Controlling texture and shape bias in deep learning classifiers

We know that neural networks when giving their predictions can be biased towards the shape of objects(general topology), or towards the texture(high frequency repeated patterns within object edges). Being biased towards either shape or texture is not necessarily bad but it could be useful or harmful depending on the specifics of the task. We propose to allow the freedom to control the amount of shape or texture bias.

We combine ideas from style transfer and *L2 regularization* to attempt to control the shape-texture bias.In style transfer we iteratively try to match the gram matrix of the intermediate representations of the content image to match the intermediate representations of the style image. The gram matrix of the representations captures the necessary texture information. While training networks: feed images throughout the networks, then collect intermediate representations at specific layers, compute gram matrices of these representations, and compute the norms of the gram matrices.

Our goal is to introduce a hyper-parameter b in the loss function such that Having a large value of b will force the model towards focusing into the shape information, because we will encourage the model to have a small norm of Gram matrix of representations and thus force it to lose texture information. Having small values of b will allow the model more freedom to capture whatever texture information it wants. We are currently experimenting.