# Literature Review for Ingur Thesis References

Reza

August 22, 2024

# BinaryViT [1]

## 0.1 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.

## 0.2 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:

### 0.2.1 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.

### 0.2.2 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.

### 0.2.3 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.

### 0.2.4 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.

### 0.2.5 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.

### 0.2.6 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.

## 0.3 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.

## 0.4 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:

### 0.4.1 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.

### 0.4.2 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 $0.79 \times 10^8$ operations compared to ReActNet's $1.93 \times 10^8$ 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 $0.19 \times 10^8$ FLOPs, much lower than ReActNet and other competing models, further highlighting its efficiency.

### 0.4.3 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 $1.93 \times 10^8$ operations and 21.8 million parameters.
2. BinaryViT: Matched ReActNet's accuracy of 67.7%, but with significantly fewer operations ($0.79 \times 10^8$ vs. $1.93 \times 10^8$) 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.

### 0.4.4 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%**.

### 0.4.5 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.

## 0.5 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.

# 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.