Generalization and Data Transformation Invariance of Visual Attention Models

Pepijn de Kruijff
Supervisor: Wendelin Böhmer
EEMC, Delft University of Technology, The Netherlands

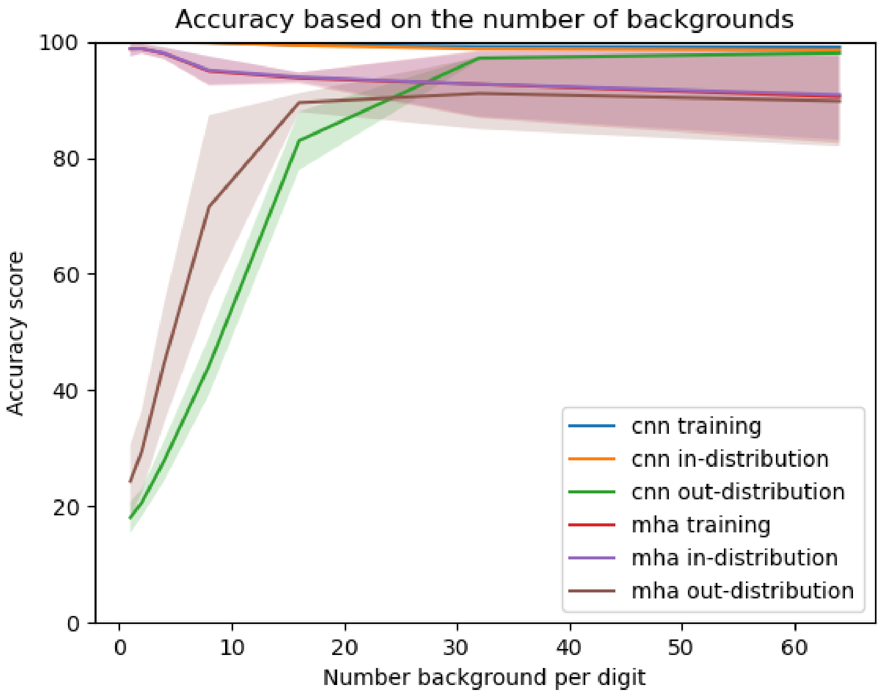
Introduction

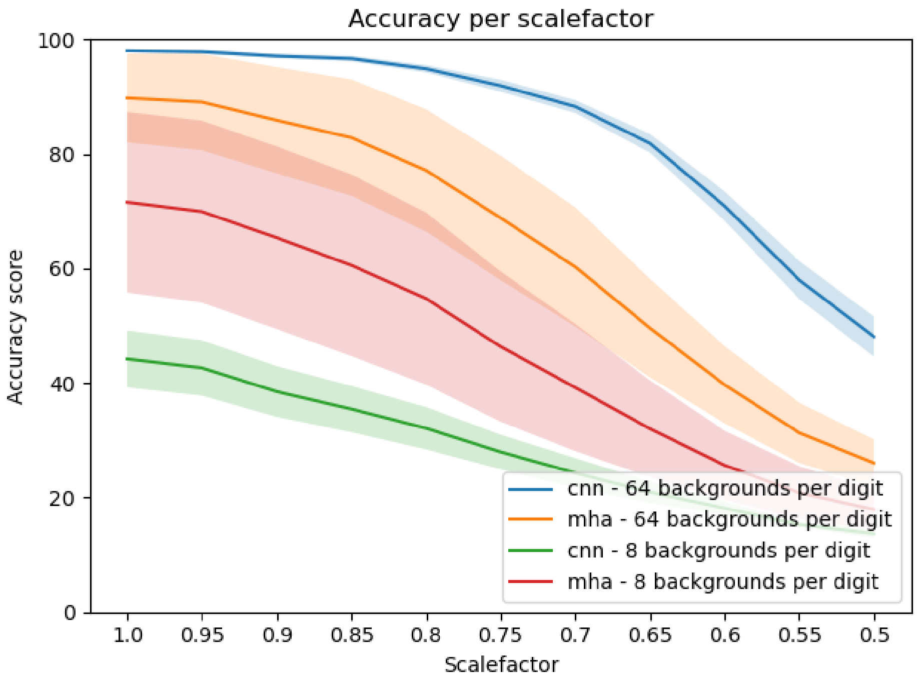
The current standard for image classification tasks are Convolutional Neural Networks (CNNs). Modern CNNs can efficiently classify MNIST digits with about 99.9 percent accuracy but are unable to classify out-of-distribution data well.

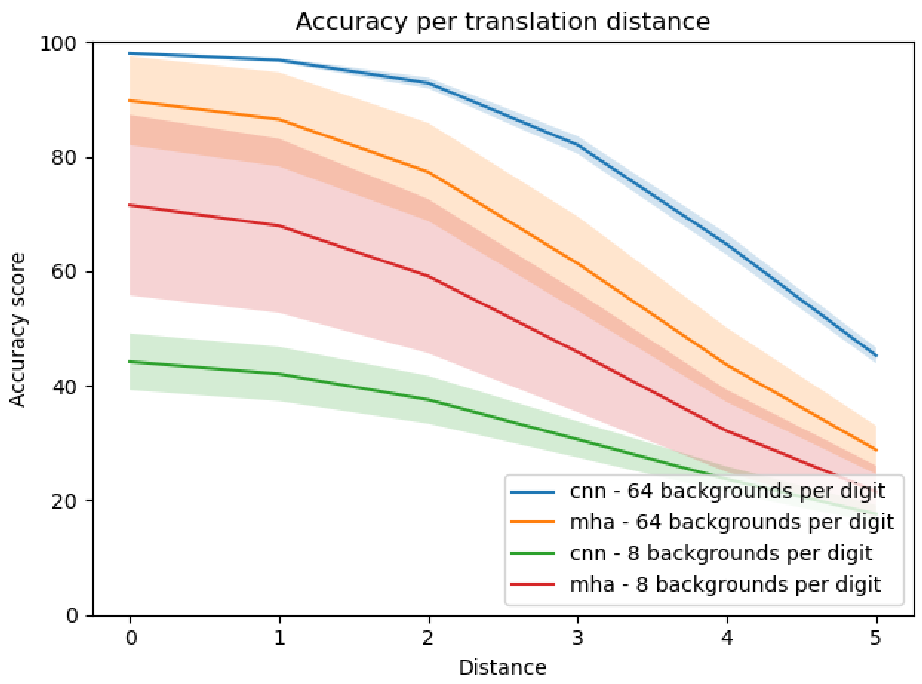
A relatively new architecture using Multi-Head Attention (MHA) layers has been introduced. This paper investigates how the following properties of MHA layers compare to those of CNNs:

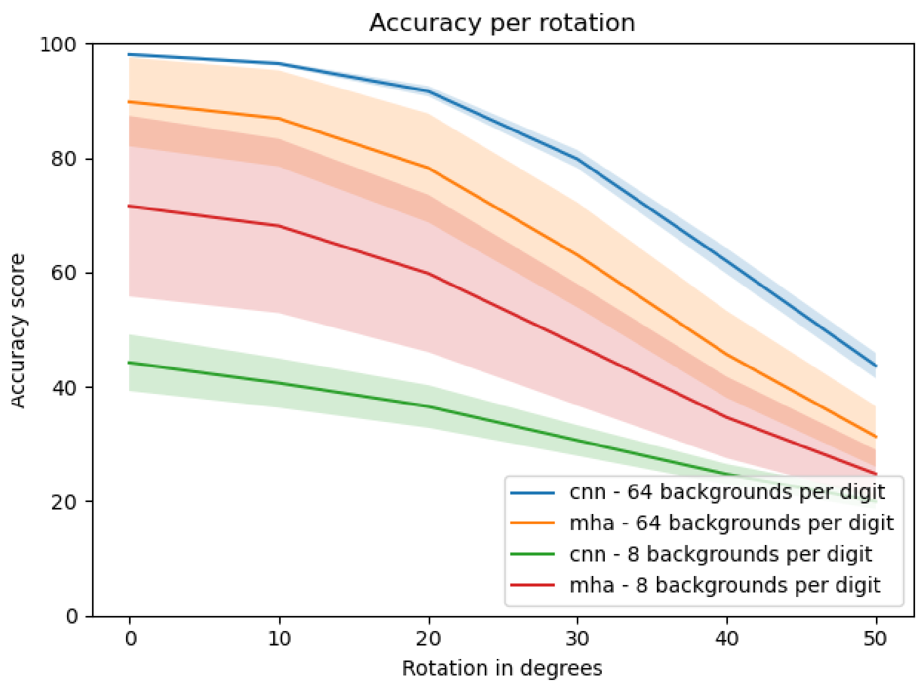
1. Are MHAs are better able to generalize on out-of-distribution data than CNNs?
2. How invariant are MHAs to affine data transformations?

Results & Conclusions





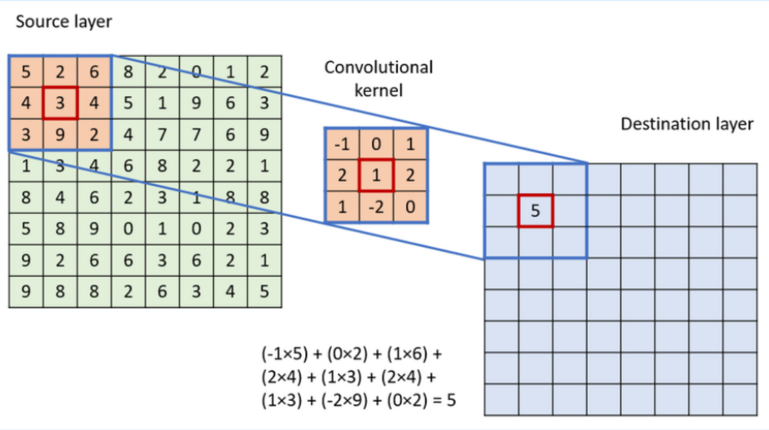




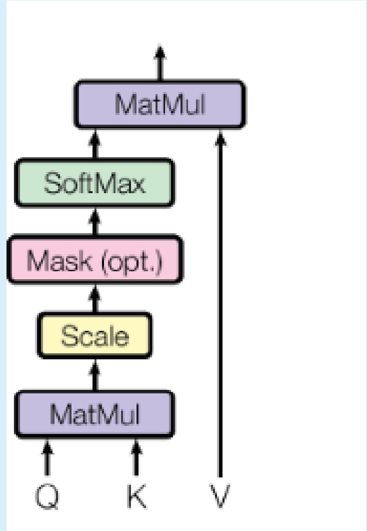
- The MHA does perform better when the model is trained with more images per digit, suggesting that it is better able to generalize to unseen data.
- The CNN performs better for higher amounts of backgrounds per image.
- The CNNs are more invariant to affine transformations than MHAs.

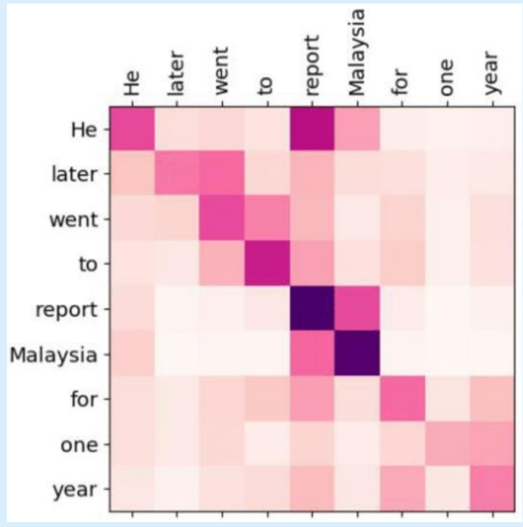
Scientific Background

CNNs are a sparsely connected neural networks that make use of the convolution operation.



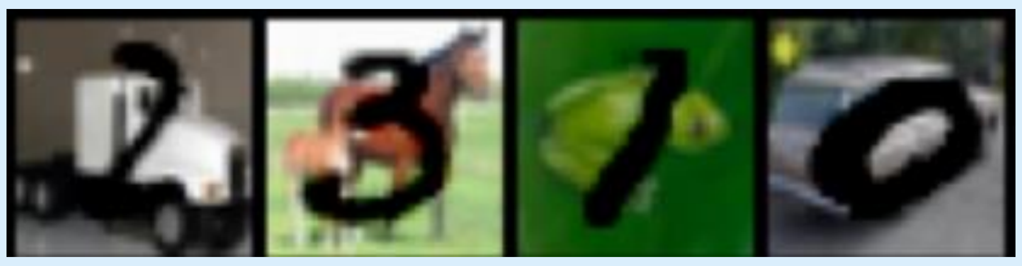
MHAs are able to use an attention mechanism to pay attention to several parts of the image. This is done by calculating the attention between input patches.





Methodology

Dataset is constructed using MNIST and CIFAR-10 images. The models are trained with a variable amount of backgrounds per digit. The generalizability of both models is compared by testing them on an out-of-distribution sample: one with a background not seen during training.



The following Transformations have been applied to the digits in the test-set:

1. Translation. In a random direction by a variable amount of pixels.
2. Rotation. Average of clockwise and counterclockwise rotation.
3. Scaling.