ON THE REGULARIZATION OF CNNS AND TRANSFORMERS UNDER DISTRIBUTION SHIFTS

Author: Leo Zi-You Assini
Email: l.z.assini-1@student.tudelft.nl
Supervisor: Wendelin Böhmer

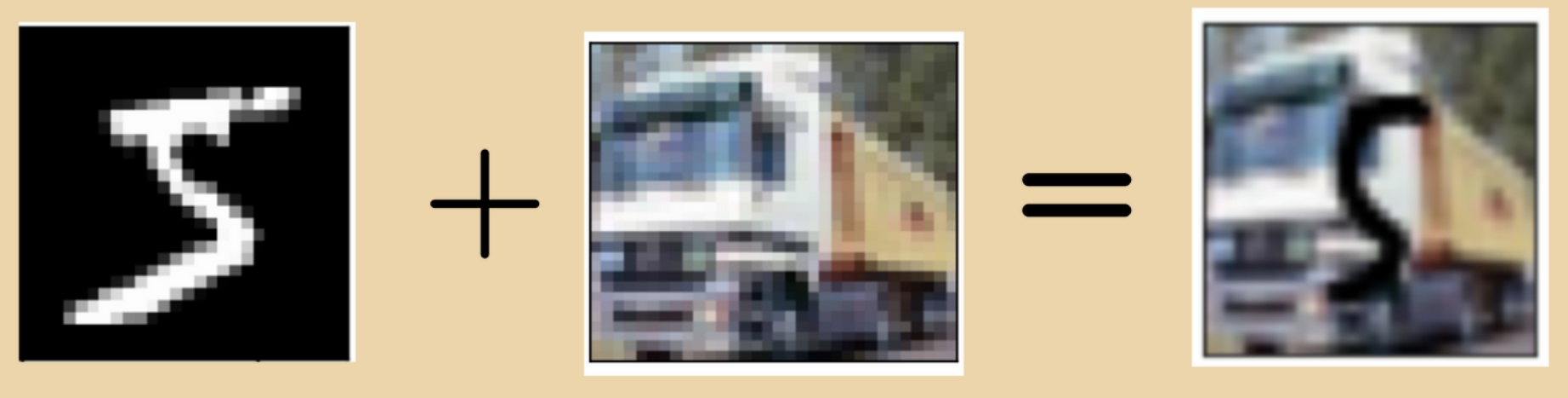
TU Delft Delft University of Technology

BACKGROUND

- Image classification is a hugely important area of research which has given rise to many applications, ranging from the detection of diseases in medical images to self-driving cars.
- Great success has been achieved using models such as Convolutional Neural Networks.
- Accuracy drops rapidly when the test images have the same type of information as the training data but with systematic modifications, known as distribution shift. We say that these images are out-of-distribution.
- Regularization is any modification made to a learning algorithm to reduce its generalization error.

RESEARCH OBJECTIVES

- Does an unregularized MHA perform better than an unregularized CNN under distribution shift?
- Do regularization schemes, with tuned hyperparameters, improve the accuracy of the CNN and MHA architectures under distribution shift?
- Does a regularized MHA have a better accuracy than a regularized CNN under distribution shift?



DATASET USED

The dataset was a MNIST dataset with CIFAR10 images used as backgrounds

EXPERIMENTAL SETUP

The architectures were tested by applying each regularization scheme in isolation and then combining best results for each. The experiments were run on a number of different background images, ranging from 1 to 64.

The Regularization schemes applied to both architectures

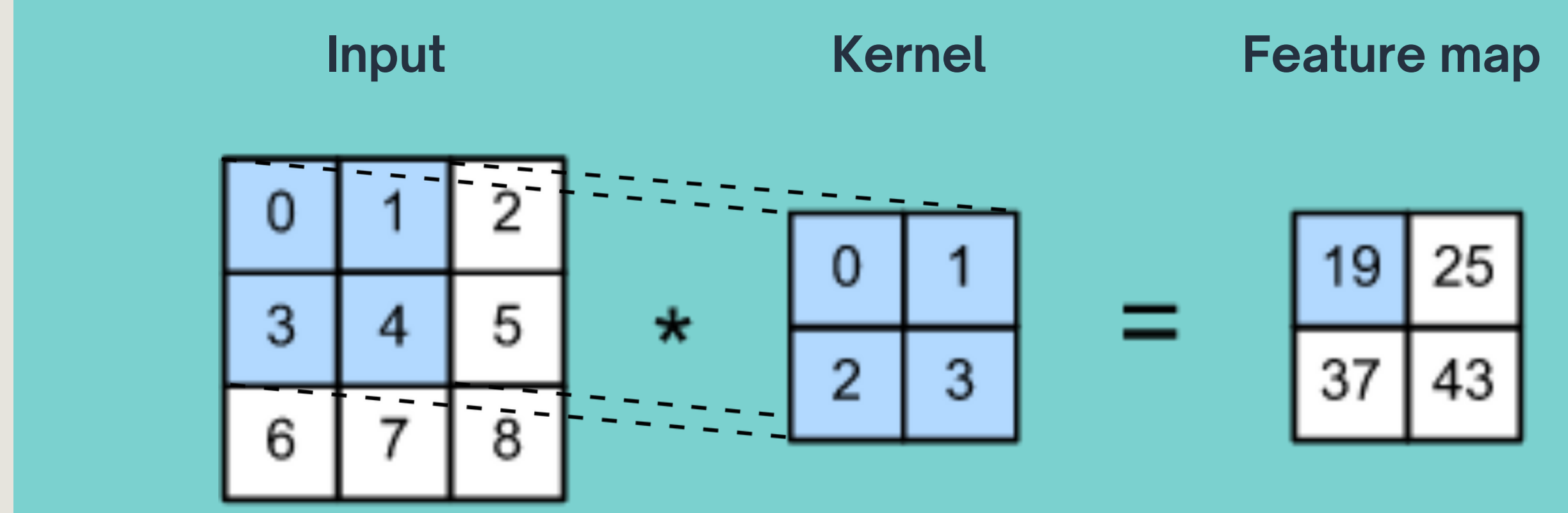
- Normalization
- Dropout
- L2 Norm

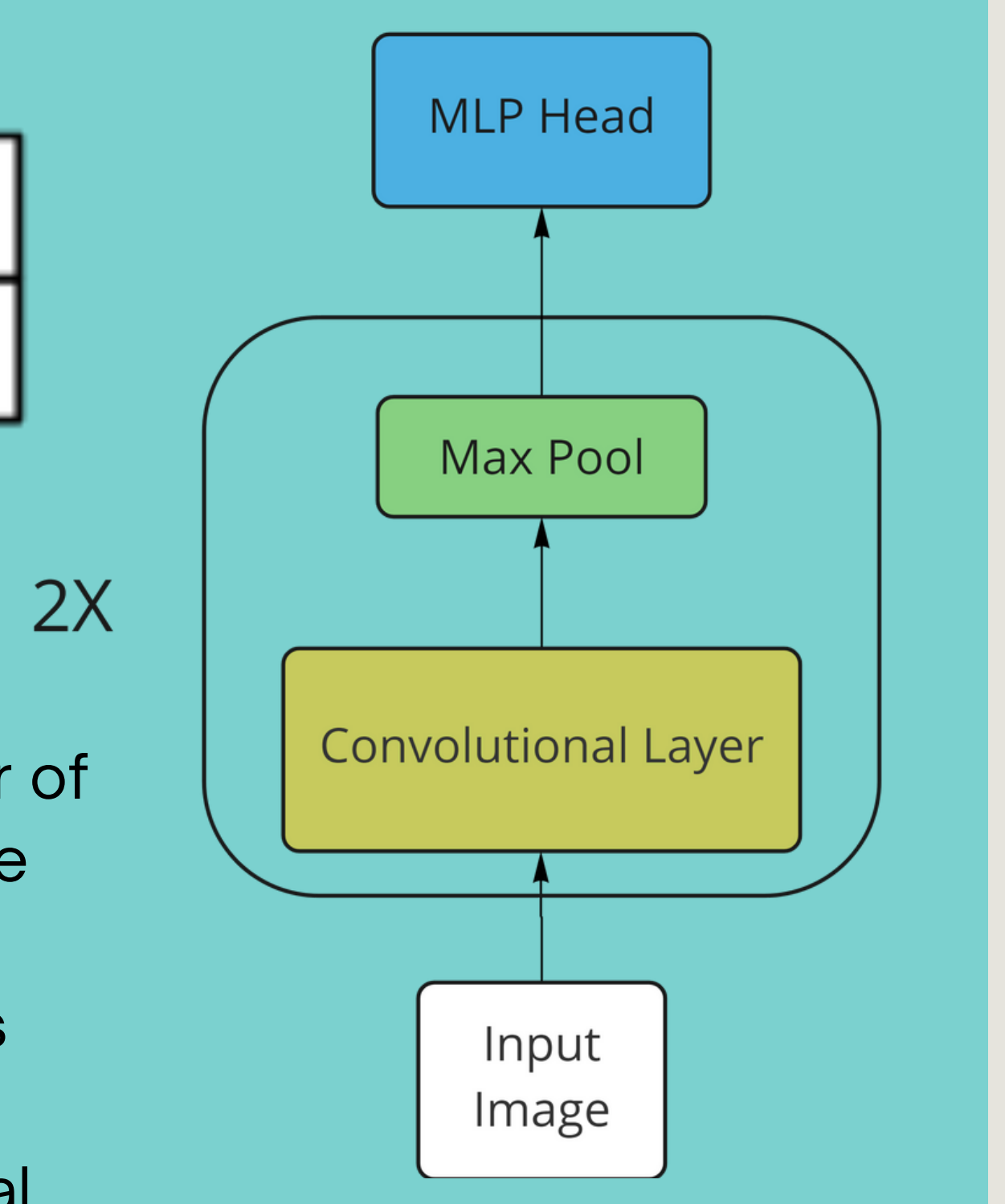
Additional schemes applied to the multi-head attention architecture

- Positional encoding
- Disagreement regularization

The architectures on the left are the baseline architectures to which regularization will be applied.

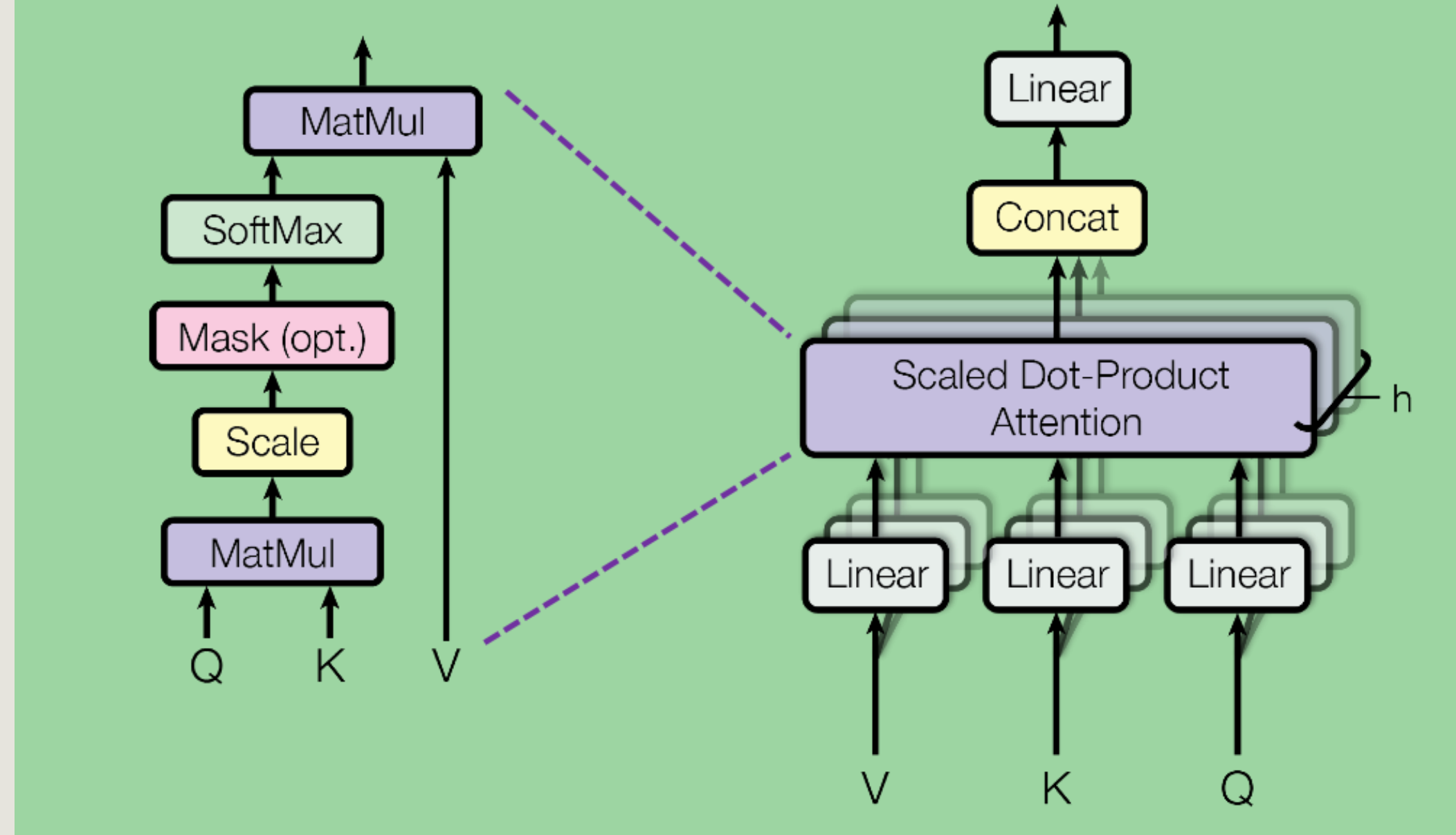
CONVOLUTIONAL NEURAL NETWORK (CNN)

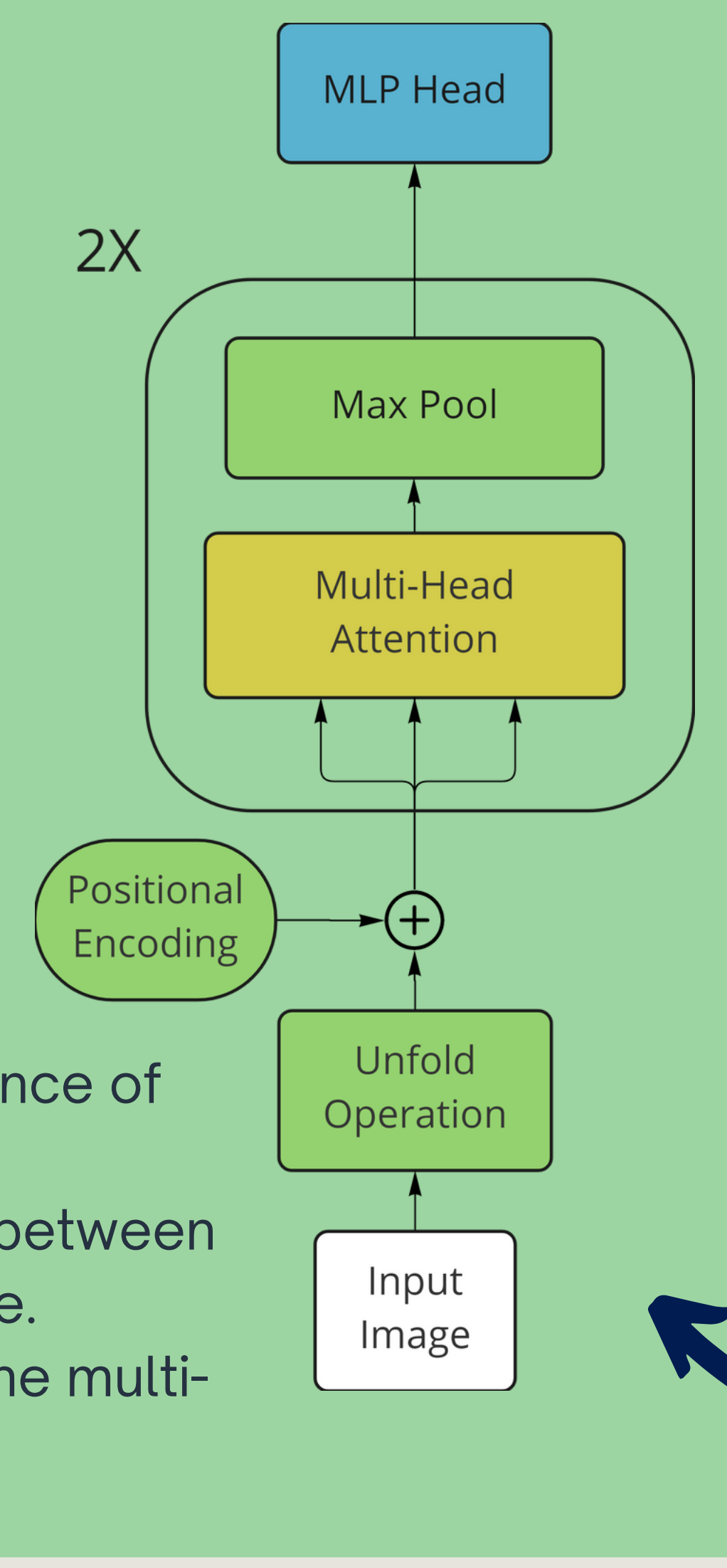




The convolutional layer places a kernel in the top left corner of the image, computes the weighted sum and then shifts the kernel to the right.
This operation is done row by row until the feature map is completed.
A pooling layer replaces areas of the input with a statistical summary. For max pool this is the maximal value.

MULTI-HEAD ATTENTION (MHA)

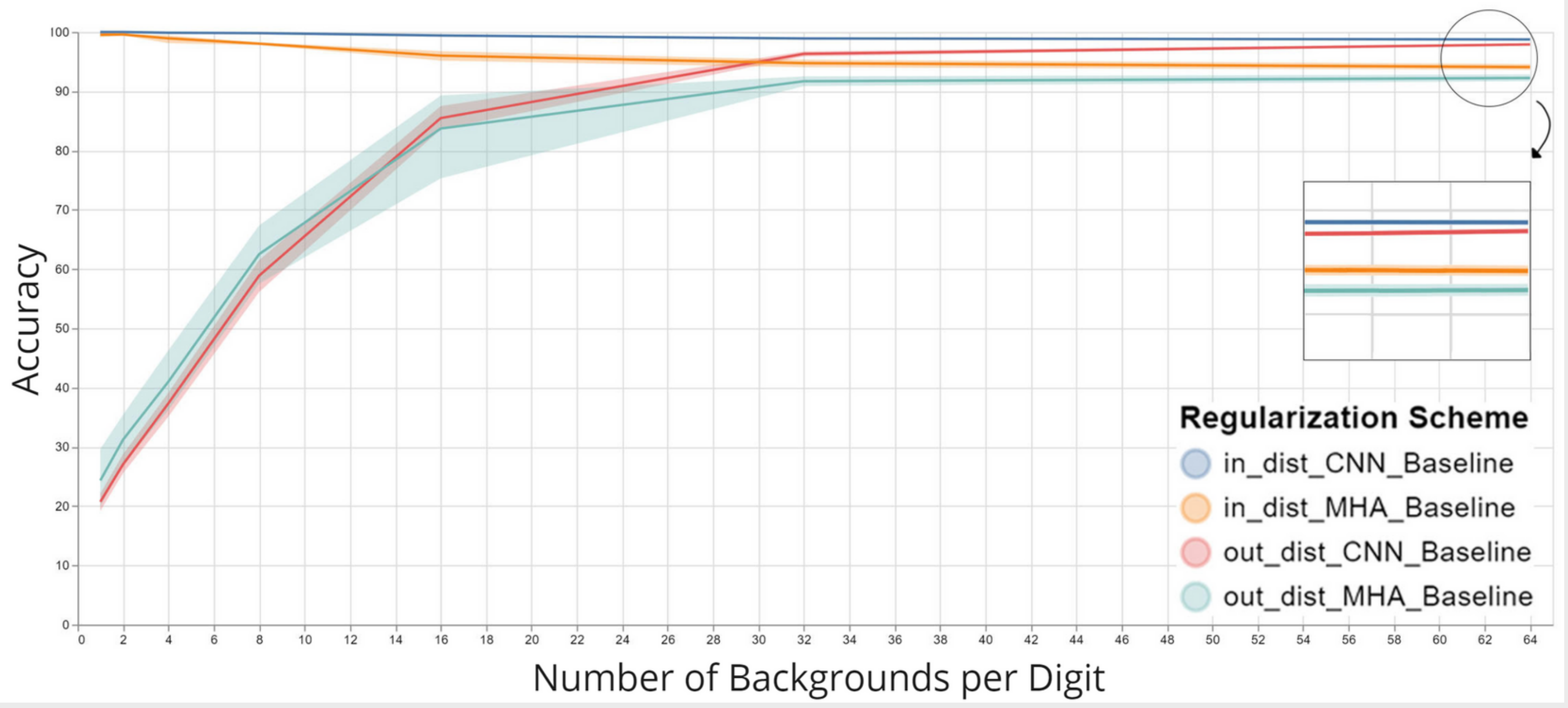




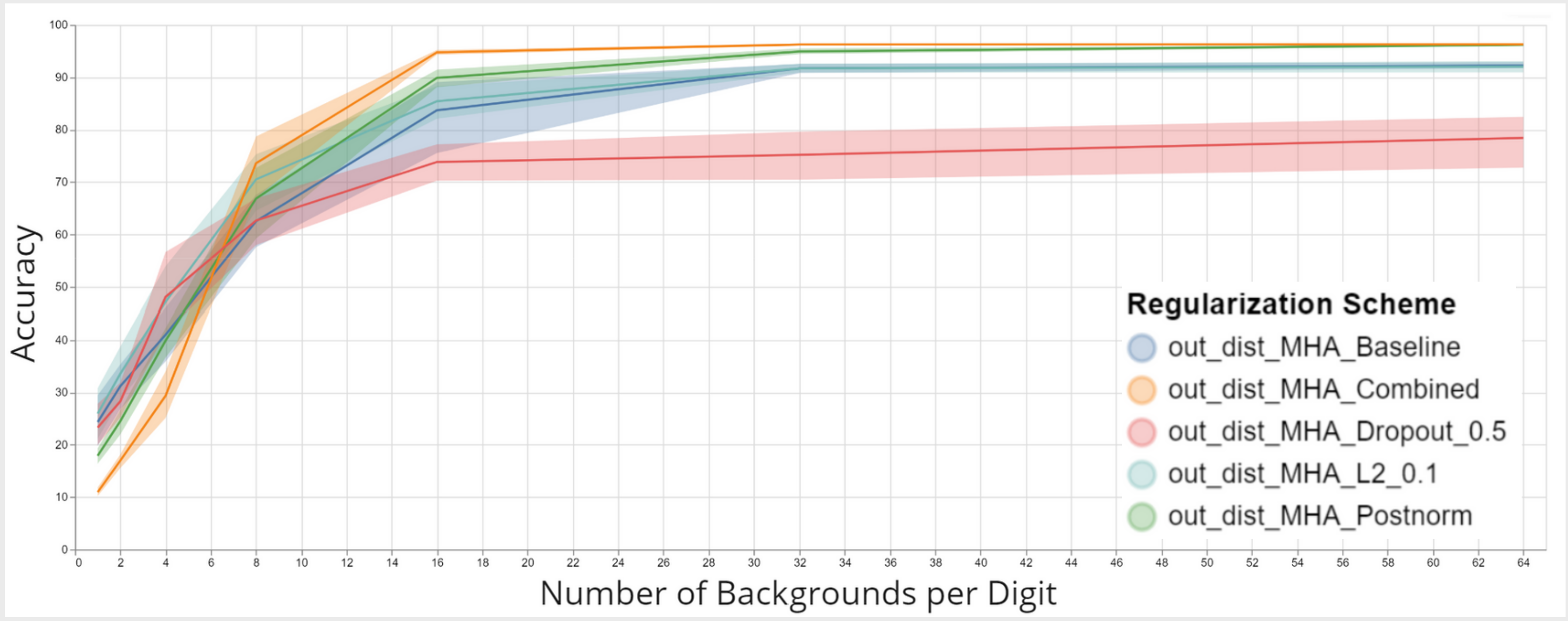
The Multi-Head Attention layer takes as input a sequence of elements, tokens.
These tokens modified to take in account the similarity between each token and every other token in the sequence.
The figure above shows the computations occuring in the multi-head attention module.

RESULTS

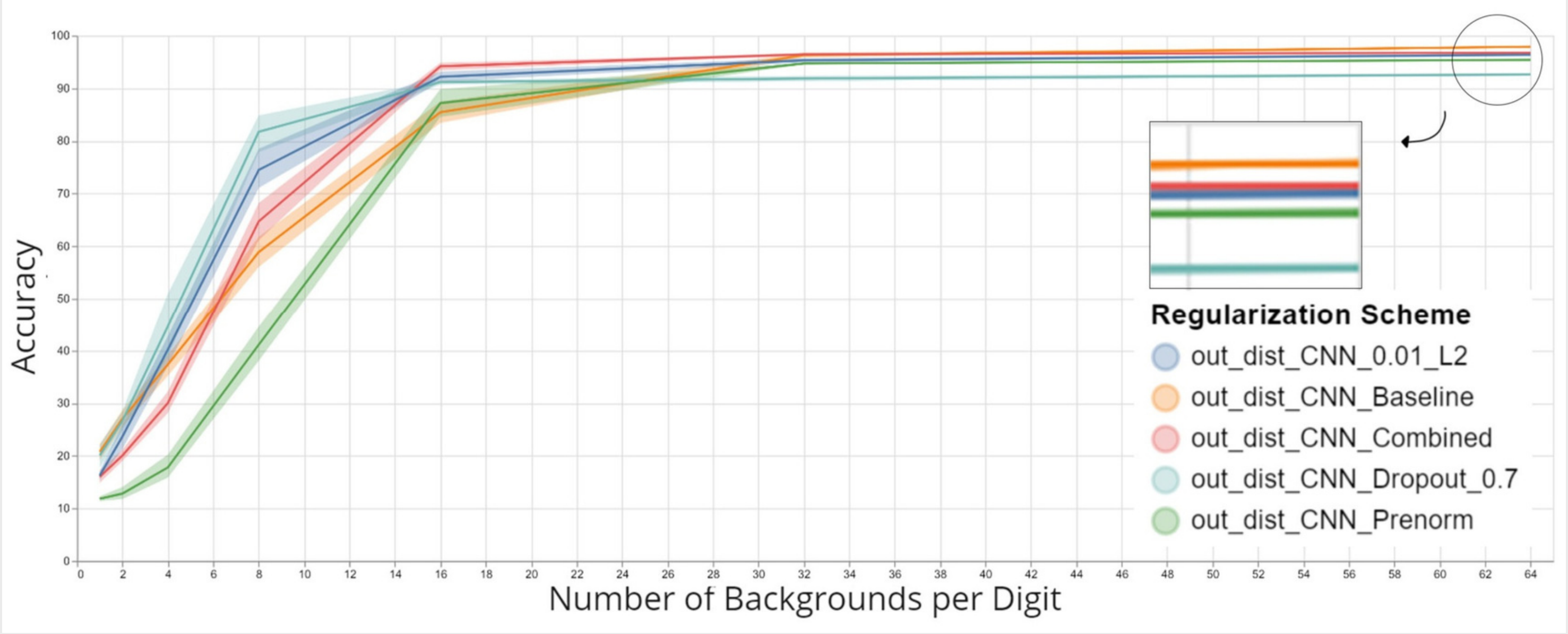
BASELINE ARCHITECTURES COMPARED



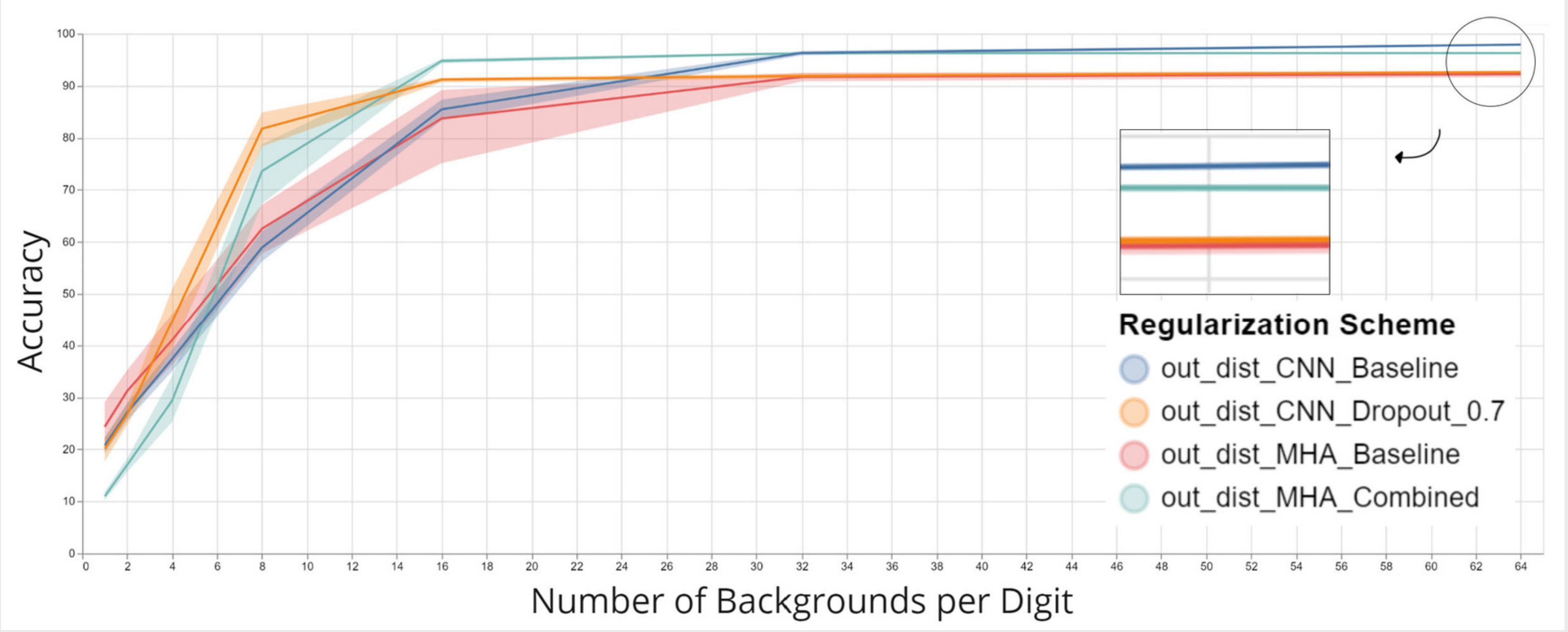
REGULARIZATION APPLIED TO MHA



REGULARIZATION APPLIED TO CNN



BEST RESULTS COMPARED



CONCLUSION

- Regularization is an effective tool for improving the generalization capabilities of neural networks, significantly increasing out-of-distribution accuracy.
- An important trade off is that sometimes in-distribution accuracy is decreased.
- No clear winner between CNN and MHA with both regularized architectures showing similar final performance.
- MHA out-of-distribution accuracy is improved only when regularization schemes are combined.

FUTURE WORK

- Using much larger data sets, for example ImageNet.
- Using other regularization techniques such as drophead, specific to MHAs, or the many different positional encodings that have been proposed for Transformers.
- Using different combinations of regularization techniques.