4. **Write a 2 page (excluding figures) summary of MixUp regularization and how this may help (30 points).**

For fine tuning (FT), you only need to grid search over one hyper parameter, which then decides on training schedule length, resolution, and whether to use MixUp regularization.

MixUp regularization is a technique that helps when you have little data and it interpolates between datapoints and trains on datapoints from half one class and half another class to make more data available.

We found that MixUp is not useful for pre-training BiT, likely due to the abundance of data. However, it is sometimes useful for transfer. Interestingly, it is most useful for mid-sized datasets, and not for few-shot transfer.

Large deep neural networks are powerful but exhibit undesirable behaviors such as memorization and sensitivity to adversarial examples. Mixup trains a neural network on convex combinations of pairs of examples and their labels. By doing so, mixup regularizes the neural network to favor simple linear behavior in-between training examples.

Experiments on the ImageNet-2012, CIFAR-10, CIFAR-100, Google commands and UCI datasets show that mixup improves the generalization of state-of-the-art (SOTA) neural network architectures. We also find that mixup reduces the memorization of corrupt labels, increases the robustness to adversarial examples, and stabilizes the training of generative adversarial networks.

To attain a low per-task adaptation cost, we do not perform any hyperparameter sweeps downstream. Instead, we present BiTHyperRule, a heuristic to determine all hyperparameters for fine-tuning. Most hyperparameters are fixed across all datasets, but schedule, resolution, and usage of MixUp depend on the tasks image resolution and training set size.

We use MixUp, with $\alpha = 0.1$, for medium and large tasks