Generalization by Visual Attention?

CNNs vs Transformers on out-of-distribution performance

Author
Baptiste Colle

Supervisor
Wendelin Böhmer

Want to learn more? Read the paper!

01 Background Information

Out-of-distribution

This research investigates if a neural network trained on specific distribution can generalize its world's understanding in a new distribution. We call this difference between training and test environment "out-of-distribution".

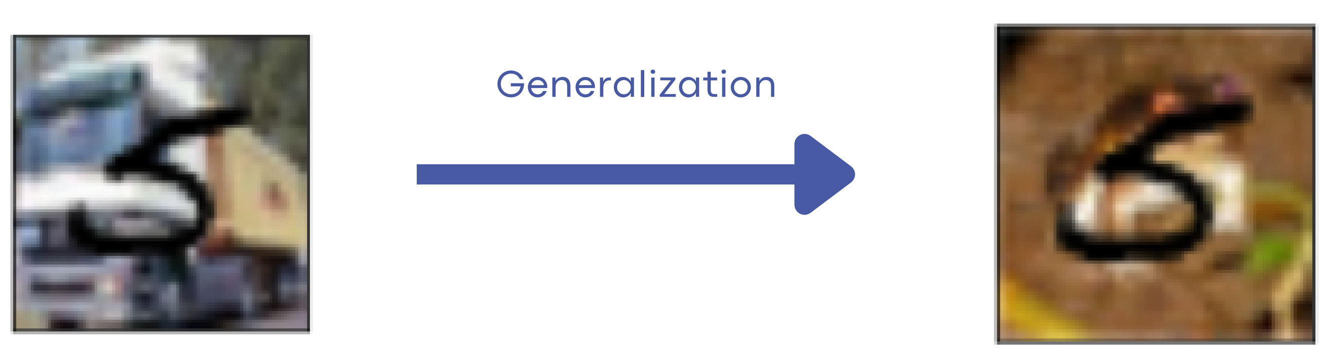

Figure 1: Out-of-distribution

To illustrate our statement, as we can observe in figure 1 we wondered whether a network that has only seen the number "5" associated with trucks would still recognize the number "5" in a new environment like on top of a flower.

Convolutional neural network (CNN)

The CNN architecture is the most common architecture in computer vision. It works by extracting features based on its kernel, which are a set of learnable weights. This kernel is then rolled over the image to produce an output, which allows the network to extract the meaningful feature out of the image.

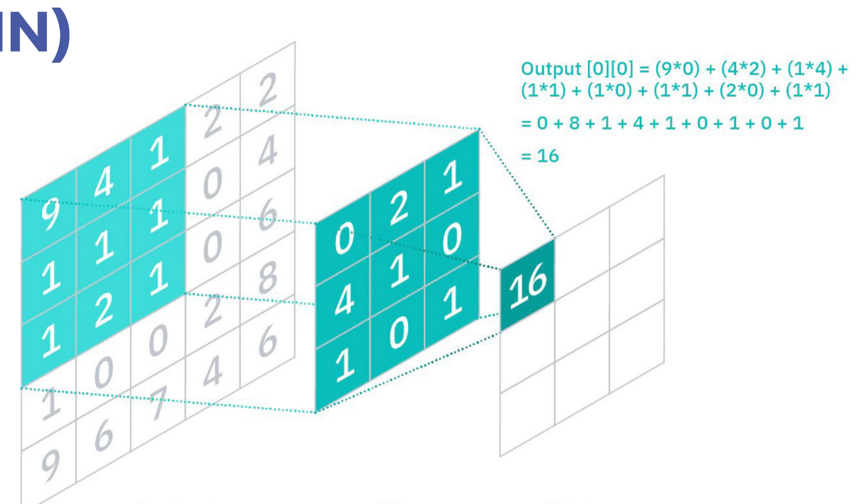

Figure 3: CNN Architecture [2]

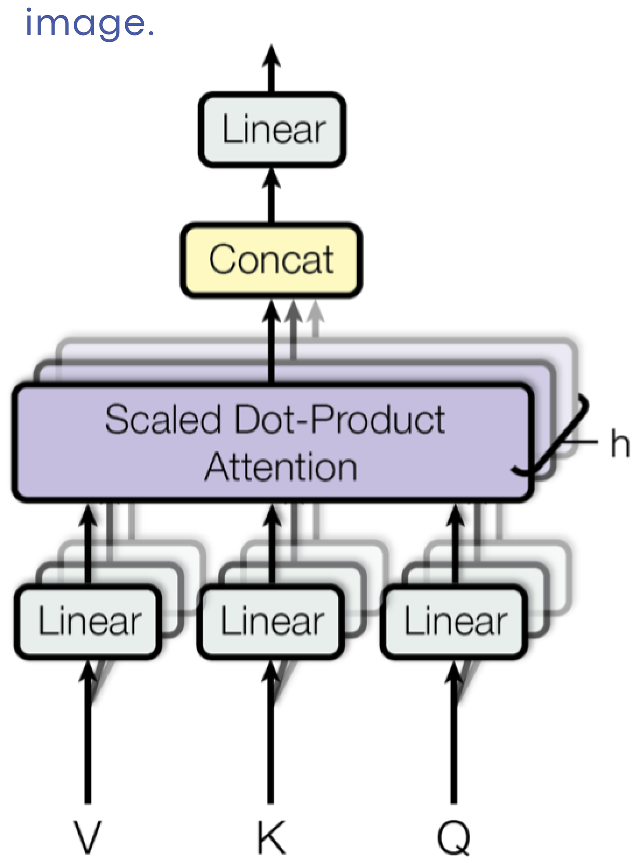

Figure 2: MHA Architecture [1]

Transformer

The transformer is a novel alternative to CNNs for computer vision. Instead of being restricted by the association allowed inside the kernel, it can exploit the full input to make connections by using its attention mechanism (see figure 2). This attention is then duplicated and encapsulated into a single module that we call **multi-head attention** (MHA). We believed that the attention mechanism of the transformer should lead to better performance on out-of-distribution.

02 Objective

Research Question

Which network configurations have the largest impact on out-of-distribution performance in both architectures?

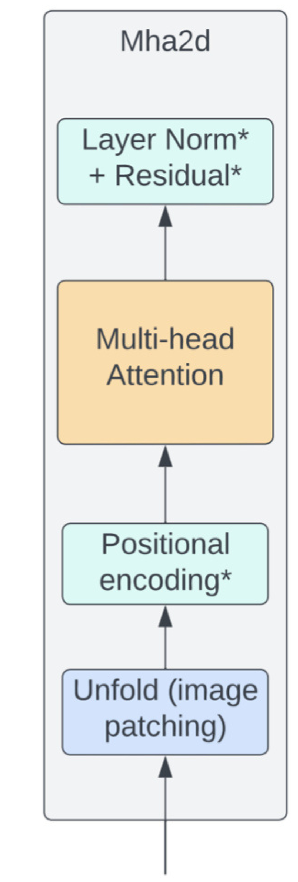

Figure 4: Custom mha2d module equivalent to conv2d

03 Methodology

To investigate this question, we started by creating datasets with a customisable number of background per digit. I then implemented a custom module fully interchangeable with a convolution operation, named **Mha2d**. Lastly, I tested different network configuration for both models to see what configuration can improve out-of-distribution performance.

04 Analysis

1. Baseline

I used the LeNet [3] architecture as a baseline comparison between both architecture by swapping the convolutional blocks with the Mha2d module. As we can see in figure 5, our multi-head attention based model performed significantly better than CNNs for out-of-distribution

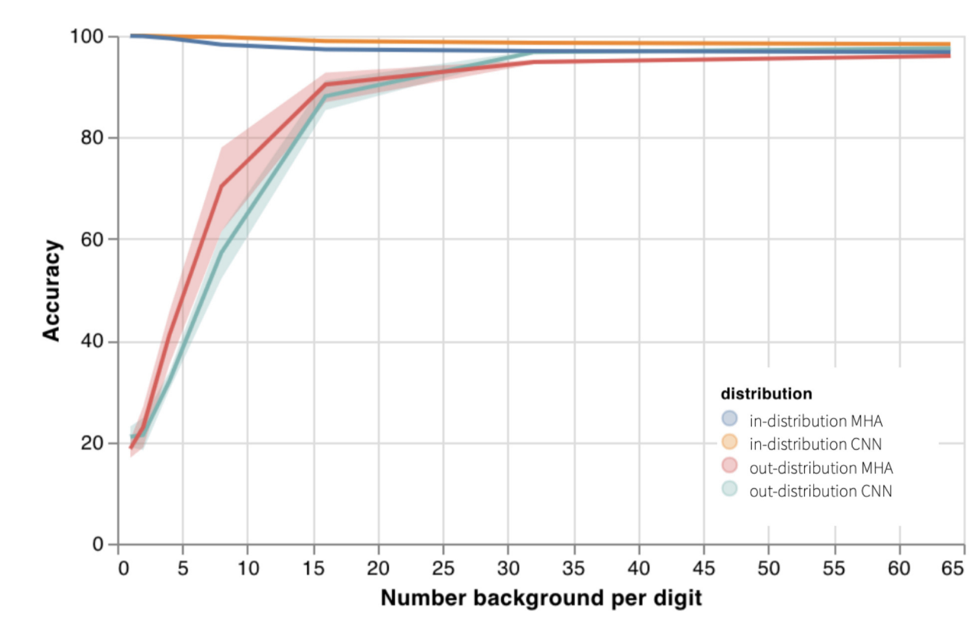

Figure 5: Baseline Performance

2. Network depth

In this experiment, I tested the impact of numbers of convolutional or attention layers on out-of-distribution performance. We can see in figure 6 that with a higher number of layers the performance collapses, as the networks are not able to learn. However, CNNs are better able to train with more layer compared to MHA.

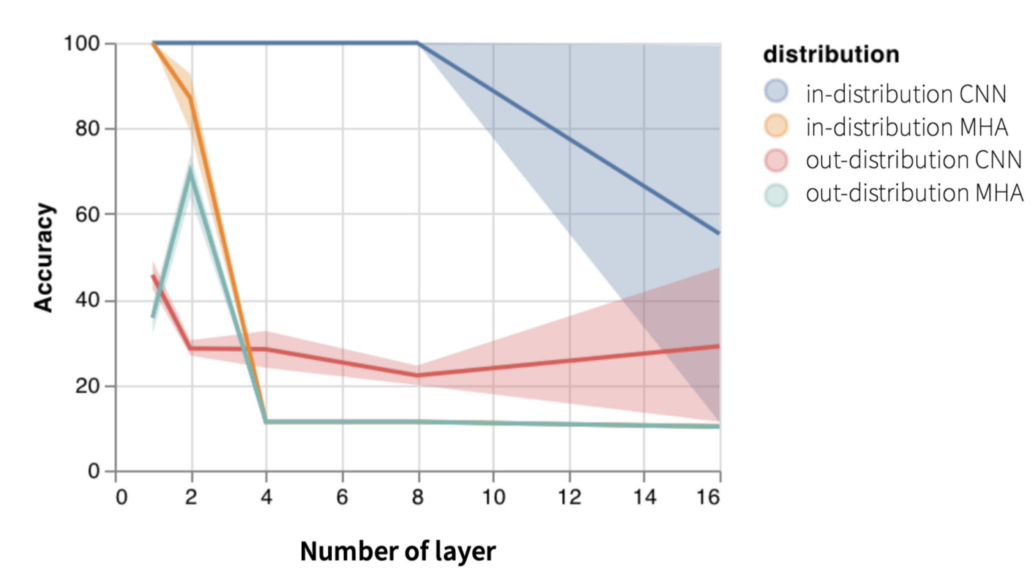

Figure 6: Accuracy of both models according to network depth

3. Number of heads

The multi-head attention is controlled by a hyperparameter called the number of heads. This determines how many focal points (attention) the network has. We observed that for our task this parameter does not influence performance.

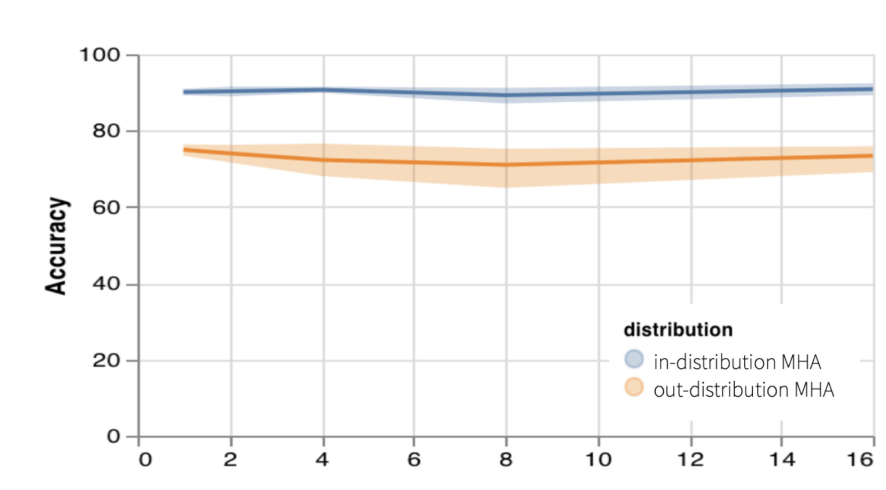

Figure 7: Accuracy of MHA based on the number of heads

Bibliography

- [1] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 2017-December, 2017.
- [2] IBM Cloud Education, What are Convolutional Neural Networks? Oct. 20. [Online]. Available: https://www. ibm.com/cloud/learn/convolutional-neural-networks.
- [3] Y. LeCun, B. Boser, J. S. Denker, et al., "Backpropagation Applied to Handwritten Zip Code Recognition," Neural Computation, vol. 1, no. 4, 1989, ISSN: 08997667. DOI: 10.1162/neco.1989.1.4.541.

4. Transformer specific component

This paper further investigated the below mentioned transformer-specific elements to see if they had an impact on out-of-distribution performance:

- Positional Encoding
  - This element adds locality information to the attention head.
- Layer Normalization
  - A regularisation technique used to normalize the output of a layer.
- Residual Connection
  - A connection between current and deeper layers, that allow us to skip these and add the input directly to a deeper layer.

After experimenting, I then came to the conclusion that only the layer normalization was shown to significantly improve out-of-distribution performance.

05 Conclusion

	CNN	Transformer
Baseline		
Network Depth		
Number of Heads	X	
Positional Encoding	X	
Layer Normalization	X	
Residual Connection	X	
Conclusion		

Recommendation for further research

Firstly, it would be interesting to investigate out-of-distribution performance with images of bigger size (our experiment is only with 32x32 images). Secondly, stabilizing learning for depth network and studying the impact of residual connection on depth network could lead to more conclusions on the impact of network depth.