after attention layers with limited flexibility in processing features.

Proposed Change: Inspired by CNNs, where convolutional layers capture different spatial information, the authors introduce multiple average pooling branches in each block. Each branch has different kernel sizes (e.g., 1x3, 3x1, 1x5, 5x1), allowing the model to process and aggregate spatial information in multiple directions. This change enhances the binary ViT's ability to represent more complex information, without adding significant computational overhead.

1.2.3 Affine Transformation Before Residual Connections

Incorporating a pyramid structure to process high-resolution features early on and reduce them as the model goes deeper, increasing its flexibility and performance.

Previous Approach: In ViTs, the scale of hidden states grows deeper in the network layers, often causing the residual branches to overwhelm the main branches, leading to a decrease in the model's effectiveness. Binary CNNs, such as ResNet, use batch normalization before residual connections, which helps balance the scale of different layers and improves performance.

Proposed Change: To counter the issue of overwhelming residual connections, the authors introduce an affine transformation before each residual addition in the ViT architecture. This technique is inspired by batch normalization in CNNs, which helps maintain a balance between the main and residual branches. The transformation prevents residual connections from dominating the main branches and allows the binary ViT to maintain better feature flow and representation through deeper layers.

1.2.4 Pyramid Structure

Previous Approach: Binary ViTs (like DeiT) typically use a fixed resolution for the feature maps throughout the network, unlike CNNs that progressively downsample the feature maps and increase the number of channels as the network goes deeper. In CNNs, this pyramid structure is important for capturing features at different resolutions and improving representational capacity.

Proposed Change: The authors introduce a pyramid structure in the binary ViT. In this architecture, the feature map size progressively decreases (downsampling) while the hidden dimension (number of channels) increases as the network goes deeper. This mirrors the pyramid structure found in CNNs, allowing the model to capture features at high resolution in the early stages and focus on more abstract, lower-resolution features in the later stages. This significantly improves the model's ability to handle complex visual tasks, especially when binarized.

1.2.5 Binary Fully-Connected Layers with Enhanced Attention

Previous Approach: Standard ViTs rely on attention mechanisms, where matrix multiplications for query, key, and value operations are computationally expensive and prone to significant performance drops when binarized.

Proposed Change: In the proposed BinaryViT model, the authors optimize the binary attention mechanism by modifying how attention probabilities are calculated. They apply scaling factors and rounding techniques to improve the binary attention probability matrix's accuracy, using methods inspired by prior works like ReActNet and Bi-RealNet in binary CNNs. This enhancement ensures that the binary ViT can more effectively process information during self-attention, resulting in better performance.

1.2.6 Distillation from Full-Precision Models

Previous Approach: Previous methods for binary ViTs did not consistently use teacher-student knowledge distillation methods to reduce the performance gap between binary and full-precision models.

Proposed Change: The authors use a full-precision ViT model as a teacher to guide the training of the binary ViT. They distill knowledge by minimizing the soft cross-entropy loss between the binary student model's logits and the full-precision teacher's logits. While distillation techniques were used in some prior works, the authors tailor it specifically to improve binary ViT performance, focusing on logits rather than other components like attention scores or feed-forward outputs, which caused performance degradation in previous experiments.

1.3 Impact of the Changes

These architectural modifications collectively improve the performance of binary ViTs, making them competitive with binary CNNs. The proposed BinaryViT model achieves a significant performance boost on the ImageNet-1k dataset, outperforming earlier binary transformer models. By integrating CNN-inspired architectural features into ViTs, the authors have managed to retain the benefits

of transformer models while reducing the computational cost and maintaining high accuracy in a binarized setting.

These changes provide a more efficient and flexible architecture for tasks requiring high performance on resource-constrained devices such as smartphones and edge devices.

1.4 Results and Improvements

The results and improvements announced in the BinaryViT method, as detailed in the paper, demonstrate significant advancements in the performance of binary Vision Transformers (ViTs) compared to previous approaches. Below is a breakdown of the results and the improvements achieved by this method:

1.4.1 Performance Improvement on ImageNet-1k

The proposed BinaryViT was evaluated on the ImageNet-1k dataset, a standard benchmark for image classification. The model showed significant performance improvements over baseline binary ViTs and previous state-of-the-art (SOTA) binary models. The key results include:

1. **Baseline binary DeiT-S (previous work):** 48.5% top-1 accuracy on ImageNet-1k.

2. **BinaryViT (proposed method):** Achieved **67.7% top-1 accuracy** using the proposed enhancements, representing a large leap of **19.2% improvement** over the baseline binary ViT (DeiT-S).

3. The modified BinaryViT architecture with full-precision patch embedding layers (BinaryViT*) achieved an even higher **70.6% top-1 accuracy**, making it competitive with top binary CNNs like ReActNet.

1.4.2 Efficiency in Terms of Operations and Parameters

BinaryViT not only improves accuracy but also maintains computational efficiency, making it suitable for deployment on edge devices with limited resources. Key findings include:

1. **Operations (OPs):** The proposed BinaryViT model performed fewer operations compared to many SOTA binary models. For example, BinaryViT had 0.79×10^8 operations compared to ReActNet's 1.93×10^8 operations, making it nearly **2.5× more efficient**.

2. **Parameters:** BinaryViT contains around 22.6 million parameters, which is comparable to the baseline binary ViT (DeiT-S) but significantly less than other competitive models such as ReActNet (21.8 million parameters for ResNet-34 backbone and 29.3 million parameters for MobileNet backbone).

3. **FLOPs (Floating-Point Operations):** BinaryViT performed 0.19×10^8 FLOPs, much lower than ReActNet and other competing models, further highlighting its efficiency.

1.4.3 Comparisons with State-of-the-Art (SOTA) Binary Models

The BinaryViT method was directly compared with other leading binary models, and the results are as follows:

1. ReActNet (ResNet-34 backbone): Achieved 67.5% top-1 accuracy on ImageNet-1k with 1.93×10^8 operations and 21.8 million parameters.
2. BinaryViT: Matched ReActNet's accuracy of 67.7%, but with significantly fewer operations (0.79×10^8 vs. 1.93×10^8) and similar parameter count (22.6 million).
3. ReActNet (MobileNet backbone): Achieved 70.1% top-1 accuracy, while BinaryViT* (with full-precision patch embedding layers) closely followed with 70.6% top-1 accuracy and fewer operations, making it highly competitive.

1.4.4 Impact of Architectural Enhancements

The authors tested the impact of each architectural enhancement introduced in BinaryViT. The individual contributions to the model's performance are detailed as follows:

1. **Global Average Pooling:** Replacing token pooling with global average pooling increased the top-1 accuracy from **48.5% to 56.4%**, demonstrating the value of incorporating information from all input tokens.
2. **Multiple Pooling Branches:** Adding multi-branch average pooling layers further improved the accuracy to **60.2%**, showing that this design helps to enrich the representational power of the model.
3. **Affine Transformations:** Introducing affine transformations before residual connections (to balance feature scales) increased accuracy to **61.8%**.
4. **Pyramid Structure:** Implementing the pyramid structure, which mimics CNNs by processing higher-resolution features early on, provided the biggest improvement, bringing accuracy up to **67.7%**.

1.4.5 Reduction in Computational Complexity

One of the key improvements announced by the authors is the ability of BinaryViT to reduce computational complexity without sacrificing performance:

1. **Lower Bit-Operations (BOPs)**: BinaryViT achieved a balance between bit-operations and floating-point operations, outperforming other methods in terms of efficiency.

2. **Efficient Scaling:** The pyramid structure, multi-branch pooling, and affine transformations ensure that the model remains computationally efficient while handling large-scale image datasets like ImageNet-1k.

Comparison Between Full-Precision and Binary Versions

The authors demonstrated that the proposed BinaryViT model maintains performance close to its full-precision counterpart:

1. The full-precision DeiT-S achieves **79.9%** top-1 accuracy.

2. The binary version of BinaryViT achieved **70.6%** accuracy, closing much of the gap between full-precision and binary ViT models.

1.5 Overall Improvements

1. **Significant performance boost:** BinaryViT improves the accuracy of binary ViTs by 19.2% compared to the baseline, making it competitive with binary CNNs.

2. **Reduced operations and parameters:** BinaryViT achieves competitive performance with a lower computational cost, making it ideal for edge devices.

3. **Innovative architecture:** The introduction of CNN-inspired elements such as global average pooling, multiple branches, affine transformations, and pyramid structures enhances the performance of binary ViTs without introducing convolutions.

Chapter 2

Vision Transformer for Small-Size Datasets [2]

This paper focuses on improving the performance of Vision Transformers (ViTs) on small datasets. ViTs, which have shown remarkable success in large-scale datasets, often struggle with small datasets due to their weak locality inductive bias. This bias is critical in image classification tasks as it allows models to focus on local relationships between pixels, which CNNs do well but ViTs lack.

The authors propose two main techniques to address this issue:

2.1 Shifted Patch Tokenization (SPT)

This method aims to improve the tokenization process by spatially shifting image patches in different directions before feeding them into the model. This shift increases the receptive field of each token, allowing the ViT to capture more spatial relationships between neighboring pixels, which enhances the model's ability to understand local features in an image.

2.1.1 Previous Approach:

Traditional Vision Transformers divide an input image into non-overlapping patches and treat each patch as a token, which is then fed into the transformer for processing. This method lacks spatial awareness between adjacent patches because the patches are non-overlapping. In CNNs, the use of convolutional filters ensures that neighboring pixels are processed together, allowing the network to capture local spatial information. However, ViTs, without such mechanisms, have limited capacity to capture local context.

2.1.2 Proposed Change:

The authors introduce Shifted Patch Tokenization (SPT), which enhances the spatial relationship between image patches. The core idea behind SPT is to spatially shift an image in multiple directions (up-left, up-right, down-left, down-right) before dividing it into patches. These shifted versions of the image are then concatenated with the original image and passed through the tokenization process. This results in a larger receptive field for each patch, enabling the model to capture more spatial relationships between neighboring pixels.

1. **Impact:** SPT improves the model's ability to understand local pixel interactions, which is particularly important for smaller datasets where capturing fine details is crucial. By increasing the locality inductive bias, the ViT performs more like a CNN in terms of capturing local information, while still leveraging the benefits of self-attention.

2.2 Locality Self-Attention (LSA)

This technique adjusts the attention mechanism in ViTs to focus more on local regions of an image. LSA uses two strategies: diagonal masking (removing the attention between a token and itself) and learnable temperature scaling (sharpening the attention score distribution). These adjustments prevent the attention from becoming too smooth, forcing it to focus more locally, thus boosting the model's ability to differentiate between important regions in an image.

2.2.1 Previous Approach:

In standard ViTs, the self-attention mechanism evaluates the relationship between all tokens in an image. While this approach is effective for large datasets, it tends to be inefficient for small datasets because it results in a uniform distribution of attention across tokens. This means that ViTs often fail to focus on the most relevant tokens, especially in smaller images where local details matter more. Additionally, the attention scores tend to be smoothed due to the use of high temperatures in the softmax function, making it harder for the model to attend to important local regions.

2.2.2 Proposed Change:

The authors introduce Locality Self-Attention (LSA), which modifies the attention mechanism in two significant ways:

1. **Diagonal Masking:** This method excludes self-tokens from the attention process. In standard attention mechanisms, tokens often pay too much attention to themselves (self-tokens). Diagonal masking forces the model to focus on relationships between different tokens rather than giving undue weight to each token itself.

2. **Learnable Temperature Scaling:** The authors propose adding a learnable temperature parameter to the softmax function, allowing the model to sharpen the attention distribution. A lower temperature sharpens the attention scores, helping the model focus on the most important tokens, particularly in the local regions of an image.

3. **Impact:** These two changes together reduce the tendency of ViTs to spread attention too broadly across the entire image. Instead, the attention becomes more focused on local regions, improving the ability of the model to recognize patterns and details within smaller datasets. LSA makes the attention mechanism more fine-tuned, thus improving performance on small-scale data.

2.3 Comparison to Other Data-Efficient ViTs

The paper compares the proposed SPT and LSA techniques to prior data-efficient ViT models, such as:

1. **DeiT (Data-efficient Image Transformer):** DeiT introduced techniques like knowledge distillation and data augmentations to make ViTs more efficient for training on mid-sized datasets like ImageNet. While effective, it still relies on large datasets and does not specifically address issues with small datasets.

2. **T2T-ViT (Tokens-to-Tokens ViT):** T2T-ViT introduced overlapping patches to improve the spatial relationship between patches. However, it did not fully solve the locality inductive bias issue as it only slightly increased the receptive field of the tokens.

3. **PiT (Pooling-based Vision Transformer):** PiT introduced a hierarchical pooling structure similar to CNNs to generate multi-scale features, allowing for better generalization on smaller datasets. However, it still does not effectively capture fine-grained local spatial information like SPT and LSA.

In contrast, the SPT and LSA techniques specifically address the locality inductive bias in a more targeted way by increasing the receptive field during tokenization (SPT) and making attention more locally focused (LSA). These changes allow the proposed ViT to learn from small datasets effectively without relying on external large-scale pre-training, which was a limitation of previous models.

2.4 Efficiency vs. Performance Trade-offs

2.4.1 Previous Models:

Many of the prior ViT-based models aimed to improve performance but often at the cost of computational efficiency. For example, DeiT used knowledge distillation, and T2T employed a complex overlapping tokenization method, both of which added computational overhead.

2.4.2 Proposed Model:

The proposed BinaryViT maintains competitive performance without a significant increase in computational cost. The SPT technique increases the receptive field without introducing convolutions or pooling layers, and LSA fine-tunes the attention mechanism with minimal additional parameters. As a result, the authors claim that BinaryViT improves accuracy on small datasets while maintaining acceptable overhead in terms of computational complexity.

2.5 Performance Gains

The experimental results in the paper show that the proposed BinaryViT model achieves substantial performance improvements over both the standard ViT and prior data-efficient ViTs when tested on small datasets like CIFAR-100, Tiny-ImageNet, and ImageNet. The model achieves these gains primarily due to its improved ability to capture local spatial information, a limitation that previous models struggled with.

For example:

1. In CIFAR-100, the use of SPT and LSA leads to an accuracy improvement of around 3-4% compared to the baseline ViT model.

2. In Tiny-ImageNet, BinaryViT improves accuracy by up to 4.08%, making it highly competitive with state-of-the-art CNNs on small datasets.

3. Even on a mid-sized dataset like ImageNet, the proposed changes result in a performance boost of 1.06% to 1.60%, demonstrating that the improvements are not limited to only small datasets.

2.6 Overall Impact of the Proposed Changes

The changes proposed by the authors—Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA)—represent significant architectural improvements that specifically address the limitations of Vision Transformers on small datasets. By increasing the locality inductive bias, these techniques make ViTs more efficient and effective at capturing the fine details that are crucial for tasks involving smaller datasets, bridging the gap between CNNs and transformers in this space.

2.7 Results and Improvements

2.7.1 Performance Improvements on Small Datasets

The authors evaluated their methods on various small datasets, such as CIFAR-10, CIFAR-100, Tiny-ImageNet, and SVHN. They compared the performance of standard ViT models with and without the proposed SPT and LSA modules. The key findings are:

1. **CIFAR-100:** The accuracy improved by up to 3.43% for the CaiT model and 4.01% for the PiT model when using SPT and LSA.
2. **Tiny-ImageNet:** ViTs saw significant performance boosts, with up to 4.08% improvement for the Swin Transformer and 4.00% improvement for the baseline ViT.
3. **SVHN and CIFAR-10:** Moderate improvements were observed, with a maximum gain of around 1-2% for some models.

These results highlight that the proposed methods effectively improve ViT performance on small datasets, where the original ViT architectures struggle.

2.7.2 Improvements in ImageNet Performance

While the methods were primarily designed for small datasets, they were also tested on the larger ImageNet dataset to verify if the improvements generalize to mid-sized data. The results show that the proposed methods also enhance ViT performance on ImageNet:

1. **ViT:** Performance increased by 1.60%, achieving a top-1 accuracy of 71.55% (compared to 69.95% for the baseline ViT).
2. **PiT:** Improved by 1.44%, reaching 77.02% accuracy.
3. **Swin Transformer:** Gained 1.06% in accuracy, reaching 81.01%.

These results indicate that SPT and LSA can enhance ViTs even on larger datasets like ImageNet, although their primary benefit is seen in smaller datasets.

2.7.3 Efficiency and Computational Overhead

One of the key advantages of the proposed methods is their minimal computational overhead. Despite the performance improvements, the added complexity from SPT and LSA is modest:

1. **Throughput:** The proposed methods cause only slight reductions in throughput. For example, the addition of SPT and LSA caused a 1.12% latency overhead for the ViT model, and similar small increases for other models.

2. **FLOPs and Parameters:** The increase in FLOPs (Floating Point Operations) and parameters was minimal, ensuring that the models remain efficient and deployable, even with the added improvements in locality inductive bias.

2.7.4 Ablation Study Results

The authors conducted an ablation study to demonstrate the individual contributions of SPT and LSA:

1. SPT (Shifted Patch Tokenization): Improved performance independently by +1.43% in Tiny-ImageNet.
2. LSA (Locality Self-Attention): Provided an independent boost of +3.60% in Tiny-ImageNet.
3. Combining SPT and LSA: When both methods were applied together, the performance improvement reached +4.00% in Tiny-ImageNet, showing a strong synergy between the two methods.

This shows that each technique effectively increases the model's ability to capture local details, and when used together, they yield even greater performance gains.

2.7.5 Qualitative Improvements

In addition to quantitative results, the authors provided qualitative visualizations of the ViT models' attention maps. They compared the attention scores of final class tokens with and without the proposed methods:

1. **Object Shapes:** When SPT and LSA were applied, the attention maps better captured the object shapes, focusing more on the relevant parts of the image, and avoiding excessive attention on background elements.
2. **Sharper Attention:** The learnable temperature scaling in LSA sharpened the attention distribution, leading to more focused and accurate attention on the target objects in images.

These qualitative results visually demonstrate that the proposed changes help the model better understand the structure of the images, especially on smaller datasets where fine-grained details are essential.

2.7.6 Comparison with State-of-the-Art (SOTA) Models

The authors compared their proposed ViT models (with SPT and LSA) against several state-of-the-art (SOTA) models, including CNN-based models like ResNet and EfficientNet. The results showed that:

1. **SL-CaiT:** Achieved better performance than ResNet and EfficientNet on most small datasets (except CIFAR-10).

2. **SL-Swin:** Provided comparable or better performance than CNNs while maintaining higher throughput.

These comparisons highlight the ability of the modified ViTs to close the performance gap with CNNs on small datasets, a space where CNNs have traditionally outperformed transformers.

2.8 Key Takeaways:

1. **Substantial accuracy improvements:** The proposed SPT and LSA methods significantly enhance the performance of ViTs on small datasets, with gains of up to 4.08% on Tiny-ImageNet and 3-4% on CIFAR-100.

2. **Minimal computational overhead:** Despite the improvements, the increase in latency and computational cost is minimal, making these methods practical for deployment.

3. **Generalization to larger datasets:** While primarily aimed at small datasets, SPT and LSA also improve ViT performance on mid-sized datasets like ImageNet, with gains of up to 1.60%.

4. **ViT competitiveness with CNNs:** The proposed methods make ViTs competitive with CNNs in small dataset tasks, both in terms of accuracy and computational efficiency.

In conclusion, the results and improvements from the proposed methods mark a significant advancement for ViTs in handling small datasets, overcoming their limitations in local feature extraction, and making them competitive with traditional CNN architectures.

Chapter 3

How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers [3]

This paper, explores the best ways to train Vision Transformers (ViTs) effectively by balancing the use of data augmentation, regularization, model size, and available computational resources. ViTs are powerful models for computer vision tasks like image classification, but they tend to rely heavily on large datasets and regularization techniques to avoid overfitting. This article aims to provide practical insights for improving ViT performance, especially for practitioners with limited data and computational budgets.

3.1 Data Augmentation and Regularization ("AugReg")

The study shows that using proper data augmentation and regularization can yield models that perform as well as those trained on much larger datasets. By fine-tuning these techniques, smaller datasets can be used effectively, making the training process more efficient.

3.1.1 Previous Works:

Earlier studies on Vision Transformers, such as the original ViT paper, focused heavily on the need for large datasets like ImageNet-21k or JFT-300M to achieve competitive performance. The use of data augmentation and regularization was acknowledged, but the specific impact of these techniques on different dataset

sizes, model configurations, and compute budgets was not systematically explored.

3.1.2 Proposed Changes

The authors of this paper shift the focus to a systematic study of how data **augmentation and regularization** can act as powerful tools to improve the performance of ViTs, even when the available dataset is smaller. The idea is that well-applied augmentation techniques (like Mixup and RandAugment) and regularization methods (such as dropout and stochastic depth) can compensate for the lack of large datasets, mimicking the effects of increasing the dataset size.

This approach differs from previous work by providing empirical evidence showing that with carefully chosen augmentation and regularization settings, models can achieve results comparable to those trained on much larger datasets. This is particularly relevant for practitioners with limited access to massive datasets.

3.2 Trade-offs Between Data, Augmentation, and Compute Budget

The article systematically investigates how the size of training data, the use of augmentation and regularization, and the compute budget interact. It demonstrates that well-designed regularization and augmentation strategies can mimic the effect of significantly increasing the dataset size.

3.2.1 Previous Works:

Many earlier studies on ViTs, such as the DeiT (Data-efficient Image Transformers) work, emphasized the importance of using teacher-student distillation to enhance the performance of ViTs on smaller datasets. While this approach improved results, it added complexity to the training pipeline. Additionally, previous work often considered fixed trade-offs between model size and dataset size, without systematically exploring the effect of compute budget and regularization across a wide range of scenarios.

3.2.2 Proposed Changes:

The authors go beyond distillation-based techniques and offer a more comprehensive investigation into the **interplay between model size, data size, and compute budget**. They conduct experiments across different ViT variants (from small to large models) and different dataset sizes (ImageNet-1k and ImageNet-21k) while systematically adjusting the amount of compute and AugReg techniques.

This approach offers a more nuanced understanding of how to balance **model complexity, data augmentation, and regularization** to achieve optimal performance under various constraints, helping practitioners make better decisions based on their available resources.

3.3 Regularization Techniques and Their Impact

3.3.1 Previous Works:

The role of regularization in ViTs was relatively underexplored in previous work, with most efforts focusing on training larger models on massive datasets. Dropout and stochastic depth were sometimes applied, but their effects were not systematically tested across different model sizes and dataset conditions.

3.3.2 Proposed Changes:

The authors explore the use of **regularization techniques** like dropout and stochastic depth in greater detail. They find that regularization primarily benefits larger models when trained for extended periods, and actually harms performance in smaller models or when training on smaller datasets like ImageNet-21k. They also conduct ablation studies to identify the best settings for regularization, determining that a peak dropout/stochastic depth probability of 0.1 works best.

This more detailed exploration of regularization sets their work apart by offering actionable insights into when and how to apply regularization effectively in ViTs, providing a deeper understanding of its benefits and drawbacks.

3.4 Impact of Model Size

Larger models tend to benefit more from regularization techniques, but this comes with the cost of requiring more training time and computational resources. Smaller models, on the other hand, might not benefit as much from regularization and could even suffer a loss in performance.

3.4.1 Previous Works:

Earlier studies on ViTs often treated model size as a static factor, with larger models generally preferred when training on large datasets. However, there was little guidance on how to adapt model size based on the available compute or dataset size.

3.4.2 Proposed Changes:

The authors provide specific **model size recommendations** based on their findings. They suggest that **larger patch sizes (e.g., 32×32)** are often more

effective than reducing model size (e.g., using "Tiny" ViT models) when compute resources are limited. This means that instead of making the model smaller, increasing the patch size can help maintain performance without increasing the computational load significantly.

This recommendation is based on a systematic analysis of throughput, model size, and patch size, providing a practical guide for selecting the right model configuration depending on the computational constraints.

3.5 Pre-training and Transfer Learning

The study finds that models pre-trained on larger datasets, like ImageNet-21k, perform better across a variety of tasks, including transfer learning. However, practitioners are advised to carefully choose augmentation and regularization settings to match their available compute budget and dataset size.

3.5.1 Previous Works:

Earlier ViT papers focused heavily on large-scale pre-training (e.g., JFT-300M) and emphasized the importance of pre-training on massive datasets for transfer learning. However, this left smaller organizations or researchers without access to such data at a disadvantage.

3.5.2 Proposed Changes:

The authors show that **transfer learning from pre-trained models on smaller datasets (like ImageNet-1k or ImageNet-21k)**, combined with strong AugReg, can yield results similar to those obtained from pre-training on massive datasets. They provide practical recommendations for selecting pre-trained models and fine-tuning them on specific tasks, which is particularly helpful for practitioners without access to large pre-training datasets.

This change democratizes the use of ViTs, making them more accessible to a broader range of users and use cases, and helping users achieve competitive performance without requiring access to extremely large datasets for pre-training.

3.6 Practical Recommendations

The authors offer several practical guidelines, such as preferring data augmentation over regularization for smaller datasets, and choosing larger models with proper augmentation for the best results in a transfer learning setup.

3.7 Overall Impact of Changes

These changes collectively offer a more flexible and practical approach to training Vision Transformers, making them more applicable to real-world scenarios

with constrained resources. The authors provide a comprehensive guide on how to balance data augmentation, regularization, and compute budget, allowing practitioners to achieve top-tier performance without relying on enormous datasets or computational resources. This is a significant shift from earlier ViT models, which focused primarily on large-scale data and heavy compute environments.

Chapter 4

Training data-efficient image transformers & distillation through attention [4]

This paper focuses on making Vision Transformers (ViTs) more accessible and efficient by reducing their reliance on large datasets and expensive computing resources. Vision Transformers are highly effective for image classification, but their performance typically depends on massive datasets and extended training times on large infrastructure. This paper aims to train transformers effectively using only the ImageNet dataset on standard hardware, making these models more usable for a broader audience.

4.1 Data-Efficient Image Transformers (DeiT)

The authors propose a new training method for Vision Transformers, called DeiT, which allows these models to achieve high performance using only the ImageNet-1k dataset, without the need for the massive datasets like JFT-300M used in earlier ViT models. They demonstrate that it is possible to train ViTs efficiently on a single computer with 4 to 8 GPUs within just a few days, making them competitive with convolutional neural networks (CNNs) on standard benchmarks.

4.1.1 Previous Works:

The original ViT model by Dosovitskiy et al. (2020) showed exceptional performance, but it required training on extremely large datasets like JFT-300M, which contains 300 million labeled images. This makes the model less accessible

for researchers and organizations with limited computational resources. The training of ViT models also demanded significant infrastructure, often involving many GPUs over long periods.

4.1.2 Proposed Changes:

The authors propose the DeiT (Data-efficient Image Transformer), a novel approach for training ViTs more efficiently. Key changes in the training strategy include:

1. **Training on ImageNet-1k only:** The authors demonstrate that it is possible to train ViTs using only the 1.28 million images in the ImageNet-1k dataset, instead of relying on large-scale private datasets.

2. **Shorter Training Time:** The authors successfully trained their models on a single machine with 4 to 8 GPUs in less than three days, significantly reducing the computational cost and making it accessible to a wider range of users.

3. **Use of Repeated Augmentation:** Repeated augmentation was introduced to provide more data variations during training, allowing the model to generalize better on smaller datasets. This is crucial for data efficiency in training without relying on external datasets.

By optimizing the training setup, the DeiT model becomes more practical for real-world use cases, where large-scale datasets and high-end computing resources are unavailable.

4.2 Distillation Through Attention

A unique contribution of this paper is the introduction of a distillation token, a new method that enables the student model (transformer) to learn from a teacher model (either a CNN or another transformer). This token is added to the transformer's input and interacts with the class token during training. The distillation token improves the model's performance by helping it mimic the predictions of the teacher, allowing the transformer to learn more efficiently from the teacher's inductive biases, especially when the teacher is a CNN.

4.2.1 Previous Works:

Knowledge distillation is a common technique used in CNN models, where a smaller "student" model learns from a larger "teacher" model, typically by mimicking the teacher's output (soft labels). Previous works used this approach for model compression and transfer learning, but it was not specifically adapted to the transformer architecture.

4.2.2 Proposed Changes:

The authors introduce a novel distillation mechanism specifically designed for transformers. The key innovation is the distillation token, which operates alongside the class token. Unlike the typical class token that learns from the ground truth labels, the distillation token learns from the teacher model's predictions, allowing the student transformer to benefit from the teacher's inductive biases.

1. **Interaction Through Attention:** The distillation token interacts with the other tokens through the self-attention mechanism, ensuring that the student model receives rich, token-level information from the teacher.

2. **Inductive Bias Transfer:** Interestingly, the paper shows that a CNN teacher (such as RegNet) transfers its inductive biases (such as convolutional feature learning) to the transformer through this attention-based distillation process, making the transformer perform better on image recognition tasks.

This approach to distillation is a significant departure from earlier methods because it leverages the attention mechanism to integrate the teacher's guidance into the student's learning process, improving the model's performance.

4.3 Smaller and More Efficient Models (DeiT-S and DeiT-Ti)

4.3.1 Previous Works:

The original ViT model introduced a large transformer architecture (ViT-B) that required massive datasets and significant computational resources to achieve competitive results. Smaller versions of ViTs had not been thoroughly explored, and the effectiveness of scaling down transformers while maintaining performance was not well established.

4.3.2 Proposed Changes:

The authors introduce two smaller versions of the ViT model in their DeiT approach:

1. **DeiT-S (Small):** A smaller version with fewer parameters and heads, designed to be more efficient while still maintaining competitive performance.

2. **DeiT-Ti (Tiny):** An even smaller model, comparable to ResNet-18 in terms of parameter count, which is designed to be lightweight and fast while achieving solid accuracy on image classification tasks.

These smaller models are trained using the same data-efficient strategies and distillation techniques as the larger DeiT-B model, making them highly competitive with traditional CNNs like ResNet. The development of these smaller models is a step forward in making transformers accessible for deployment on edge devices and environments with limited computational power.

4.4 Performance and Efficiency Gains

The proposed DeiT models achieve competitive results with state-of-the-art CNNs, reaching up to 85.2% top-1 accuracy on ImageNet while requiring less compute time and fewer resources. The authors show that, with their distillation method, Vision Transformers can even outperform their teacher models in some cases, demonstrating the effectiveness of this distillation approach.

4.4.1 Previous Works:

Earlier Vision Transformer models (such as ViT-B and ViT-L) were known for their large computational overhead. Their training and inference speed were slower than state-of-the-art CNN models like EfficientNet and ResNet, limiting their usability in real-time or resource-constrained environments.

4.4.2 Proposed Changes:

The authors focus on improving the **throughput** (images processed per second) of DeiT models, ensuring that they are efficient in terms of both **accuracy** and **speed**. They show that:

1. **DeiT-S** achieves higher throughput than many large CNNs like EfficientNet-B4 while maintaining competitive accuracy.
2. **The DeiT-B** model surpasses earlier ViTs trained on ImageNet-1k in terms of throughput, closing the gap between transformers and CNNs.
3. Through their distillation process, the authors achieve **better accuracy-to-throughput ratios** than standard CNNs, meaning that transformers can now offer high performance without compromising speed.

4.5 Transfer Learning and Generalization

The authors also explore the transfer learning capabilities of the DeiT models, showing that they perform well when fine-tuned on other popular datasets like CIFAR-10, CIFAR-100, and iNaturalist, further proving their generalization power.

4.5.1 Previous Works:

Vision Transformers, especially in their original form, struggled with transfer learning performance when fine-tuned on smaller datasets, often requiring extensive fine-tuning or retraining with external datasets. This was a significant limitation compared to CNNs, which have long been effective at transferring learned features across tasks and domains.

4.5.2 Proposed Changes:

The DeiT models are shown to **generalize well** to downstream tasks through transfer learning, performing competitively on popular datasets such as CIFAR-10, CIFAR-100, and iNaturalist. Key improvements include:

1. **Fine-tuning at higher resolutions:** The authors successfully fine-tune DeiT models at different image resolutions, demonstrating that they can achieve high accuracy even when transferred to tasks that require different input sizes.

2. **Cross-dataset generalization:** DeiT models achieve competitive performance across various fine-grained classification tasks, proving their versatility beyond ImageNet. The model can be fine-tuned efficiently on smaller datasets, making it a more general-purpose model.

This shows that DeiT not only excels at training from scratch but also performs strongly in transfer learning scenarios, making it more suitable for a broader range of tasks compared to previous ViT models.

4.6 Results and Improvements

4.6.1 Competitive Performance with Smaller Datasets

The primary contribution of this work is that it shows Vision Transformers (ViTs) can achieve competitive results using only the ImageNet-1k dataset. Previous ViTs required massive datasets like JFT-300M for training. The authors achieve the following:

1. **DeiT-B** (Base model) achieves **83.1% top-1 accuracy** on ImageNet-1k, with no external data.

2. When fine-tuned at higher resolution (384×384), **DeiT-B** achieves **85.2% top-1 accuracy**, surpassing many state-of-the-art CNN models like ResNet and EfficientNet.

These results highlight that DeiT models can perform on par with models that rely on much larger datasets, making ViTs more accessible to the broader machine learning community.

4.6.2 Distillation Through Attention Enhances Model Performance

The introduction of the **distillation token** is a novel improvement in this work. This token allows the transformer model to learn from a teacher (often a CNN) in a more effective manner than traditional distillation methods. The benefits of this approach include:

1. **DeiT-B with distillation (DeiT-B)** achieves **84.4% top-1 accuracy** on ImageNet-1k, which is higher than the original ViT-B model (trained on larger datasets) and better than CNN models of similar size.
2. The distillation token outperforms standard soft and hard distillation techniques, showing that it is particularly effective in helping the transformer model adopt the inductive biases from a CNN teacher.

This result demonstrates that transformers can benefit from distilling knowledge from CNNs, making the training process more efficient and effective.

4.6.3 Improved Throughput and Computational Efficiency

Another key improvement is the **throughput and computational efficiency** of the DeiT models. Compared to earlier Vision Transformers, the DeiT models:

1. Achieve faster **throughput while maintaining high accuracy**. For example, DeiT-Tiny (DeiT-Ti), the smallest model, processes 2536 images per second, making it one of the fastest ViTs available.
2. **Outperform ViTs trained on larger datasets** in terms of the balance between accuracy and throughput. The DeiT models are competitive with CNNs like EfficientNet and ResNet in terms of both speed and accuracy.

The improved throughput and reduced computational requirements make DeiT models more suitable for real-world applications, especially in environments with limited computational resources.

4.6.4 Smaller Models with Comparable Accuracy

The paper introduces smaller and more efficient variants of the DeiT model:

1. **DeiT-S (Small model)** achieves **79.8% top-1 accuracy** on ImageNet-1k while having fewer parameters than ResNet-50, making it a strong alternative to CNNs for image classification tasks.
2. **DeiT-Ti (Tiny model)**, a smaller version comparable to ResNet-18, achieves **72.2% top-1 accuracy** on ImageNet-1k, showing that even small ViT models can be effective when trained efficiently.

These smaller models provide flexibility in choosing the right trade-off between model size, speed, and accuracy, depending on the task at hand.

4.6.5 Transfer Learning and Generalization

The authors demonstrate that DeiT models generalize well to downstream tasks, achieving competitive performance in transfer learning scenarios. Key results include:

1. **DeiT-B achieves 99.1% accuracy on CIFAR-10, 91.4% on CIFAR-100**, and **93.9% on Stanford Cars**, which are common benchmarks for transfer learning tasks.

2. The models also perform well on fine-grained classification tasks and datasets like iNaturalist, Oxford-102 Flowers, and Stanford Cars.

These results show that the DeiT models are not only effective for large-scale classification tasks like ImageNet but also excel in transfer learning, making them versatile for different types of datasets and tasks.

4.6.6 Training Time Reduction

One of the critical improvements in this method is the reduced training time:

1. **DeiT-B** is trained on **a single machine with 4 to 8 GPUs** in just three days, compared to the weeks or even months of training required for other ViT models on large datasets.

2. The fine-tuning of DeiT models at higher resolutions is also efficient, taking only around 20 hours on an 8-GPU machine for 25 epochs.

This reduction in training time, combined with efficient throughput, makes DeiT models more practical and accessible for organizations with limited resources.

4.6.7 Distillation from CNNs Is More Effective than from Transformers

A surprising finding in the paper is that **transformers learn better when the teacher is a CNN rather than another transformer**. The authors observed that:

1. When using a **RegNetY CNN** as the teacher, the student DeiT model outperforms models distilled from transformer teachers.

2. The distillation process allows the transformer to learn important inductive biases from the CNN, improving its ability to capture spatial and hierarchical information in images.

This result suggests that CNNs can still play a valuable role in improving transformer models through distillation, leveraging the strengths of both architectures.

4.7 Overall Improvements

1. **Data Efficiency:** The DeiT models reduce the need for large datasets, showing that ViTs can perform well with ImageNet-1k only.

2. **Distillation Innovation:** The novel distillation token mechanism leads to significant performance gains, especially when CNNs are used as teachers.

3. **Improved Throughput:** DeiT models achieve a better balance between accuracy and throughput, making them faster and more efficient for real-time applications.

4. **Smaller, Faster Models:** The introduction of DeiT-S and DeiT-Ti offers more flexibility with smaller models that perform competitively.

5. **Transfer Learning Success:** DeiT models generalize well across various downstream tasks, proving to be versatile for different datasets.

Chapter 5

Going deeper with Image Transformers [5]

This paper explores how to enhance the performance of Vision Transformers (ViTs) by making them deeper and more efficient for image classification tasks. The authors focus on improving the architecture of ViTs to address their limitations in handling complex visual tasks, especially when compared to Convolutional Neural Networks (CNNs).

5.1 Key Ideas:

5.1.1 Deeper Vision Transformers (ViTs):

The paper investigates the performance of ViTs when scaling them to deeper layers. While transformers have excelled in natural language processing, the same depth and complexity have not been fully explored in vision tasks. The authors experiment with significantly deeper ViTs to better understand how increasing depth affects the model's ability to capture detailed visual information.

5.1.2 Class-Attention Mechanism:

One of the major contributions of this work is the introduction of **class-attention layers**. These layers are added to the architecture to improve how the model attends to important image features. The class-attention mechanism allows the model to focus more on class-relevant areas of the image, leading to better performance, especially when dealing with complex, high-resolution images.

5.1.3 Distillation with Class-Attention:

The authors explore **knowledge distillation**—a technique where a smaller or less powerful "student" model learns from a larger, pre-trained "teacher"

model. While previous works used a distillation token in Vision Transformers to enhance performance, this paper primarily uses hard-label distillation, which averages the teacher's prediction with the true label. The class-attention layers, however, did not benefit from the distillation token as much as expected, so the authors opted for **hard-label distillation**, which provided better results.

5.1.4 Efficient Training and Generalization:

The paper also focuses on how these deeper ViTs, equipped with class-attention, can be trained efficiently and still generalize well to new tasks. By fine-tuning their model architecture, they show that transformers can handle large-scale vision tasks effectively, even with deeper networks. They also achieve faster convergence through the use of hard-label distillation.

5.1.5 Performance on Benchmarks:

The authors demonstrate that their deeper ViTs with class-attention perform exceptionally well on various image classification benchmarks, showing improvements over standard ViTs and CNNs. Their models achieve higher accuracy while maintaining computational efficiency, making them more practical for real-world applications.

5.2 Results and Improvements

5.2.1 Performance Improvement with Depth

A major finding of the paper is that scaling ViTs to deeper architectures improves performance on image classification tasks. By extending the depth of Vision Transformers, the model is able to capture more complex visual patterns and hierarchical features, leading to better results. The authors demonstrate that:

1. Deeper ViTs with **up to 100+** layers outperform shallower models, especially when trained on large datasets.
2. **Deeper transformers** are shown to be highly competitive with state-of-the-art CNN architectures while retaining the flexibility and scalability that ViTs offer.

This improvement highlights the ability of ViTs to handle complex visual tasks effectively when scaled to deeper architectures.

5.2.2 Introduction of Class-Attention Layers

The authors propose **class-attention layers**, which are a significant architectural improvement. These layers are designed to:

1. Focus more on class-relevant areas of an image, helping the model attend to important features while ignoring irrelevant information.

2. Improve the model's ability to distinguish between fine-grained classes, making it more effective for high-resolution and complex images.

The introduction of class-attention layers results in **improved accuracy** in image classification benchmarks compared to standard ViT models. By better aligning the model's attention with class labels, the classification performance is enhanced, particularly for tasks that require precise attention to detail.

5.2.3 Hard-Label Distillation for Faster Convergence

In contrast to previous work that relied on distillation tokens, the paper shows that **hard-label distillation**—a method where the model's predictions are averaged with the teacher's output—provides better performance, particularly when used with class-attention layers. The authors found that:

1. **Hard-label distillation** led to **faster convergence** during training, reducing the overall training time required to achieve high accuracy.

2. The method is more effective for deeper transformers, as it simplifies the distillation process without sacrificing accuracy.

This improvement demonstrates that hard-label distillation can help models learn more efficiently from pre-trained teachers, particularly when deeper architectures are used.

5.2.4 Training Efficiency and Generalization

The deeper ViTs proposed in this paper are not only more accurate but also **more efficient in training**. The authors optimized the training process to handle the deeper architectures, ensuring that the model can still converge quickly and generalize well to new tasks. Key results include:

1. **Faster convergence** during training compared to traditional ViTs, thanks to the combined use of class-attention layers and hard-label distillation.

2. The deeper models generalize well to different datasets and achieve strong performance across various benchmarks without overfitting, demonstrating the robustness of the proposed architecture.

5.2.5 Benchmark Results and Competitive Performance

The authors tested their proposed deeper ViTs on several image classification benchmarks, showing **improved performance over baseline models:**

1. The model achieves higher accuracy compared to standard Vision Transformers, ResNet-based CNNs, and previous state-of-the-art models.

2. When combined with class-attention and hard-label distillation, the deeper ViTs achieve **competitive results on benchmarks like ImageNet** and other challenging datasets.

These results position deeper ViTs with class-attention as a strong alternative to CNNs, making them viable for large-scale and complex visual tasks.

5.2.6 Improved Attention Mechanism for Class Prediction

The introduction of the class-attention mechanism ensures that the **model pays more attention to class-relevant features**, reducing the chance of focusing on background noise or irrelevant parts of an image. This improves the accuracy of class predictions, especially in complex datasets where fine-grained differences between classes are crucial.

1. The **class-attention mechanism** ensures better feature extraction and significantly enhances the model's performance in distinguishing between closely related classes.

Chapter 6

Attention is All you need [6]

This paper introduces the Transformer model, a new architecture for sequence transduction tasks like machine translation, which relies entirely on self-attention mechanisms instead of the traditional recurrent or convolutional layers commonly used in previous models. This approach aims to address the limitations of recurrent neural networks (RNNs), particularly their inefficiency in parallelization and the challenges they face in learning long-range dependencies within sequences.

6.1 Transformer Architecture:

The Transformer is designed with both an encoder and decoder, which are stacked in layers. Each layer uses a self-attention mechanism that allows the model to focus on different parts of the input sequence, regardless of their distance. This is different from RNNs and Convolutional Neural Networks (CNNs), where longer sequences pose challenges for learning relationships between distant elements.

6.2 Self-Attention and Multi-Head Attention:

At the core of the Transformer is the **self-attention (SA)** mechanism, where each word in a sequence attends to every other word, learning the relationships between them. The model uses **multi-head attention (MHA)**, which runs several attention mechanisms in parallel, allowing the model to jointly focus on different parts of the sequence in multiple ways. This enables the model to capture richer contextual information and improve performance on tasks like language translation.

6.3 Positional Encoding:

Since the model is not based on sequences or recurrence, it lacks an inherent sense of order. The authors introduce **positional encodings** to provide the model with information about the position of each word in the input sequence. This helps the Transformer distinguish between different words' positions within a sentence and ensures that word order is maintained.

6.4 Advantages of Transformers:

The Transformer architecture is highly **parallelizable** and can be trained much faster than RNN-based models, which process one word at a time. By using self-attention, the model reduces the computational complexity and makes it easier to train on long sequences. This leads to faster training times, greater scalability, and the ability to achieve high performance with significantly fewer resources compared to other models.

6.5 Results:

The Transformer achieves state-of-the-art results on machine translation tasks, including **English-to-German** and **English-to-French** translations. It significantly outperforms previous models like RNNs and CNNs in terms of both accuracy (measured by BLEU score) and efficiency (measured by training speed). The model can be trained in just a few days on standard hardware while delivering superior performance.

6.6 Summary

In essence, the Transformer model redefines how sequence modeling is approached by eliminating recurrence and relying entirely on attention mechanisms. This new architecture is not only more efficient and parallelizable but also capable of learning long-range dependencies better than previous models. The paper's results demonstrate the Transformer's superiority in machine translation tasks, making it a breakthrough in the field of natural language processing.

Chapter 7

Deepfake Video Detection Using Convolutional Vision Transformer [7]

This paper addresses the growing concerns around deepfakes—hyper-realistic videos created through deep learning techniques that can manipulate or replace faces in videos. While deepfakes have useful applications in fields like entertainment, education, and virtual reality, they also pose serious risks, such as being used for identity theft, misinformation, or fraud.

The authors propose a new model for detecting deepfakes, called the **Convolutional Vision Transformer (CViT)**. This model combines the strengths of two powerful architectures: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). CNNs are effective at learning local features (such as textures or small patterns), while ViTs excel at capturing global features and understanding the relationship between different parts of an image. By combining both models, CViT is able to detect subtle visual differences between real and fake videos.

7.1 Key Components of the Model

7.1.1 Feature Learning through CNNs:

Feature Learning through CNNs: The CNN module extracts local features from video frames, such as facial details and textures, which are crucial for detecting visual artifacts commonly found in deepfakes.

7.1.2 Global Feature Understanding through ViTs:

The Vision Transformer component processes the features extracted by the CNN, learning the relationships between different parts of the image. This

helps the model detect manipulations at both a local and global level.

7.1.3 Comprehensive Data Preprocessing:

The authors emphasize the importance of data preprocessing, which ensures that the input data is well-prepared for training. This step helps improve the model's accuracy and robustness.

7.1.4 Testing and Results:

The CViT model was trained and tested on the DeepFake Detection Challenge (DFDC) dataset, one of the largest and most diverse datasets for detecting deepfakes. The model achieved an accuracy of 91.5%, demonstrating strong performance in identifying fake videos. However, the authors acknowledge that there is room for improvement and plan to expand their research by using more diverse datasets in the future.

7.2 Results and Improvements

This paper presents several key results and improvements from the proposed Convolutional Vision Transformer (CViT) model, which combines CNNs and Vision Transformers for detecting deepfake videos. Below is a detailed overview of the results and improvements:

7.2.1 High Accuracy in Deepfake Detection

The CViT model achieved 91.5% accuracy on the DeepFake Detection Challenge (DFDC) dataset, demonstrating its effectiveness in distinguishing between real and fake videos. This high accuracy reflects the model's ability to detect subtle artifacts and inconsistencies in manipulated videos, which are often difficult for simpler models to detect.

7.2.2 AUC and Loss Metrics

Along with the high accuracy, the model also achieved an AUC (Area Under the Curve) value of 0.91 and a loss value of 0.32. These metrics further validate the robustness of the model in detecting deepfakes. The AUC score indicates that the model performs well in distinguishing true positives (real videos) from false positives (deepfakes), and the low loss value shows that the predictions are close to the actual values.

7.2.3 Combination of CNN and ViT for Local and Global Feature Learning

One of the main improvements of the CViT model is its ability to combine the strengths of CNNs and Vision Transformers (ViTs). CNNs are highly effective

at extracting local features such as textures and small details, while ViTs are powerful in understanding global relationships across an image. The fusion of these two methods allows CViT to detect deepfakes more effectively by capturing both fine-grained details and broader spatial relationships in video frames.

This combined approach is a notable improvement over models that rely solely on CNNs or Transformers, as it provides a more comprehensive analysis of the video content.

7.2.4 Generalized Model for Different Deepfake Scenarios

The authors describe their model as "generalized" for several reasons:

1. **Local and Global Feature Learning:** The combination of CNN and ViT enables the model to learn from both small details and larger spatial contexts, making it more versatile in detecting various types of deepfakes.

2. **Thorough Data Preprocessing:** The emphasis on proper data preprocessing ensures that the input data is of high quality, leading to better detection performance.

3. **Diverse Dataset Training:** The CViT model was trained on the DFDC dataset, which includes a wide variety of videos created in different environments, lighting conditions, and orientations. This makes the model more adaptable to real-world deepfake scenarios.

7.2.5 Comparison with Other Models

The paper compares the performance of the CViT model with other existing deepfake detection models and shows that:

1. The CViT model performs better than simpler CNN-based models, such as those using RNNs or shallow CNNs.

2. The model demonstrates a higher accuracy on most subsets of the DFDC dataset, such as FaceForensics++, where it achieves high detection rates on multiple types of manipulations (Deepfake, FaceSwap, and others).

However, the authors note that the model performed less effectively on certain datasets, such as FaceForensics++ FaceShifter, indicating room for further improvement.

7.2.6 Data Preprocessing and Face Extraction

The authors emphasize the importance of data preprocessing, especially in the extraction of facial images from video frames. By carefully pre-processing the data, including extracting faces in a standardized format (224x224 pixels), the model is able to focus on the relevant parts of the video, improving detection accuracy.

The use of tools like BlazeFace and MTCNN for face detection and alignment ensures that the model works with high-quality input data. This careful preprocessing approach is a key factor in the model's strong performance, setting it apart from other deepfake detection models that may not prioritize preprocessing to the same extent.

7.2.7 Future Improvements and Expansion

While the model shows strong results, the authors acknowledge that there is still room for improvement. They plan to enhance the model by:

1. Adding more diverse datasets for training, which will help the model detect deepfakes in even more varied scenarios.
2. Improving artifact detection in more challenging deepfake videos, such as those created using advanced techniques like FaceShifter, where the model currently struggles.

7.3 Summary of Improvements:

1. High accuracy (91.5%) and AUC score (0.91) on the DFDC dataset, showing effective deepfake detection.
2. CNN and ViT combination enables the model to capture both local and global features, improving the ability to detect deepfakes.
3. Generalized model design allows the CViT to perform well across different deepfake scenarios, including varying video qualities and environments.
4. Comprehensive data preprocessing ensures the input quality is high, leading to better model performance.
5. Future improvements aim to address current limitations, such as improving detection on more complex deepfake techniques and expanding the dataset.

Bibliography

[1] P.-H. C. Le and X. Li, "Binaryvit: pushing binary vision transformers towards convolutional models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4664–4673.

[2] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," *CoRR*, vol. abs/2112.13492, 2021. [Online]. Available: https://arxiv.org/abs/2112.13492

[3] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *CoRR*, vol. abs/2106.10270, 2021. [Online]. Available: https://arxiv.org/abs/2106.10270

[4] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *CoRR*, vol. abs/2012.12877, 2020. [Online]. Available: https://arxiv.org/abs/2012.12877

[5] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," *CoRR*, vol. abs/2103.17239, 2021. [Online]. Available: https://arxiv.org/abs/2103.17239

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[7] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *CoRR*, vol. abs/2102.11126, 2021. [Online]. Available: https://arxiv.org/abs/2102.11126