Literature Review for Ingur Thesis References

Reza

August 25, 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.

Chapter 8

Visual Transformer Pruning [8]

This paper focuses on making Vision Transformers (ViTs) more efficient by reducing their size and computational demands without significantly sacrificing performance. Vision Transformers are known for their high accuracy in computer vision tasks like image classification, but they are resource-intensive, requiring substantial memory, computation, and storage. This makes them difficult to deploy on devices with limited resources, such as mobile phones and embedded systems.

The authors propose a method to prune Vision Transformers, which means removing less important parts of the model to reduce the number of parameters and computational operations. This is done through L1 regularization, a technique that encourages sparsity in the model's dimensions. During training, this method automatically identifies which parts of the model (such as individual dimensions in the layers) are less important and can be removed. By applying this pruning technique, the authors manage to create smaller, faster, and more efficient Vision Transformers while maintaining competitive accuracy.

8.1 Key Components of the Approach

1. L1 Regularization for Sparsity: The method applies L1 regularization to the transformer's dimensions, which forces the model to focus on the most important components while ignoring others. This results in a model with fewer, but more critical, dimensions.

2. Pruning Process: After training with L1 regularization, the dimensions with low importance are pruned. This means they are effectively removed from the model, reducing its size and computational complexity.

3. Fine-tuning After Pruning: Once the model is pruned, it is fine-tuned to

ensure that the performance drop is minimized. This final step helps the model regain any accuracy lost during pruning.

4. Experiments and Results: The authors test their pruning method on two large datasets—ImageNet-100 and ImageNet-1K. The results show that the pruned models significantly reduce the number of parameters and computations, while maintaining similar accuracy to the original, unpruned models. For example, pruning 40% of the model's dimensions reduces computational costs by 43% while only dropping accuracy by about 1

8.2 Results and Improvements

8.2.1 Significant Reduction in Parameters and Computation Costs

The proposed pruning method successfully reduces both the number of parameters and the computational costs (measured in FLOPs) of Vision Transformers. For example:

1. 40% pruning of the model's dimensions resulted in a 45.3% reduction in parameters and a 43.0% reduction in FLOPs (computational cost).

2. 20% pruning reduced the parameters by 23.5% and FLOPs by 22.0%, showing the method's ability to effectively compress the model.

This demonstrates that the pruning approach drastically cuts the resource requirements of Vision Transformers, making them more suitable for deployment on devices with limited resources.

8.2.2 Maintaining High Accuracy with Minimal Loss

Despite reducing the model's size and complexity, the accuracy of the pruned models remained high, showing only minor drops in performance. Specifically:

1. After pruning 40% of the model's dimensions, the accuracy on ImageNet-1K dropped by only 1.1%, from 81.8% to 80.7%.

2. Pruning 20% of the dimensions resulted in an accuracy drop of just 0.5%, maintaining a high level of performance with significantly fewer parameters and lower computational cost.

This result demonstrates the strength of the pruning method, as it allows for a substantial reduction in resources while keeping the model highly accurate.

8.2.3 Effectiveness on Large Datasets

The pruning method was tested on both ImageNet-100 (a subset of ImageNet-1K) and ImageNet-1K, two large-scale image classification datasets. The results were consistent across both datasets:

1. On ImageNet-100, the pruned models achieved similar reductions in parameters and FLOPs with minimal accuracy loss. For instance, the 40% pruned model maintained an accuracy of 92.58% on ImageNet-100.
2. On the larger ImageNet-1K dataset, the pruned model remained highly competitive with only a 1.1% accuracy loss despite the significant reduction in computational complexity.

This demonstrates that the method generalizes well to large datasets, making it viable for use in real-world applications involving complex data.

8.2.4 Flexibility of Pruning Rates

The authors conducted ablation studies to test different pruning rates, allowing for flexible reductions in model size. Depending on the desired trade-off between resource savings and accuracy, users can select different pruning rates:

1. 20% pruning saves a modest amount of computational resources while maintaining nearly the same level of accuracy.
2. 40% pruning significantly reduces resource usage while still keeping the accuracy drop minimal.

This flexibility enables users to adjust the pruning rate based on the specific requirements of their application, such as maximizing efficiency for mobile devices or embedded systems.

8.2.5 Simplicity and Efficiency of the Pruning Process

The method is described as simple yet efficient. It uses L1 regularization during training to automatically identify less important dimensions in the Vision Transformer. Afterward, the model is pruned based on the learned importance scores, and a final fine-tuning step ensures that the performance remains high.

The simplicity of this pruning pipeline—training with L1 regularization, pruning, and fine-tuning—makes it easy to implement and apply to various Vision Transformer models. It is less complex than other compression techniques, such as quantization or knowledge distillation, but achieves comparable improvements in resource efficiency.

8.2.6 Promising Future Improvements

The authors suggest that this pruning method could be extended in future research to further reduce other components of Vision Transformers, such as:

1. Pruning attention heads within the multi-head self-attention layers.
2. Pruning transformer layers, which could lead to even more significant reductions in both parameters and computational costs.
3. These potential improvements open the door for even more efficient and compact Vision Transformers that can be deployed on a wider range of devices.

8.3 Overall Improvements:

1. Reduced parameters and FLOPs: Pruning 20-40% of the model's dimensions results in a significant reduction in both parameters and FLOPs, making the model more efficient for real-world applications.
2. Minimal accuracy loss: The method achieves large savings in computational resources while maintaining high accuracy, with only minor performance drops (as low as 0.5% in some cases).
3. Flexibility of pruning rates: Users can adjust the pruning rate depending on the trade-off they prefer between model size and performance, making the method adaptable to different use cases.
4. Simplicity and ease of implementation: The pruning method is straightforward to apply, relying on L1 regularization and a simple pruning and fine-tuning process.
5. Potential for future enhancements: The method can be expanded to further reduce other components like attention heads and transformer layers, providing even more opportunities for efficient Vision Transformers.

Chapter 9

Scalable MatMul-free Language Modeling [9]

This paper introduces a new approach to language modeling that focuses on reducing the computational cost and memory usage of deep learning models. Traditional models, like transformers, rely heavily on matrix multiplication (MatMul) operations, which are computationally expensive and require large amounts of memory. This makes them inefficient, especially for deployment on devices with limited resources, such as mobile phones or edge devices.

The authors propose a MatMul-free language model that replaces traditional matrix multiplication operations with bitlinear layers, which are much more efficient. These bitlinear layers use ternary weights (values of -1, 0, and 1) instead of continuous weights, significantly reducing the computational demands of the model. By avoiding matrix multiplication, the model becomes faster and consumes less memory, making it scalable and suitable for real-time applications in resource-constrained environments.

9.1 Key Contributions

9.1.1 Bitlinear Layers:

Instead of relying on standard matrix multiplication, the authors introduce bitlinear layers that perform more efficient operations using ternary weights. These layers simplify the model's calculations while maintaining a similar level of performance to traditional models.

9.1.2 Ternary Weights:

The use of ternary weights (-1, 0, 1) further reduces computational complexity. Ternary weights reduce the precision of the model's parameters without significantly affecting accuracy, enabling faster inference and lower memory usage.

9.1.3 Scalability and Efficiency:

The model is scalable, meaning it can handle larger language modeling tasks effectively, despite its simpler structure. It is particularly well-suited for use on devices with limited resources, where traditional models would be too computationally expensive to run efficiently.

9.1.4 Experimental Results:

Through extensive experiments on benchmark datasets, the authors demonstrate that their MatMul-free model achieves performance comparable to traditional models, such as Transformer++, while being significantly more efficient in terms of memory and computation.

9.2 Results and Improvements

9.2.1 Reduction in Computational Complexity

One of the main improvements of this method is the reduction in computational complexity:

1. By eliminating matrix multiplication, which is traditionally the most computationally expensive operation in models like transformers, the proposed MatMul-free architecture significantly reduces both computation time and memory usage.

2. The introduction of bitlinear layers replaces traditional dense layers, allowing for more efficient processing of input data. This results in faster inference, which is crucial for real-time applications, especially on resource-constrained devices.

9.2.2 Memory Efficiency

Another key result is the reduction in memory usage:

1. The model's use of ternary weights (-1, 0, 1) in bitlinear layers reduces the precision of the model's parameters. However, this lower precision leads to a substantial reduction in memory requirements without significantly affecting model accuracy.

2. This memory efficiency makes the model particularly well-suited for deployment on edge devices, where memory resources are often limited.

9.2.3 Comparable Performance with Traditional Models

Despite the simpler architecture and fewer computations, the MatMul-free model achieves competitive performance when compared to traditional language models like Transformer++. The key findings include:

1. On various language modeling benchmark tasks, the proposed model performs at a level that is comparable to traditional MatMul-based models.

2. Even with fewer parameters and reduced computation, the model maintains accuracy and is able to perform large-scale language modeling tasks effectively.

This demonstrates that the proposed model can offer the same level of accuracy while using fewer computational resources, which is a significant advancement in terms of efficiency.

9.2.4 Faster Inference Times

The MatMul-free model provides a significant improvement in inference speed:

1. The absence of matrix multiplication enables faster processing of inputs, which results in faster inference times compared to traditional transformer-based models. This makes the model suitable for real-time language modeling applications, where speed is critical.

2. The faster inference is particularly beneficial for deployment in low-power environments such as mobile devices, IoT devices, and edge computing, where computational power and speed are often constrained.

9.2.5 Scalability Across Model Sizes

The scalability of the MatMul-free model was evaluated through experiments on models with varying parameter sizes. The authors demonstrated that the architecture is highly scalable, meaning it can be used for both small-scale and large-scale language modeling tasks:

1. The model performs well with different parameter sizes, indicating that it can be scaled up or down depending on the specific application requirements.

2. This scalability proves that the MatMul-free approach is flexible and adaptable to various use cases, from small language tasks to more complex, large-scale applications.

9.2.6 Potential for Further Optimization Through Quantization

While the proposed model is already efficient, the authors highlight the potential for further improvements by integrating more advanced quantization techniques:

1. The use of ternary weights in the current model is a step towards quantization, but the authors suggest exploring more sophisticated quantization methods to further reduce computational and memory costs.

2. This opens the possibility of additional gains in efficiency, especially for tasks that require even more resource-constrained environments, such as embedded systems or real-time NLP applications.

9.2.7 Wider Applicability Beyond Language Modeling

The authors also hint at the possibility of extending their MatMul-free architecture to a broader range of natural language processing (NLP) tasks, beyond language modeling. Some potential areas for future exploration include:

1. Machine translation: The model's efficiency could be leveraged in translation tasks where speed and low resource consumption are critical.

2. Question answering: The model's scalability and efficiency make it a promising candidate for real-time question-answering systems, where both accuracy and quick response times are essential.

9.3 Overall Improvements:

1. Reduction in computational complexity: The elimination of matrix multiplication leads to faster and more efficient computation, making the model suitable for devices with limited computational resources.

2. Memory efficiency: By using ternary weights, the model reduces memory consumption without sacrificing much accuracy, enabling it to run on memory-constrained devices.

3. Comparable performance to traditional models: Despite its simpler architecture, the MatMul-free model achieves competitive results with traditional models like Transformer++, proving that it can handle large-scale language tasks effectively.

4. Faster inference times: The absence of matrix multiplication results in faster inference, which is critical for real-time applications, particularly in edge devices and mobile applications.

5. Scalability: The model scales well with different parameter sizes, demonstrating its adaptability for both small and large-scale tasks.

6. Opportunities for further optimization: The authors suggest further gains can be made by exploring additional quantization techniques, offering more potential for increased efficiency.

7. Wider applicability: The approach can be expanded to other NLP tasks beyond language modeling, such as machine translation and question answering, making it a versatile solution for various natural language processing applications.

Chapter 10

SpikeGPT: Generative Pre-trained Language Model with Spiking Neural Networks [10]

Spiking Neural Network??? (WTF!!!!!!!) I think this paper is outlier and i don't write literature review for this.

Chapter 11

Deepfake Detection Scheme Based on Vision Transformer and Distillation [11]

This paper focuses on developing a more effective and robust method for detecting deepfake videos, which are digitally manipulated videos created using advanced techniques like GANs (Generative Adversarial Networks). As deepfake videos become more widespread, especially in harmful applications like misinformation and fraud, the need for accurate detection methods has grown.

The authors propose a Vision Transformer-based model combined with a distillation technique for improved deepfake detection. The model leverages the strengths of both Vision Transformers (which are effective at capturing global image features) and EfficientNet (a powerful CNN that extracts local features). By combining these two architectures and adding a distillation token (which helps the model learn better by mimicking the output of a teacher network), the proposed scheme aims to achieve more accurate and robust detection of deepfake videos.

11.1 Main Components

11.1.1 Vision Transformer and EfficientNet Combination:

The model uses patch embedding to break the image into smaller sections, which are then processed using both EfficientNet (for local feature extraction) and Vision Transformer (for global feature analysis). This combination allows the model to capture both fine-grained details and broader contextual information in the video frames.

11.1.2 Distillation Token:

The authors introduce a distillation token, which improves the model's ability to learn from a teacher network (EfficientNet in this case). This method enhances the model's robustness and generalization, allowing it to perform better on unseen deepfake videos.

11.1.3 Performance Comparison:

The proposed model is tested against the state-of-the-art EfficientNet-based deepfake detection model. Through rigorous experiments, the authors demonstrate that their Vision Transformer model outperforms the previous best models in terms of accuracy, achieving a higher AUC score of 0.978 (compared to 0.972) and a better f1 score of 91.9% (compared to 90.6%).

11.2 Results and Improvements

11.2.1 Higher Accuracy and Better Performance Metrics

The proposed model achieved better accuracy compared to the previous state-of-the-art (SOTA) models for deepfake detection:

1. The model reached an AUC score of 0.978, which is an improvement over the 0.972 AUC achieved by the previous SOTA EfficientNet-based model.

2. In terms of the f1 score, which balances precision and recall, the new model achieved a score of 91.9%, outperforming the 90.6% f1 score of the previous best model.

These metrics indicate that the proposed method provides more reliable and accurate results in distinguishing between real and fake videos.

11.2.2 Improved Deepfake Detection Robustness

The combination of Vision Transformer (ViT) and EfficientNet in the detection model helps to capture both local and global features of the video frames, making the detection more accurate:

1. EfficientNet extracts local features, which are useful for identifying small artifacts or irregularities in individual frames.

2. The Vision Transformer captures global relationships across the entire image, improving the model's ability to identify deepfakes based on broader patterns.

This integration of local and global feature extraction results in a more robust detection model that performs well even in complex deepfake scenarios.

11.2.3 Reduction in False Negatives (Improved Detection of Fake Videos)

The model showed a significant improvement in detecting fake videos, as evidenced by a reduction in false negatives compared to the previous SOTA model. The confusion matrix presented in the paper shows that:

1. The proposed model reduced the number of false negatives (incorrectly classifying a fake video as real) from 335 to 187.

This is a critical improvement since reducing false negatives means the model is better at identifying deepfakes, which is essential for preventing the spread of harmful or misleading content.

11.2.4 Distillation Token for Better Generalization

The use of a distillation token plays a significant role in improving the model's learning capabilities and generalization:

1. The distillation token helps the student model (Vision Transformer) learn from the teacher network (EfficientNet), which has been pretrained on the same dataset.

2. During testing, the distillation token leads to better classification results than relying solely on the class token, improving the model's generalization to unseen data.

This technique helps prevent overfitting and makes the model more adaptable to new, unseen deepfake videos.

11.2.5 Clearer Prediction of Fake Videos

The paper includes a comparison of the model's predictions with the previous SOTA model. The results show that the proposed model provides clearer and more confident predictions when identifying fake videos:

1. The precision-recall curve shows that the proposed model consistently achieves better precision, meaning it is more confident when it identifies a video as fake.

2. The ROC-AUC curve also demonstrates that the proposed model has a larger area under the curve, reflecting improved performance in distinguishing between real and fake videos.

This indicates that the model is more reliable in its predictions, particularly when identifying deepfake content.

11.2.6 Better Loss Reduction in Training

The paper compares the training and validation loss of the proposed model against the SOTA EfficientNet-based model:

1. The loss curves show that the proposed model achieves a lower validation loss for both real and fake videos, demonstrating that it is better at learning from the training data and generalizing to new examples.

2. The lower loss indicates that the model is less prone to overfitting and performs better during testing.

11.3 Overall Improvements:

1. **Higher accuracy:** The model achieves a higher AUC score (0.978 vs. 0.972) and a better f1 score (91.9% vs. 90.6%) than the previous state-of-the-art model.

2. **Robustness to deepfake videos:** The integration of Vision Transformer and EfficientNet improves the model's ability to detect deepfakes by capturing both local and global features.

3. **Reduction in false negatives:** The model significantly reduces the number of false negatives, meaning it is better at identifying deepfake content and not misclassifying it as real.

4. **Distillation token for better generalization:** The use of the distillation token enhances the model's learning process, helping it generalize better to unseen deepfake videos.

5. **Clearer predictions:** The proposed model provides more confident and reliable predictions, as demonstrated by improved precision-recall and ROC-AUC curves.

6. **Lower validation loss:** The model's lower training and validation loss indicates better learning and generalization, making it more robust in real-world deepfake detection scenarios.

Chapter 12

DeepFakes: a New Threat to Face Recognition? Assessment and Detection [12]

This paper explores the growing challenge that deepfake videos pose to face recognition systems. Deepfakes are videos created using generative adversarial networks (GANs) to swap one person's face with another's, often in a realistic and convincing way. This technology has led to serious concerns, especially due to its use in creating misleading and harmful content, such as fake videos of celebrities in inappropriate situations. The paper investigates how vulnerable current face recognition systems are to deepfakes and evaluates different methods for detecting these tampered videos.

12.1 Deepfake Video Generation:

The authors used a GAN-based method to create deepfake videos by swapping faces in the publicly available VidTIMIT database. They generated both low-quality (64x64 pixels) and high-quality (128x128 pixels) versions of deepfakes, resulting in 620 deepfake videos. These videos are used to assess the impact on face recognition systems and to test the performance of detection methods.

12.2 Vulnerability of Face Recognition Systems:

The paper shows that two popular face recognition systems, based on VGG and Facenet neural networks, are highly vulnerable to deepfake videos. These

systems were fooled by high-quality deepfakes in 85-95% of cases, showing that they often cannot distinguish real faces from fake ones in manipulated videos.

12.3 Deepfake Detection Methods:

Several baseline methods were tested for detecting deepfakes. The study found that simple audio-visual approaches, such as lip-sync inconsistency detection, were not effective in distinguishing real videos from deepfakes. However, image-based methods using Image Quality Metrics (IQM) combined with Support Vector Machine (SVM) classifiers performed better, especially on high-quality deepfakes, achieving an equal error rate (EER) of 8.97%.

12.4 Challenges for Detection Systems:

The paper emphasizes that as deepfake technology continues to improve, it will become even harder to detect fake videos. The current detection methods, although promising, still face challenges, especially with high-quality deepfakes that are harder to detect.

12.5 Results and Improvements

12.5.1 Vulnerability of Face Recognition Systems

The paper highlights that face recognition systems, particularly those based on VGG and Facenet neural networks, are highly vulnerable to deepfake videos:

1. On high-quality deepfakes, the false acceptance rates (FAR) for VGG-based recognition was 85.62%, and for Facenet-based recognition, it was even higher at 95.00%.

2. These results indicate that current face recognition systems are often unable to distinguish deepfake videos from real ones, demonstrating the need for improved detection mechanisms.

12.5.2 Creation of a Public Deepfake Database

One of the major contributions of this paper is the creation of a publicly available database of deepfake videos:

1. The authors generated 620 deepfake videos from the VidTIMIT database using a GAN-based face-swapping approach. The database includes both low-quality (64x64 pixels) and high-quality (128x128 pixels) deepfakes, allowing researchers to evaluate the impact of deepfakes on face recognition systems and to develop new detection methods.

2. This database is an important resource for the research community, providing a standardized dataset for benchmarking deepfake detection methods.

12.5.3 Deepfake Detection Methods

Several baseline methods for detecting deepfakes were tested in the paper, with varying levels of success:

1. Lip-syncing detection (detecting inconsistencies between lip movements and speech in the audio) was found to be ineffective for detecting deepfake videos, as GAN-generated videos are often good at mimicking realistic lip movements.

2. The most effective method was based on image quality metrics (IQM) combined with a Support Vector Machine (SVM) classifier. This approach performed best for high-quality deepfakes, achieving an equal error rate (EER) of 8.97%, which is considered a reasonably good detection result for such challenging cases.

12.5.4 Performance on Low-Quality vs. High-Quality Deepfakes

The paper compares the effectiveness of detection methods on both low-quality and high-quality deepfakes:

1. The IQM + SVM approach achieved better detection results for low-quality deepfakes, with an EER of 3.33%, compared to 8.97% for high-quality deepfakes.

2. This shows that detecting high-quality deepfakes remains a more difficult challenge, even with the best-performing detection methods.

12.5.5 Improvements for Future Detection Systems

The paper discusses the limitations of current detection methods and suggests that:

1. As face-swapping and deepfake generation technology continues to improve, future detection methods will need to become more robust and generic to handle increasingly sophisticated deepfakes.

2. The development of new databases with more complex deepfakes and advanced detection algorithms will be essential for keeping up with the evolution of this technology.

12.6 Overall Results and Improvements:

1. **High vulnerability of face recognition systems:** The paper demonstrates that current face recognition systems, such as VGG and Facenet, are highly vulnerable to high-quality deepfake videos, with false acceptance rates as high as 95

2. **Public deepfake database:** The creation of a publicly available deepfake video database is a valuable resource for researchers to evaluate and develop deepfake detection methods.

3. **Effectiveness of image-based detection methods:** The IQM + SVM method is the most effective approach for detecting deepfakes, particularly on low-quality deepfakes, with an EER of 3.33% for low-quality videos and 8.97% for high-quality videos.

4. **Challenges with high-quality deepfakes:** The results show that detecting high-quality deepfakes is more challenging, and current detection methods still face difficulties with these more sophisticated fake videos.

5. **Future directions for detection methods:** The paper highlights the need for more robust and advanced detection techniques to cope with future developments in deepfake generation technology.

Chapter 13

Networks of spiking neurons: The third generation of neural network models [13]

Spiking Neural Network Again! I think this paper is outlier and i don't write literature review for this.

Chapter 14

DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance [14]

This paper provides a detailed study on how deepfake detection systems have evolved in recent years and examines the effectiveness of various detection methods. The authors focus on understanding how facial regions play a role in detecting manipulated videos and analyze different generations of deepfake datasets.

14.1 Key Idea

The paper examines the performance of state-of-the-art deepfake detection systems using two main approaches:

1. **Entire Face Analysis:** Traditional deepfake detection models that use the entire face as input for detecting manipulation.
2. **Facial Region Analysis:** A novel approach that focuses on analyzing specific facial regions (such as the eyes, nose, and mouth) to determine whether these areas can improve detection performance.

14.2 Key Contributions

1. **Comparing First- and Second-Generation Deepfakes:** The paper contrasts the differences in detection performance between older, lower-quality deepfake videos (1st generation) and newer, more realistic ones (2nd generation). It shows that while existing systems perform well on

older, less realistic deepfakes, they struggle to detect more sophisticated ones from the latest datasets.

2. **Evaluation of Specific Facial Regions:** The study analyzes the importance of different facial regions for detecting deepfakes. The authors find that focusing on regions such as the eyes and mouth can improve detection performance, especially when combined with traditional full-face analysis.

3. **Challenges with Second-Generation Deepfakes:** Even the best performing detection systems struggle with high-quality deepfakes from newer databases. Equal Error Rates (EERs) range from 15% to 30%, highlighting that more research is needed to improve the robustness of detection methods against these advanced deepfakes.

14.3 Results and Improvements

14.3.1 Improved Detection Using Facial Region Analysis

One of the key findings of the paper is that focusing on specific facial regions can enhance deepfake detection performance:

1. The study shows that concentrating on areas such as the eyes and mouth yields better detection results compared to analyzing the entire face alone. These regions are often harder to manipulate realistically in deepfake videos, making them useful focal points for detection systems.

2. Combining facial region analysis with traditional full-face approaches leads to improved performance in identifying deepfakes, particularly in challenging second-generation deepfakes.

14.3.2 Comparison of First and Second Generation Deepfakes

The paper examines the differences between older, less realistic deepfakes (first-generation) and newer, high-quality deepfakes (second-generation):

1. The results show that first-generation deepfakes are easier to detect, with detection systems achieving higher accuracy and lower error rates.

2. In contrast, second-generation deepfakes—which are more sophisticated and harder to distinguish from real videos—pose a significant challenge to current detection models. The study highlights that modern detection systems struggle with these newer, more advanced deepfakes, resulting in Equal Error Rates (EERs) ranging from 15% to 30%.

This finding emphasizes the growing difficulty of detecting second-generation deepfakes and the need for more advanced techniques to address this issue.

14.3.3 Importance of Facial Artifacts in Detection

The paper identifies certain artifacts that are more pronounced in specific facial regions, which can be leveraged for better detection:

1. For example, regions like the eyes and mouth exhibit subtle inconsistencies in deepfake videos, such as unnatural blinking patterns, lip-sync mismatches, or texture anomalies. These artifacts provide valuable clues that can improve detection accuracy.

2. By focusing on these facial artifacts, the detection model can better differentiate between real and fake videos, especially when combined with a full-face analysis.

14.3.4 Challenges with High-Quality Deepfakes

The results show that detecting second-generation deepfakes remains a major challenge, even for state-of-the-art detection systems:

1. While current models perform well on older, lower-quality deepfakes, their performance significantly drops when tested against high-quality deepfakes from newer datasets.

2. The EERs of modern detection systems range from 15% to 30% for second-generation deepfakes, indicating that these deepfakes are becoming increasingly difficult to identify accurately.

This highlights the need for continued research and improvement in detection methods to keep up with the evolution of deepfake technology.

14.3.5 Benchmarking and Dataset Evaluation

The paper contributes to the research community by benchmarking detection models on first- and second-generation deepfake datasets:

1. It shows that while detection models trained on older deepfake datasets perform well in detecting similar types of deepfakes, they often fail to generalize to second-generation deepfakes.

2. This benchmarking is essential for identifying the gaps in current detection methods and for driving future improvements in detecting more advanced deepfakes.

14.4 Overall Results and Improvements

1. Facial region analysis improves detection: By focusing on specific areas like the eyes and mouth, detection models can achieve better results, particularly on more challenging deepfakes.

2. Detection challenges with second-generation deepfakes: The paper highlights the significant drop in performance when detecting high-quality, second-generation deepfakes, with EERs ranging from 15% to 30%. This underscores the growing sophistication of deepfake technology.

3. Importance of facial artifacts: Certain facial regions, such as the eyes and mouth, often contain subtle but detectable artifacts in deepfake videos, which can be used to improve detection accuracy.

4. Benchmarking on evolving datasets: The paper provides an important evaluation of detection models across both first- and second-generation deepfake datasets, revealing weaknesses in current methods and emphasizing the need for ongoing research.

5. Future research directions: The study suggests that more advanced methods focusing on specific facial regions and artifacts should be developed to cope with the increasing quality of deepfakes in future generations.

Chapter 15

Exposing Deep Fakes Using Inconsistent Head Poses [15]

This paper presents a novel method to detect deepfake videos by analyzing inconsistencies in 3D head poses. Deepfakes are created by manipulating and splicing synthetic face regions onto real videos, and this process often introduces errors in how the face aligns with the natural head movement of the person.

The authors observed that many deepfakes fail to maintain consistent head poses due to the synthetic manipulation of the facial regions. By estimating 3D head poses from the face images in videos, they found that deepfakes exhibit noticeable inconsistencies that can be used as a cue for detection. The paper develops a detection method that uses these inconsistencies to distinguish real videos from AI-generated deepfakes.

15.1 Method

1. **3D Head Pose Estimation:** The method estimates the 3D orientation of the head from facial landmarks in both real and deepfake videos. It uses these estimates to check if the head movement is consistent with the facial orientation, which is often misaligned in deepfakes.

2. **Classification Based on Inconsistencies:** After detecting these pose inconsistencies, the method applies a classification approach that distinguishes between real videos and deepfakes based on the presence of such errors.

3. **Experimental Results:** The authors tested their method on a dataset containing both real face images and deepfakes. The results show that this

approach can effectively expose deepfakes by identifying misalignments in head poses.

15.2 Results and Improvements

15.2.1 Effective Detection of Deepfakes Using Head Pose Inconsistencies

The core result of the paper is the demonstration that inconsistent 3D head poses can be a reliable indicator of deepfake videos:

1. The method estimates the 3D orientation of the head from facial landmarks and identifies mismatches between the facial region and the overall head pose.

2. The study shows that deepfakes often exhibit significant misalignments in head poses because the synthetic facial regions do not align naturally with the original head movement in the video.

This finding highlights that head pose analysis can be an effective, low-cost method for detecting deepfakes, particularly those created with simpler or less sophisticated techniques.

15.2.2 High Detection Accuracy

The method proposed in the paper achieves high accuracy in distinguishing real videos from deepfakes:

1. The classification method based on head pose inconsistencies was able to successfully expose deepfakes in the tested dataset, demonstrating its reliability as a detection tool.

2. The approach is particularly effective at identifying deepfakes that involve subtle or small facial manipulations, which are often harder to detect with other methods.

This result confirms that head pose analysis is a promising direction for improving the accuracy of deepfake detection systems.

15.2.3 Simplicity and Efficiency

One of the main advantages of the proposed method is its simplicity and computational efficiency:

1. The method does not require complex models or extensive computational resources, making it suitable for real-time applications or scenarios where computational power is limited.

2. By focusing on a specific feature—head pose consistency—the approach reduces the need for large datasets or deep learning models, which are often required by other detection techniques.

This improvement in efficiency makes the method accessible for broader use, including on devices with limited processing capabilities.

15.2.4 Applicability Across Different Types of Deepfakes

The paper demonstrates that the method is effective across various types of deepfakes:

1. Whether the deepfakes involve simple face-swapping or more complex facial region synthesis, the head pose inconsistency approach was able to detect anomalies.

2. The method's robustness across different manipulation techniques suggests that it can serve as a versatile tool in the ongoing battle against deepfake videos.

This versatility is a significant improvement over some detection methods that may be tailored to specific types of deepfakes but struggle with others.

15.2.5 Potential for Integration with Other Detection Techniques

While the method is effective on its own, the paper suggests that it could be integrated with other detection methods to further enhance overall detection performance:

1. By combining head pose inconsistency analysis with other techniques, such as texture analysis or temporal inconsistencies, the detection system could achieve even higher accuracy and robustness.

2. This potential for integration makes the method a valuable component in a multi-layered defense against deepfakes.

15.3 Overall Results and Improvements

1. **Reliable detection through head pose inconsistencies:** The method effectively identifies deepfakes by analyzing misalignments in 3D head poses, offering a reliable indicator of manipulation.

2. **High accuracy:** The proposed approach achieves high detection accuracy, particularly for deepfakes that involve subtle facial manipulations.

3. **Simplicity and efficiency:** The method is computationally efficient and simple to implement, making it suitable for real-time detection and use in resource-constrained environments.

4. **Versatility across different deepfake types:** The technique is effective across various deepfake generation methods, enhancing its applicability and robustness.

5. **Potential for enhanced detection systems:** The method can be integrated with other detection techniques to create a more comprehensive and effective deepfake detection system.

Chapter 16

Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers [16]

This paper investigates the effectiveness of two deep learning approaches—Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs)—in detecting deepfake videos. Deepfakes, which use AI to create realistic manipulated videos, pose a significant threat in spreading misinformation and harming individuals, making effective detection methods critical.

16.1 Key Ideas

1. **Comparison of CNNs and Transformers:** The authors compare CNNs and ViTs to understand which method is more effective at identifying deepfakes. CNNs have traditionally been used in image and video analysis tasks, while transformers, initially popular in natural language processing, have recently shown promise in computer vision tasks, including deepfake detection.

2. **Deepfake Datasets:** The study tests both CNN and Transformer models on several prominent deepfake datasets, including FaceForensics++ 2020 (FF++ 2020), Google DeepFake Detection (DFD), Celeb-DF, Deeper Forensics, and DeepFake Detection Challenge (DFDC). These datasets contain a variety of deepfake videos, from simpler to more sophisticated manipulations.

3. **Evaluation Metrics:** The models' performance is evaluated based on accuracy and Area Under the Curve (AUC) scores. These metrics measure how well the models distinguish between real and fake videos.

16.2 Key Findings:

1. **CNNs vs. Transformers:** The study shows that while both CNNs and Transformers perform well, Vision Transformers (ViTs) tend to outperform CNNs in detecting deepfakes, particularly on newer and more complex datasets like Celeb-DF and DFDC. This indicates that transformers may be better at capturing subtle patterns in manipulated videos.

2. **Best Detection Results:** The highest detection accuracies and AUC scores were achieved for various datasets, with 99.53% accuracy and 100% AUC for detecting deepfakes from the Google DFD dataset, and 99.73% accuracy and 99.99% AUC on the Deeper Forensics dataset, showing the models' high effectiveness in detecting manipulations.

3. **Strengths of Each Model:** CNNs are still highly effective and faster in terms of computation, making them suitable for real-time detection on simpler datasets. However, Transformers demonstrate superior performance on more sophisticated deepfakes, suggesting that they are better at detecting subtle irregularities in the manipulated content.

Chapter 17

Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks [17]

This presents a method that combines face detection and facial landmark alignment in a single, unified framework. The authors propose using a cascaded structure of Convolutional Neural Networks (CNNs) to improve the accuracy and efficiency of both tasks. This approach is designed to handle challenging conditions such as various poses, lighting conditions, and occlusions.

17.1 Main Idea

The main idea is to jointly perform face detection and facial landmark alignment using a multi-task learning framework. By integrating these two tasks, the model can exploit their natural correlation, which leads to improved performance. The proposed framework is made up of three stages of CNNs:

1. **Stage 1:** A shallow CNN, called P-Net, quickly generates candidate face windows.
2. **Stage 2:** The R-Net refines these candidates by rejecting false positives and further calibrating the bounding boxes.
3. **Stage 3:** The O-Net outputs the final refined face bounding boxes and facial landmark locations.

The method also introduces online hard sample mining, which automatically selects difficult samples during training to improve the model's robustness, without the need for manual sample selection.

17.2 Main Contributions

1. **Unified Framework:** The authors propose a unified cascaded CNN framework that jointly performs face detection and facial landmark alignment, leveraging the inherent connection between these tasks.

2. **Improved Performance:** The model significantly outperforms previous methods on key benchmarks such as FDDB and WIDER FACE for face detection, and AFLW for facial landmark alignment.

3. **Efficiency:** The cascaded structure ensures that the model is lightweight and achieves real-time performance, making it suitable for practical applications.

17.3 Results and Improvements

17.3.1 Improved Accuracy on Challenging Benchmarks

The proposed multi-task cascaded CNN framework achieves superior accuracy in both face detection and facial landmark alignment tasks. Key results include:

1. On the FDDB (Face Detection Data Set and Benchmark) and WIDER FACE datasets for face detection, the method outperforms state-of-the-art techniques by a significant margin.

2. For facial landmark alignment, the framework shows strong results on the AFLW (Annotated Facial Landmarks in the Wild) benchmark, where it consistently outperforms previous methods.

These results demonstrate the model's robustness and ability to handle challenging real-world conditions like varying lighting, poses, and occlusions.

17.3.2 Joint Face Detection and Alignment Improves Both Tasks

By combining face detection and facial landmark alignment into a single framework, the model leverages the inherent correlation between these tasks, leading to improvements in both areas:

1. Face detection benefits from the additional information provided by facial landmarks, making the detection process more accurate.

2. Facial landmark alignment is improved by the enhanced face detection process, which provides better bounding box localization for the landmarks.

This joint approach enables the model to perform better in real-world scenarios, where detecting faces and aligning facial features simultaneously can boost accuracy.

17.3.3 Efficiency and Real-Time Performance

The proposed cascaded CNN structure is designed to be lightweight and computationally efficient:

1. The model achieves real-time performance, processing at 16 frames per second (fps) on a CPU and 99 fps on a GPU (Nvidia Titan Black).
2. This efficiency makes the framework suitable for practical applications, where speed is critical, such as in facial recognition systems, video surveillance, or mobile device deployment.

The model's real-time performance is an important improvement over existing techniques, which often require significant computational resources and processing time.

17.3.4 Online Hard Sample Mining

The introduction of online hard sample mining during training improves the model's performance:

1. The method automatically selects the hardest samples during the training process, focusing the learning on challenging cases. This reduces the need for manual sample selection and helps the model learn more effectively.
2. As shown in experiments, this approach leads to better convergence and overall performance without the additional complexity of offline sample mining.

The inclusion of this technique provides an important improvement in the model's training efficiency and accuracy.

17.3.5 Lightweight CNN Architecture

The model's CNN architecture is specifically designed to balance performance and computational cost:

1. By using smaller 3×3 filters instead of larger 5×5 filters, the model reduces the number of parameters and the overall computational complexity.
2. This architecture allows for deeper networks without increasing the processing time, leading to better performance in both face detection and alignment tasks.

The architectural improvements result in a model that is both more powerful and more efficient than previous designs.

17.3.6 Versatility Across Multiple Datasets

The framework has been tested across multiple benchmarks and datasets, showing its versatility:

1. It performs well on both face detection benchmarks (FDDB and WIDER FACE) and facial landmark alignment benchmarks (AFLW), proving that it can generalize across different tasks and environments.

2. The method handles various challenges, such as occlusions, pose variations, and extreme lighting conditions, making it highly adaptable to real-world scenarios.

17.4 Overall Results and Improvements

1. **Higher accuracy:** The model achieves state-of-the-art results in face detection and facial landmark alignment tasks on challenging datasets like FDDB, WIDER FACE, and AFLW.

2. **Joint task improvement:** Combining face detection and facial landmark alignment improves the performance of both tasks, leveraging their inherent correlation.

3. **Real-time performance:** The model is highly efficient, achieving up to 99 fps on a GPU, making it suitable for real-time applications.

4. **Online hard sample mining:** This innovative technique improves the model's learning process by focusing on hard samples during training, leading to better accuracy without the need for manual intervention.

5. **Efficient CNN architecture:** The lightweight CNN design with 3×3 filters provides a good balance between depth and computational efficiency, enhancing performance while reducing complexity.

6. **Versatility across datasets:** The model performs well across a variety of datasets and conditions, making it adaptable to real-world face detection and alignment challenges.

Chapter 18

BiLLM: Pushing the Limit of Post-Training Quantization for LLMs [18]

This paper focuses on improving the efficiency of Large Language Models (LLMs) through a process called post-training quantization. As LLMs like GPT, BERT, and others continue to grow in size and computational requirements, deploying these models on devices with limited resources, such as mobile phones and edge devices, becomes challenging. Quantization is a method used to reduce the size and computational cost of these models by compressing their parameters, typically by reducing the precision of the numerical data used to represent them (for example, converting 32-bit floating-point numbers to 8-bit integers).

The main goal of this paper is to push the limits of post-training quantization for LLMs without sacrificing accuracy. BiLLM introduces new techniques that allow for aggressive quantization, making it possible to run large models on smaller devices while maintaining competitive performance.

18.1 Main Contributions

1. **Improved Post-Training Quantization:** The paper presents a new approach to quantizing LLMs after they have been trained, which is different from quantization-aware training that happens during the training process. By focusing on post-training quantization, the authors aim to make it easier to apply to models that have already been trained.
2. **Maintaining Model Accuracy:** A major challenge with quantization is the loss of accuracy that can occur when reducing precision. The authors propose techniques that minimize the impact of quantization on model performance, ensuring that LLMs can still deliver high-quality results even after being compressed.

3. **Wide Applicability:** The BiLLM method is designed to be applied to various LLM architectures, meaning it can work across different types of models, making it a versatile solution for reducing computational requirements in a wide range of applications.

18.2 Results and Improvements

18.2.1 Significant Model Size Reduction

The BiLLM method achieves substantial model size reduction by applying aggressive post-training quantization:

1. The paper demonstrates that BiLLM can compress LLMs by reducing the precision of the weights from 32-bit to lower-bit representations (such as 8-bit or even lower), dramatically decreasing the model's memory footprint.

2. This reduction enables the deployment of large language models on devices with limited resources, such as mobile phones, edge devices, or low-powered hardware, making LLMs more accessible for real-world applications.

18.2.2 Minimal Accuracy Loss

One of the key improvements of BiLLM is that it achieves this compression with minimal loss in accuracy:

1. The paper shows that despite aggressive quantization, the accuracy drop is small, allowing the models to maintain high-quality performance even after compression.

2. BiLLM achieves this through techniques that carefully handle the quantization process, ensuring that the critical information in the model is preserved while reducing less significant elements.

This is a crucial improvement over traditional quantization methods, where the performance drop can often be significant, making BiLLM a more effective solution for practical use.

18.2.3 Wide Applicability Across Model Architectures

The BiLLM method is designed to be versatile and applicable to various large language models:

1. The authors tested BiLLM on different LLM architectures, and the results show that the method works across a wide range of models, making it a general solution for compressing LLMs.

2. Whether applied to transformer-based models like BERT, GPT, or other types of architectures, BiLLM consistently provides compression with minimal performance degradation.

This wide applicability makes BiLLM a robust technique for optimizing LLMs in diverse settings.

18.2.4 Improved Computational Efficiency

In addition to reducing model size, BiLLM significantly improves the computational efficiency of LLMs:

1. With fewer bits required to represent the model parameters, the quantized models can be processed faster, leading to reduced computational load and quicker inference times.

2. This improvement is especially beneficial for real-time applications or scenarios where latency is critical, such as in mobile applications or cloud-based services with limited computational resources.

The increased efficiency in terms of both speed and resource usage makes BiLLM an attractive option for deploying LLMs in practical environments.

18.2.5 Post-Training Quantization Simplicity

The paper highlights that BiLLM focuses on post-training quantization, which simplifies the process of optimizing LLMs:

1. Since the method is applied after the model has been trained, it can be easily implemented on any pretrained model without modifying the training process. This is a significant advantage over quantization-aware training, which requires adjusting the model during the training phase.

2. This simplicity makes BiLLM a more flexible and convenient method for compressing existing large language models, offering a straightforward way to reduce their size and improve efficiency without retraining.

18.2.6 Potential for Future Extensions

The paper concludes with the potential for further extensions of BiLLM:

1. The authors suggest that the method can be further refined to push the limits of quantization even more, exploring lower-bit precision and more sophisticated techniques for maintaining accuracy while aggressively reducing the model's size.

2. This opens up possibilities for future research to improve the scalability and effectiveness of post-training quantization for increasingly larger and more complex models.

18.3 Overall Results and Improvements

1. **Significant size reduction**: BiLLM achieves substantial compression of large language models by applying aggressive post-training quantization, reducing memory usage and enabling deployment on resource-constrained devices.

2. **Minimal accuracy loss:** Despite the aggressive quantization, BiLLM maintains high model performance with only a small drop in accuracy, making it a reliable method for compressing LLMs without sacrificing quality.

3. **Versatility across models:** The method is effective on a wide range of LLM architectures, making it a general solution for post-training quantization that can be applied to various transformer-based and other models.

4. **Improved computational efficiency:** BiLLM reduces computational load and speeds up inference times, providing practical benefits for real-time and latency-sensitive applications.

5. **Simplicity of post-training quantization:** The method can be applied to pretrained models without modifying the training process, making it easy to use and implement.

6. **Future potential:** The paper highlights opportunities for further optimization, including exploring more aggressive quantization techniques to push the limits of model compression.

Chapter 19

RepQuant: Towards Accurate Post-Training Quantization of Large Transformer Models via Scale Reparameterization [19]

This paper introduces a new method called RepQuant to improve the efficiency and accuracy of post-training quantization for large transformer models. As transformer models grow larger and more complex, reducing their size and computational needs without losing performance is essential for deploying them in real-world applications, especially on devices with limited resources.

The authors propose a quantization-inference decoupling approach to separate the complex quantization process from the inference stage, which is the part that runs on hardware. RepQuant uses scale reparameterization to connect the complex and hardware-friendly quantizers, ensuring that the model's behavior remains consistent even after quantization. This method makes it possible to use simple, hardware-compatible quantizers during inference while maintaining the accuracy of more complex quantization during training.

19.1 Main Contributions

1. **Post-Training Quantization (PTQ):** RepQuant is designed to work after the model is trained, making it easy to apply to existing large models without requiring retraining. This is critical for practical deployment.

2. **Scale Reparameterization:** By introducing scale reparameterization, the method ensures that quantization errors are minimized, and the model's accuracy remains close to that of the original model, even after applying aggressive quantization.

3. **Hardware-Friendly Inference:** The method enables the use of simple quantizers during inference, which are compatible with hardware accelerators, ensuring efficient computation without compromising performance.

19.2 Results and Improvements

19.2.1 Improved Accuracy in Post-Training Quantization

RepQuant achieves significant improvements in maintaining model accuracy after quantization:

1. Post-training quantization often leads to a drop in performance, but RepQuant minimizes this accuracy loss through its scale reparameterization technique, which ensures a closer alignment between the original and quantized models.

2. This framework allows for more precise quantization of large-scale transformers, such as ViTs (Vision Transformers) and language models, without requiring additional retraining or complex adjustments.

The paper reports high accuracy retention even when using lower-bit quantization, making the approach effective for reducing model size and computational demands while maintaining model performance.

19.2.2 Scale Reparameterization Technique

One of the main innovations in RepQuant is the introduction of scale reparameterization, which helps to bridge the gap between training and inference quantizers:

1. This technique decouples the training quantizers from the inference quantizers, allowing for the use of complex quantizers during the training process while keeping hardware-friendly quantizers for inference.

2. The mathematically equivalent transformation between these two quantizers ensures consistency and improves the overall efficiency of the model post-quantization.

This improvement results in better quantization performance, especially for large transformer models, where the precision of quantization plays a critical role in maintaining accuracy.

19.2.3 Versatility Across Large Transformer Models

RepQuant is designed to be versatile and applicable across various transformer architectures:

1. The method has been tested on Vision Transformers (ViTs), and other large-scale transformer models, demonstrating its broad applicability to different architectures.

2. By applying RepQuant to different model types, the authors show that the method can be used to optimize a wide range of transformer models without the need for architecture-specific modifications.

This versatility makes RepQuant a general solution for quantizing large transformers in both computer vision and natural language processing applications.

19.2.4 Significant Reduction in Computational Complexity

RepQuant enables a significant reduction in computational complexity while preserving performance:

1. By using low-bit quantization with minimal loss of accuracy, the method drastically reduces the computational overhead required to deploy large models on resource-constrained devices.

2. This makes RepQuant particularly useful for applications requiring real-time performance, such as mobile devices or edge computing systems, where resource efficiency is critical.

The reduction in computational complexity also translates to lower energy consumption, making it an environmentally friendly solution for deploying large transformer models.

19.2.5 Hardware Compatibility and Efficiency

One of the strengths of RepQuant is its focus on hardware compatibility:

1. The framework is designed to work seamlessly with hardware-friendly quantizers, making it easier to deploy quantized models on various hardware platforms, including GPUs, TPUs, and specialized AI chips.

2. The ability to use simple quantizers for inference ensures that the models are compatible with existing hardware while maintaining high performance.

This hardware efficiency makes RepQuant an attractive option for industries looking to optimize the deployment of large models on embedded systems or other constrained environments.

19.2.6 Potential for Future Extensions

The paper concludes with the possibility of extending RepQuant for further quantization optimizations:

1. The authors suggest that additional research could explore even lower-bit quantization or other transformer architectures to continue pushing the boundaries of post-training quantization.

2. This opens the door for further advancements in model compression, making RepQuant a promising foundation for future work in the field of model optimization.

19.2.7 Overall Results and Improvements

1. Improved post-training quantization accuracy: RepQuant minimizes accuracy loss by using scale reparameterization, achieving superior precision in low-bit quantization of large transformer models.

2. Introduction of scale reparameterization: This novel technique bridges the gap between training and inference quantizers, enabling more accurate and efficient post-training quantization.

3. Versatility across transformer models: RepQuant is applicable to a variety of large transformer models, demonstrating its flexibility and broad utility across different architectures.

4. Reduction in computational complexity: The method enables significant reductions in computational demands, making it ideal for deploying large models in resource-constrained environments.

5. Hardware compatibility: By ensuring that the quantized models are compatible with hardware-friendly quantizers, RepQuant facilitates easy deployment on various hardware platforms.

6. Potential for further improvements: The framework offers opportunities for future research to push the limits of quantization, potentially achieving even greater compression with minimal accuracy loss.

Chapter 20

Post-Training Quantization for Vision Transformer [20]

The paper introduces a new approach to post-training quantization for vision transformers, aimed at making these models more efficient in terms of memory usage and computational speed without sacrificing accuracy. Vision transformers, while powerful, are resource-heavy, and this paper tackles the problem of reducing their complexity to allow deployment on devices with limited resources, such as mobile phones or embedded systems.

The proposed method adjusts the bit-widths of different layers based on their sensitivity, as measured by the nuclear norm of the attention maps and output features, rather than applying a uniform quantization across the model. The core of the approach involves finding optimal quantization intervals to minimize the difference between the original and quantized outputs. The researchers also introduce a ranking loss function to preserve the relative order of attention values, which is critical to the performance of transformers, and a bias correction mechanism to further reduce quantization errors.

Extensive experiments on several datasets show that this method significantly outperforms existing post-training quantization techniques, delivering better accuracy with lower memory and computational costs. This makes it easier to deploy vision transformers in real-world applications where hardware constraints are an issue.

20.1 Results and Improvements

20.1.1 Mixed-Precision Quantization:

1. The method dynamically assigns different bit-widths to each layer based on the layer's sensitivity, calculated using the nuclear norm of the attention map and output features.

2. This approach results in higher compression rates while maintaining or improving accuracy, making the models more efficient in terms of memory and computation.

20.1.2 Higher Accuracy on Quantized Models:

1. The proposed method achieves a Top-1 accuracy of 81.29% on the DeiT-B model (8-bit quantization) using the ImageNet dataset. This is comparable to the full-precision baseline and outperforms other quantization methods such as EasyQuant and Bit-Split.

2. The proposed method achieves a Top-1 accuracy of 81.29% on the DeiT-B model (8-bit quantization) using the ImageNet dataset. This is comparable to the full-precision baseline and outperforms other quantization methods such as EasyQuant and Bit-Split.

3. On the ViT-B model, the 8-bit quantized version achieves 97.79% accuracy on CIFAR-10 and 85.76% on CIFAR-100, improving on existing quantization techniques by a significant margin.

20.1.3 Improved Performance on Smaller Models

1. For smaller models like DeiT-S, the proposed method reduces accuracy loss significantly when compared to percentile-based quantization, with accuracy losses of 5.22% (6-bit) and 2.33% (8-bit), improving model efficiency without major accuracy degradation.

20.1.4 Bias Correction for Error Reduction:

1. The paper introduces a bias correction mechanism to reduce accumulated quantization errors, which helps maintain the performance of quantized models close to their full-precision counterparts.

20.1.5 Ranking Loss for Attention Preservation:

1. A novel ranking loss function ensures that the quantization process maintains the correct order of attention values, which is crucial for the vision transformer's performance. This helps prevent significant drops in accuracy.

20.1.6 Memory and Computation Savings:

1. The mixed-precision quantization scheme saves about 25% memory and 44% computational costs compared to traditional post-training quantization methods, particularly on larger models like ViT-L.

20.1.7 Generalization to Object Detection:

1. The method also shows generalization capability for object detection tasks using the DETR model, with an improvement in mAP of 40.5 (6-bit) and 41.7 (8-bit) on the COCO2017 dataset, further proving its versatility across different tasks.

Chapter 21

The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits [21]

This paper introduces a breakthrough in optimizing large language models (LLMs) through a new approach called BitNet b1.58, a 1-bit variant of traditional LLMs. The goal is to significantly reduce the memory, energy, and computational costs of LLMs without sacrificing performance. Current LLMs use 16-bit precision for model parameters, which makes them resource-intensive and challenging to run on limited hardware like mobile devices or edge devices.

The proposed BitNet b1.58 uses a more efficient representation of weights, limiting them to -1, 0, 1, resulting in models that operate at 1.58-bit precision. This allows for much faster computations, reduced memory usage, and lower latency while maintaining the same level of accuracy as full-precision models (like FP16). The approach also enables energy-efficient matrix multiplication, which is a major cost factor in LLMs, offering substantial savings.

BitNet b1.58 matches or even surpasses the performance of LLaMA, a leading LLM, especially at larger model sizes. For example, the 3B parameter version of BitNet b1.58 outperforms its LLaMA counterpart while using significantly fewer resources. As the model scales up (7B, 13B, 70B), BitNet b1.58 offers even more efficiency, being 4.1 times faster and using 7 times less memory at the largest sizes.

The paper also calls for the development of new hardware optimized for 1-bit LLMs to further enhance performance and reduce costs, paving the way for highly efficient AI applications on a wide range of devices, including mobile and edge hardware. This innovation opens up possibilities for deploying powerful LLMs in more resource-constrained environments while maintaining high performance and reducing environmental impact.

21.1 Results and Improvements

21.1.1 Reduction in Memory Usage and Latency:

1. BitNet b1.58 significantly reduces memory usage and latency compared to full-precision models like LLaMA.
2. For instance, at 3B model size, BitNet b1.58 uses 3.55 times less GPU memory and is 2.71 times faster than the 3B LLaMA model.
3. As the model scales up, the advantages become even more pronounced. At 70B model size, BitNet b1.58 is 4.1 times faster and consumes 7.16 times less memory compared to LLaMA.

21.1.2 Energy Efficiency:

1. BitNet b1.58 significantly improves energy efficiency due to its use of integer operations (INT8) instead of floating-point operations (FP16).
2. On a 7nm chip, BitNet b1.58 reduces energy consumption by 71.4 times in terms of arithmetic operations compared to LLaMA.
3. As models grow larger, BitNet b1.58 becomes increasingly efficient, offering up to 41.2 times lower energy consumption at 70B parameters.

21.1.3 Throughput Improvement:

1. At 70B model size, BitNet b1.58 offers an 8.9 times higher throughput and can support up to 11 times larger batch sizes than LLaMA on the same hardware (two 80GB A100 cards).
2. This allows for more efficient large-scale inference, making BitNet b1.58 more suitable for demanding real-world applications.

21.1.4 Maintained Accuracy:

1. BitNet b1.58 achieves performance parity with full-precision models (FP16) in terms of perplexity and zero-shot task accuracy.
2. For example, the 3B BitNet b1.58 model matches the perplexity and accuracy of the 3B LLaMA model, while using fewer resources.
3. The 3.9B version of BitNet b1.58 even outperforms the 3B LLaMA model with lower memory and latency costs.

21.1.5 Performance on Zero-Shot Tasks:

1. The method shows excellent zero-shot performance across several tasks (ARC-Easy, Hellaswag, Winogrande, etc.), closing the gap between BitNet b1.58 and full-precision models as the model size increases.

2. For instance, at 3.9B size, BitNet b1.58 performs better than the 3B LLaMA on various language benchmarks, further highlighting its effectiveness in natural language processing tasks.

21.1.6 New Scaling Laws:

1. BitNet b1.58 introduces new scaling laws, showing that as model size increases, its efficiency improvements over traditional FP16 models become more significant.

2. For example, a 13B BitNet b1.58 model is more efficient than a 3B FP16 model, and a 70B BitNet b1.58 model is more efficient than a 13B FP16 model in terms of memory usage, latency, and energy consumption.

21.1.7 Hardware Optimization Potential:

1. The paper calls for the design of new hardware systems specifically optimized for 1-bit LLMs like BitNet b1.58. Such hardware could further reduce computation costs and enable even more efficient deployment of large models.

21.2 Overall Improvements

1. 2.71x faster inference with 3.55x less memory at the 3B model size.

2. 71.4x reduction in arithmetic operations energy consumption.

3. Up to 11x larger batch sizes and 8.9x higher throughput at the 70B model size.

4. No compromise on accuracy, with performance parity with full-precision models.

Chapter 22

Root Mean Square Layer Normalization [22]

This paper a new method designed to improve the efficiency of deep neural networks. The current widely used method, LayerNorm, helps stabilize training and improve performance but comes with a significant computational cost, particularly in larger models. The authors of this paper propose RMSNorm as a simpler and faster alternative that removes the re-centering step of LayerNorm, focusing only on re-scaling. This makes RMSNorm computationally more efficient while still stabilizing network training.

The key idea behind RMSNorm is that re-centering (adjusting inputs based on their mean) is not essential for model performance, while re-scaling (adjusting inputs based on their size) is the crucial factor. RMSNorm uses the root mean square of the inputs to normalize the data, which ensures that the model remains robust to changes in input size or weight scaling. The paper also proposes pRMSNorm, a variant of RMSNorm that estimates normalization based on only a subset of inputs, further reducing computational overhead.

Extensive experiments across various tasks like machine translation, image classification, and question answering demonstrate that RMSNorm performs on par with LayerNorm in terms of accuracy while providing significant speed improvements. Depending on the model and task, RMSNorm speeds up training by 7% to 64%, making it a promising alternative for improving the efficiency of neural networks without compromising performance.

In summary, this paper presents a simpler and faster normalization technique, RMSNorm, that maintains model performance while significantly reducing the computational costs associated with LayerNorm, making it ideal for larger and more complex networks.

22.1 Results and Improvements

22.1.1 Reduction in Computational Overhead:

1. RMSNorm simplifies the normalization process by removing the re-centering step, which is computationally expensive in LayerNorm.

2. Across different models and tasks, RMSNorm reduces running time by 7% to 64%, depending on the architecture and task. This makes it a much more efficient option without sacrificing model performance.

22.1.2 Comparable Accuracy to LayerNorm:

1. In terms of model accuracy, RMSNorm performs on par with LayerNorm across a range of tasks, including machine translation, image classification, and question answering.

2. In tasks like machine translation, RMSNorm matches LayerNorm in BLEU scores while providing speed advantages. In the Transformer model, for instance, RMSNorm achieved similar accuracy to LayerNorm but with a 7% to 9% speedup.

22.1.3 Efficiency in Machine Translation:

1. Experiments in neural machine translation showed that RMSNorm outperforms the baseline (without normalization) in terms of both convergence speed and test accuracy.

2. In particular, RMSNorm reduced training time by about 25% compared to LayerNorm in tasks using the RNNSearch model.

22.1.4 Partial RMSNorm (pRMSNorm):

1. The paper introduces pRMSNorm, a variant of RMSNorm that estimates normalization from only a portion of the inputs, which further reduces computation.

2. Although pRMSNorm can achieve additional speed-ups, the gains are dependent on implementation, with the model sometimes being slower due to non-optimal tensor slicing operations. Nonetheless, pRMSNorm reduces training time by around 33% to 40% in tasks like image-caption retrieval.

22.1.5 Applicability Across Different Architectures:

1. RMSNorm is versatile and can be applied to various neural network architectures, including RNNs, CNNs, and Transformers. This makes it an effective drop-in replacement for LayerNorm across different types of models.

2. For instance, in image-caption retrieval, RMSNorm accelerates training by 40.8% to 63.9% compared to LayerNorm while maintaining high recall scores.

22.1.6 Robustness and Stability:

1. RMSNorm maintains stability in training despite not re-centering the inputs. It stabilizes layer activations effectively, ensuring that the output remains consistent even with variations in input scaling.

2. Experiments also show that RMSNorm is more robust than LayerNorm when weight initialization is not optimal, further highlighting its stability during model training.

22.1.7 Lower Energy and Memory Costs:

1. By eliminating unnecessary computational steps, RMSNorm reduces energy consumption and memory usage, making it more efficient, especially in larger models with multiple layers.

22.2 Overall Improvements:

1. Speed-up of 7% to 64% in various tasks and models.

2. Comparable or improved performance in terms of accuracy and convergence, matching LayerNorm in most tasks.

3. pRMSNorm provides additional computational savings by using only a subset of inputs for normalization.

4. Robustness in handling weight initialization and stable training dynamics, ensuring broader applicability across different neural networks.

Chapter 23

BitNet: Scaling 1-bit Transformers for Large Language Models [23]

This paper introduces BitNet, a novel architecture designed to efficiently scale large language models (LLMs) by using 1-bit weights in transformers. The main problem addressed is the growing computational and energy costs associated with increasingly large LLMs, which make them difficult to deploy and environmentally taxing. Traditional approaches, such as post-training quantization and full-precision models (like FP16), still consume significant resources and can degrade in performance when heavily quantized.

The key innovation of BitNet is its ability to train models with 1-bit weights from scratch, significantly reducing the memory and computational footprint. Instead of using floating-point representations (FP16 or FP32), BitNet applies binarization, meaning model weights are represented with just +1 or -1 values. It also includes BitLinear, a custom matrix multiplication method designed to handle 1-bit weights efficiently. Despite this extreme quantization, BitNet retains high precision for critical operations such as gradients and optimizer states during training, ensuring model stability and accuracy.

23.1 Main Contributions

23.1.1 Reduced Energy and Memory Costs:

1. BitNet significantly reduces energy consumption and memory requirements compared to FP16 transformers, especially as model sizes scale up.

2. Energy savings increase with model size, offering up to 38.8 times lower energy consumption for large models (30B parameters) compared to FP16

counterparts.

23.1.2 Competitive Performance:

1. Despite the drastic reduction in precision, BitNet performs competitively in both language modeling and downstream tasks (e.g., Hellaswag, Winogrande, Winograd, StoryCloze), achieving accuracy similar to models using full-precision methods.

23.1.3 Efficient Scaling:

1. BitNet follows a similar scaling law to full-precision transformers, meaning it continues to scale efficiently as model size increases, without losing performance benefits.

23.1.4 Stability in Training:

1. BitNet demonstrates better training stability compared to FP16 models, allowing it to handle larger learning rates and faster convergence during training. This stability comes from techniques like LayerNorm and optimized weight quantization methods.

23.1.5 Group Quantization:

1. The paper also introduces Group Quantization, which improves model parallelism by splitting large weight matrices into smaller groups and independently estimating quantization parameters, reducing communication overhead in distributed training.

Chapter 24

DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection [24]

This paper introduces DeeperForensics-1.0, a large-scale dataset designed to improve real-world face forgery detection. With the rise of "deepfakes" and other face-swapping technologies, there is an increasing need for reliable detection systems to identify manipulated videos. Current datasets, while useful, often lack the scale, quality, and diversity required to effectively train detection models that can handle the complexities of real-world scenarios.

24.1 Main Contributions

24.1.1 Large-Scale Dataset:

DeeperForensics-1.0 is the largest face forgery detection dataset to date, comprising 60,000 videos with a total of 17.6 million frames. This is significantly larger than existing datasets, providing a more comprehensive set of data for training models.

24.1.2 Real-World Relevance:

The dataset incorporates a wide range of real-world perturbations, such as compression, noise, and blurring, which mimic conditions often found in videos uploaded to the internet. This makes the dataset more realistic and useful for developing models that can detect forgeries in real-world applications.

24.1.3 High-Quality Video Collection:

The authors carefully curated high-resolution videos of 100 paid actors with diverse appearances, under controlled lighting and camera setups, to ensure high-quality source material. These videos are then used to generate fake videos using a novel DeepFake Variational Auto-Encoder (DF-VAE), which improves the quality of the manipulated content.

24.1.4 Improved Detection:

DeeperForensics-1.0 includes not only manipulated videos but also a hidden test set of highly deceptive fake videos, specifically designed to challenge detection models. This test set contains videos that have fooled a majority of human participants in user studies, making it a robust tool for evaluating the effectiveness of detection models.

24.1.5 Benchmarking:

The authors also benchmark several existing face forgery detection methods on the DeeperForensics-1.0 dataset. The results show that while existing methods perform well on clean datasets, they struggle with the more challenging, real-world-like conditions presented in DeeperForensics-1.0. This highlights the need for further advancements in detection techniques.

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

[8] M. Zhu, K. Han, Y. Tang, and Y. Wang, "Visual transformer pruning," *CoRR*, vol. abs/2104.08500, 2021. [Online]. Available: https://arxiv.org/abs/2104.08500

[9] R.-J. Zhu, Y. Zhang, E. Sifferman, T. Sheaves, Y. Wang, D. Richmond, P. Zhou, and J. K. Eshraghian, "Scalable matmul-free language modeling," 2024. [Online]. Available: https://arxiv.org/abs/2406.02528

[10] R.-J. Zhu, Q. Zhao, G. Li, and J. K. Eshraghian, "Spikegpt: Generative pre-trained language model with spiking neural networks," 2024. [Online]. Available: https://arxiv.org/abs/2302.13939

[11] Y. J. Heo, Y. J. Choi, Y. Lee, and B. Kim, "Deepfake detection scheme based on vision transformer and distillation," *CoRR*, vol. abs/2104.01353, 2021. [Online]. Available: https://arxiv.org/abs/2104.01353

[12] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *CoRR*, vol. abs/1812.08685, 2018. [Online]. Available: http://arxiv.org/abs/1812.08685

[13] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[14] R. Tolosana, S. Romero-Tapiador, J. Fiérrez, and R. Vera-Rodríguez, "Deepfakes evolution: Analysis of facial regions and fake detection performance," *CoRR*, vol. abs/2004.07532, 2020. [Online]. Available: https://arxiv.org/abs/2004.07532

[15] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," *CoRR*, vol. abs/1811.00661, 2018. [Online]. Available: http://arxiv.org/abs/1811.00661

[16] V. L. L. Thing, "Deepfake detection with deep learning: Convolutional neural networks versus transformers," 2023. [Online]. Available: https://arxiv.org/abs/2304.03698

[17] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, p. 1499–1503, Oct. 2016. [Online]. Available: http://dx.doi.org/10.1109/LSP.2016.2603342

[18] W. Huang, Y. Liu, H. Qin, Y. Li, S. Zhang, X. Liu, M. Magno, and X. Qi, "Billm: Pushing the limit of post-training quantization for llms," 2024. [Online]. Available: https://arxiv.org/abs/2402.04291

[19] Z. Li, X. Liu, J. Zhang, and Q. Gu, "Repquant: Towards accurate post-training quantization of large transformer models via scale reparameterization," 2024. [Online]. Available: https://arxiv.org/abs/2402.05628

[20] Z. Liu, Y. Wang, K. Han, S. Ma, and W. Gao, "Post-training quantization for vision transformer," *CoRR*, vol. abs/2106.14156, 2021. [Online]. Available: https://arxiv.org/abs/2106.14156

[21] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei, "The era of 1-bit llms: All large language models are in 1.58 bits," 2024. [Online]. Available: https://arxiv.org/abs/2402.17764

[22] B. Zhang and R. Sennrich, "Root mean square layer normalization," 2019. [Online]. Available: https://arxiv.org/abs/1910.07467

[23] H. Wang, S. Ma, L. Dong, S. Huang, H. Wang, L. Ma, F. Yang, R. Wang, Y. Wu, and F. Wei, "Bitnet: Scaling 1-bit transformers for large language models," 2023. [Online]. Available: https://arxiv.org/abs/2310.11453

[24] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," 2020. [Online]. Available: https://arxiv.org/abs/2001.03024