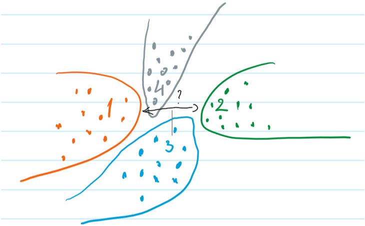
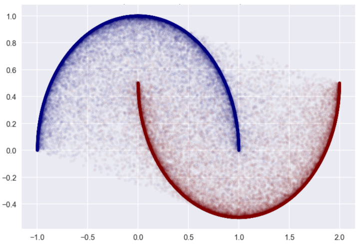
Semi-supervised Learning Models, Group of Methods "MixUp"

Outline

- MixUp (Zhang et al., ICLR 2018)
  - data augmentation method: train models on convex combination two inputs and two labels
- MixUp strategies for semi-supervised learning
  - ICT (Verma et al., IJCAI 2019)
  - MixMatch (Berthelot et al., NeurIPS 2019)
- MixUp strategies in NLP (Guo et al., 2019, Thulasidasan et al., NeurIPS 2019)
  - Word embeddings
  - Final vector representation

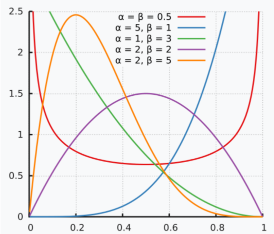
MixUp: Intuition





- ML models rely and are overconfident on existing examples
- We need to fill in the gap between the points
- The paper bases on vicinal risk minimization (VRM) theory instead of empirical risk minimization (ERM)
    - train on augmented examples neighboring original examples

MixUp: Data augmentation



- Convex combination of two examples

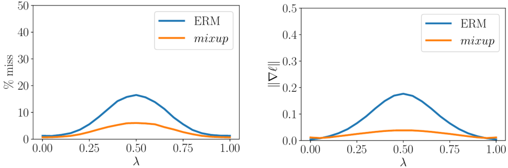
$$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$$
$$(x_i, y_i), (x_j, y_j) \sim \mathcal{D}_l$$
$${x_i}' = \lambda x_i + (1 - \lambda) x_j$$
$${y_i}' = \lambda y_i + (1 - \lambda) y_j$$

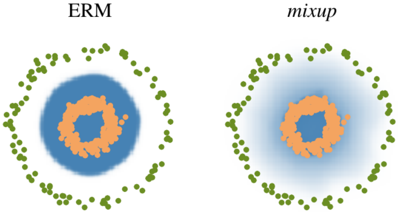
- Vicinal objective function

$$\mathcal{L}(\theta) = \ell(p(y|{x_i}'; \theta)\ , {y_i}')$$

MixUp: Benefits

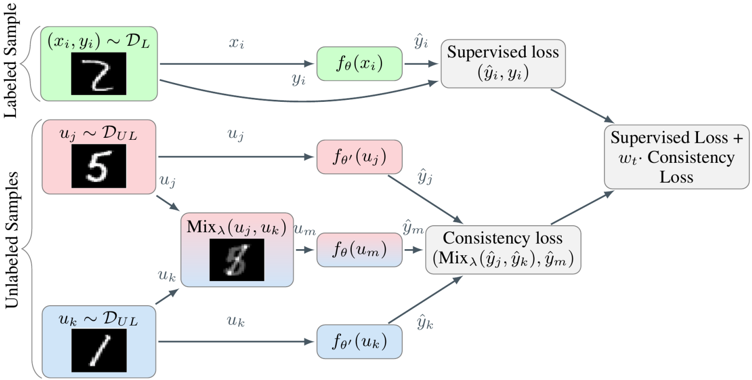
- Agnostic data augmentation: work on image, speech, text (with modification), tabular data, stablizing GANs
- Generalize better to out-of-distribution examples
- Robustness to adversarial examples
- Smooth decision boundaries, reduce memorization and overfitting





MixUp strategies for semi-supervised learning: ICT (Verma et al., IJCAI 2019)

- Mixing unlabeled examples



MixUp strategies for semi-supervised learning: MixMatch (Berthelot et al., NeurIPS 2019)

- Mixing both labeled and unlabeled examples as follows
  - Label predictions for unlabeled examples, averaging output distribution for K augmented unlabeled examples

$$y_{\mathrm{ul}} = \frac{1}{K}\sum_{k=1}^{K} p(y|\hat{u}_{\mathrm{ul}};\theta)$$

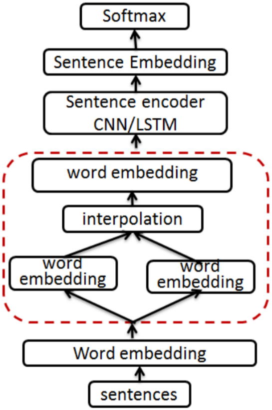
  - Sharpening (temperature anealing) label predictions, forcing output distribution towards one-hot vector by adjusting temperature $T \to 0$

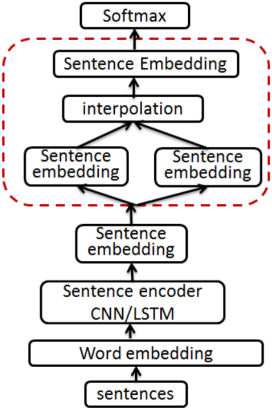
$$y_{\mathrm{ul},c} = \frac{y_{\mathrm{ul},c}^{\frac{1}{T}}}{\sum_{c=1}^{C} y_{\mathrm{ul},c}^{\frac{1}{T}}}$$

  - MixUp all labeled and unlabeled examples
- Warning: during training, we pass K duplicated unlabeled images at the same time, so we need to adjust batchnorm updates to only take one set unlabeled images

NLP MixUp Strategies

- Word embeddings
  - Mixing embeddings of aligned words between two sentences (padding sentences if one is shorter than the other)
- Final vector representation





NLP MixUp Results

RandomTune	Trec	SST-1	SST-2	Subj	MR
CNN- KIM Impl. (Kim, 2014b)	91.2	45.0	82.7	89.6	76.1
CNN- HarvardNLP Impl. [1]	88.2	42.2	83.5	89.2	75.9
CNN - Our Impl.	90.2±0.20	43.6±0.19	82.3±0.47	90.6±0.45	75.5±0.36
CNN+wordMixup	90.9±0.42	45.2±0.90	82.8±0.45	**92.9±0.41**	**78.0±0.39**
CNN+senMixup	**92.1±0.31**	**45.2±0.22**	**83.0±0.35**	92.7±0.38	77.9±0.76

RandomTune	Trec	SST-1	SST-2	Subj	MR
LSTM-StanfordNLP Impl. (Tai et al., 2015)	N/A	46.4	84.9	N/A	N/A
LSTM-AgrLearn Impl. (Guo et al., 2018a)	N/A	N/A	N/A	90.2	76.2
LSTM - Our Impl.	86.5±0.61	45.9±0.58	84.4±0.35	90.9±0.42	77.2±0.75
LSTM + wordMixup	**90.5±0.50**	48.2±0.18	86.3±0.35	**93.1±0.49**	**78.0±0.33**
LSTM + senMixup	89.4±0.40	**48.3±0.77**	**86.7±0.33**	91.9±0.34	77.9±0.33

Thank you !