Coordinate based discriminative image model (Coordinet?)

Jeremy Gillen

Idea:

A recent paper (Anokhin et al., 2020) has shown that it is possible to create a powerful generative model for images, comparable to state of the art models, without using convolutions. The model takes a random latent vector, and a pixel coordinate, and generates a single pixel from the image. Repeating for each pixel, an impressively realistic image is generated. This suggests that the model, using just pixel coordinates, contains similar inductive biases to those encoded by convolution, or contains different inductive biases that are just as useful for generative modelling. I propose investigating whether similar inductive biases can be embedded in a discriminative model that classifies images without using any convolutional layers. My initial idea for an architecture will have an encoder and decoder network, with the encoder network mapping (single pixel, coordinate) -> (latent), all latent vectors summed up into one latent vector, and passed to the decoder network that maps (latent) -> (softmax output).

The discriminative model will be compared with simple fully connected and convolutional architectures and popular out-of-the-box convolutional architectures, in the hope that it will generalize better than fully connected networks, and be comparable (but probably a little worse) to the CNNs. Architectural variations will be investigated, including the use of multiplicative conditioning and different positional encodings, as in (Anokhin et al., 2020).

The next step in this investigation will be to make a comparison between the inductive biases of convolutional architectures, fully connected architectures, and the coordinate-FCN architecture. This will be conducted by comparing the test error of each of these architectures, for various small dataset sizes on a variety of image datasets. As a stretch goal, further investigation can be done by investigating the precise differences between the inductive biases of convolutional layers and coordinate-FCN architecture, by comparing robustness to various input transformations on toy datasets (shift, rotation, scaling, noise, sheer, nonlinear warping).

Importance:

As in (Anokhin et al., 2020), processing pixels individually with their coordinates has the potential to be useful for encoding different types of prior information. With various positional encodings, images with a wide variety of topological shapes can be easily and accurately represented, including spherical or cylindrical images. Another application is encoding prior information that an image is warped in some way, like in the case of fish-eye lenses. Other potential benefits of this architectures include it's highly parallel nature, and ability to easily process different image resolutions, and even process different resolutions within the same image.

Novelty:

A 2018 paper (Liu et al., 2018), which partially inspired (Anokhin et al., 2020), investigated adding coordinates as extra channels to the initial convolution operations. This paper shows advantages of this approach for generalisation and model size. However, based on the prior research section of both (Liu et al., 2018) and (Anokhin et al., 2020), as well as a brief check of related google scholar searches, it appears there hasn't been an investigation of this idea for discriminative models without convolutions, and the specific inductive biases encoded by this coordinate based architecture.

Feasibility:

Last week it took a few hours for me to set up the most basic version of the architecture (with no advanced positional encoding), and test it on FashionMNIST, getting a test error comparable to a fully connected network. With some optimisation over architecture choices (especially positional encodings), it's likely it can be improved much further.

References:

Anokhin, I., Demochkin, K., Khakhulin, T., Sterkin, G., Lempitsky, V., & Korzhenkov, D. (2020). Image Generators with Conditionally-Independent Pixel Synthesis. *ArXiv:2011.13775 [Cs]*. http://arxiv.org/abs/2011.13775

Liu, R., Lehman, J., Molino, P., Such, F. P., Frank, E., Sergeev, A., & Yosinski, J. (2018). An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution. *ArXiv:1807.03247 [Cs, Stat]*. http://arxiv.org/abs/1807.03247