Reg. No: 2022/45/N/ST6/02222; Principal Investigator: mgr Sebastian Dariusz Jaszczur

Conditional Computation in Deep Neural Networks

Deep Neural Networks see continued increase in their usage in many real world applications. Particular impactful uses, such as image recognition and machine translation, are significant for a lot of individuals, but those machine learning models also require a significant amount of computation during both training and inference. This stems primarily from the significant size of those models, and current neural architectures.

The main hypothesis of this project is that current designs of neural network architectures use the computation inefficiently. It should be possible to improve some metrics of the neural network (such as training time, computation time during inference, quality of model's predictions, or similar) by dynamic, conditional computation in a neural network. This could lead to improvement of model's predictions given constant computation budget or faster evaluation with the same quality of predictions.

Intuitively, we can understand the idea of this project by imagining a particular problem and investigating how current neural networks operate. Assume we want to train a neural network to recognize cats and dogs on images - and more specifically, it should recognize what is a specific breed of a cat or a dog on a given image. We can separate three tasks that the model has to do: (1) recognize if an image is a cat or a dog, (2) recognize the specific breed of a cat, and (3) recognize the specific breed of a dog. Notice that task (1) is more general and seems intuitively easier than the other two. Moreover, if the neural network already solved task (1), it doesn't have to perform computation for both task (2) and (3) - depending on the answer, only one of those is relevant, and the other can be skipped.

Unfortunately, current neural networks architectures cannot compute things conditionally - they will run computation necessary for task (3) even if it already knows the image doesn't contain a dog at all. In this project, we investigate possible ways to give neural network an option to skip some computation, and train it to use this option efficiently.

This research may enable neural networks to train and run significantly faster and on cheaper hardware, without impacting quality of their predictions. This in turn would make further research and applications in the field of neural networks, and in natural language processing in particular, more accessible to everyone.