What is Mixup?

Generate additional samples by interpolating samples:

\[
\tilde{x} = \lambda x_a + (1-\lambda) x_b \quad \tilde{y} = \lambda y_a + (1-\lambda) y_b
\]

Mixing of two images

Mixing of one-hot labels

👍 Easy and effective way to enrich any training set

👍 Improves accuracy, calibration & out-of-distribution generalization

👎 Assumption: mixing should be done at output probability level

👎 Mixup wants linearity, but it's not enforced in classifier space!

Motivation

Image: 40% dog and 60% cat

Mixup: Classify as 40% dog 60% cat.

Infinite Class Mixup: Classify this image 100% to the new class: 40% dog and 60% cat.

Mixup with Infinite Classes

Main idea: for each mixed sample, we construct a unique interpolated classifier:

\[
\mathbf{w}_c = \lambda \mathbf{w}_a + (1-\lambda) \mathbf{w}_b = W\mathbf{\tilde{y}}
\]

Optimize likelihood of interpolated classes:

\[
p(C_{\mathbf{\tilde{y}}} | \mathbf{x}) \propto \exp(f_\theta(\mathbf{x})^T W\mathbf{\tilde{y}})
\]

via contrastive learning in two ways:

Class-axis

Contrast example against all interpolated classifiers in the batch.

In the spirit of conventional contrastive learning.

Pair-axis

Contrast classifier against all interpolated samples in the batch.

Exists because of the unique construction of each classifier. Not (directly) applicable in Mixup.

Optimization: \(f_\theta(\mathbf{x})^T W\mathbf{\tilde{y}}\) is a \(B \times B\) matrix for batch size \(B\), we perform cross-entropy loss across both axes.

Experiments

The class- and pair-axes are complementary.

Improves long-tailed recognition.

Insight 1: Lower confidence for ambiguous samples.

Insight 2: Better differentiation between interpolations.

Enhances recent RegMixup (NeurIPS 2022).