Literature Review for Ingur Thesis References

Reza

August 22, 2024

Contents

1 BinaryViT [1] 3
- 1.1 Introduction . . . 3
- 1.2 Proposed Model . . . 3
  - 1.2.1 Global Average Pooling Layer . . . 3
  - 1.2.2 Multiple Pooling Branches . . . 4
  - 1.2.3 Affine Transformation Before Residual Connections . . . 4
  - 1.2.4 Pyramid Structure . . . 4
  - 1.2.5 Binary Fully-Connected Layers with Enhanced Attention 5
  - 1.2.6 Distillation from Full-Precision Models . . . 5
- 1.3 Impact of the Changes . . . 5
- 1.4 Results and Improvements . . . 6
  - 1.4.1 Performance Improvement on ImageNet-1k . . . 6
  - 1.4.2 Efficiency in Terms of Operations and Parameters . . . 6
  - 1.4.3 Comparisons with State-of-the-Art (SOTA) Binary Models 7
  - 1.4.4 Impact of Architectural Enhancements . . . 7
  - 1.4.5 Reduction in Computational Complexity . . . 7
- 1.5 Overall Improvements . . . 8

2 Vision Transformer for Small-Size Datasets [2] 9
- 2.1 Shifted Patch Tokenization (SPT) . . . 9
  - 2.1.1 Previous Approach: . . . 9
  - 2.1.2 Proposed Change: . . . 10
- 2.2 Locality Self-Attention (LSA) . . . 10
  - 2.2.1 Previous Approach: . . . 10
  - 2.2.2 Proposed Change: . . . 10
- 2.3 Comparison to Other Data-Efficient ViTs . . . 11
- 2.4 Efficiency vs. Performance Trade-offs . . . 12
  - 2.4.1 Previous Models: . . . 12
  - 2.4.2 Proposed Model: . . . 12
- 2.5 Performance Gains . . . 12
- 2.6 Overall Impact of the Proposed Changes . . . 12
- 2.7 Results and Improvements . . . 13
  - 2.7.1 Performance Improvements on Small Datasets . . . 13
  - 2.7.2 Improvements in ImageNet Performance . . . 13

2.7.3 Efficiency and Computational Overhead 13
2.7.4 Ablation Study Results 14
2.7.5 Qualitative Improvements 14
2.7.6 Comparison with State-of-the-Art (SOTA) Models 14
2.8 Key Takeaways: 15

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

title

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