DMix: Distance Constrained Interpolative Mixup

Ramit Sawhney*
ShareChat
ramitsawhney@sharechat.co

Megh Thakkar*
BITS, Pilani
megh.1211@gmail.com

Shrey Pandit*
BITS Pilani, Goa Campus
f20190138@goa.bits-pilani.ac.in

Debdoot Mukherjee
ShareChat
debdoot.iit@gmail.com

Lucie Flek
Conversational AI and Social Analytics (CAISA) Lab
University of Marburg
lucie.flek@uni-marburg.de

Abstract

Interpolation-based regularisation methods have proven to be effective for various tasks and modalities. Mixup is a data augmentation method that generates virtual training samples from convex combinations of individual inputs and labels. We extend Mixup and propose DMix, distance-constrained interpolative Mixup for sentence classification leveraging the hyperbolic space. DMix achieves state-of-the-art results on sentence classification over existing data augmentation methods across datasets in four languages.

1 Introduction

Deep learning models are effective across a wide range of applications. However, these models are prone to overfitting when only limited training data is available. Interpolation-based approaches such as Mixup (Zhang et al., 2018) have shown improved performance across different modalities. Mixup over latent representations of inputs has led to further improvements, as latent representations often carry more information than raw input samples. However, Mixup does not account for the spatial distribution of data samples, and chooses samples randomly.

While randomization in Mixup helps, augmenting Mixup's simple selection strategy with logic based on the similarity of the samples to be mixed can lead to improved generalization. Further, natural language text possesses hierarchical structures and complex geometries, which the standard Euclidean space cannot capture effectively. In such a scenario, hyperbolic geometry presents a solution in defining similarity between latent representations via hyperbolic distance.

We propose **DMix**, a distance-constrained interpolative data augmentation method. Instead of choosing random inputs from the complete training distribution as in the case of vanilla Mixup, DMix samples instances based on the (dis)similarity between latent representations of samples in the hyperbolic space. We probe DMix through experiments on sentence classification tasks across four languages, obtaining state-of-the-art results over existing data augmentation techniques.

2 Methodology

Interpolative Mixup Given two data samples $x_i, x_j \in X$ with labels $y_i, y_j \in Y$, *Mixup* (Zhang et al., 2018) uses linear interpolation with mixing ratio r to generate the synthetic sample $x' = r \cdot x_i + (1-r) \cdot x_j$ and corresponding mixed label $y' = r \cdot y_i + (1-r) \cdot y_j$. Interpolative Mixup (Chen et al., 2020) performs performs linear interpolation over the latent representations of models.

Let $f_\theta(\cdot)$ be a model with parameters θ having N layers, $f_{\theta,n}(\cdot)$ denotes the n-th layer of the model and h_n is the hidden space vector at layer n for $n \in [1, N]$ and h_0 denotes the input vector. To perform interpolative Mixup at a layer $k \sim [1, N]$, we first calculate the latent representations separately for the inputs for layers before the k-th layer. For input samples x_i, x_j, we let h_n^i, h_n^j denote their respective hidden state representations at layer n,

$$\begin{aligned} h_n^i &= f_{\theta,n}(h_{n-1}^i), \quad n \in [1,k] \\ h_n^j &= f_{\theta,n}(h_{n-1}^j), \quad n \in [1,k] \end{aligned} \tag{1}$$

We then perform Mixup over individual hidden state representations h_k^i, h_k^j from layer k as,

$$h_k = r \cdot h_k^i + (1-r) \cdot h_k^j \tag{2}$$

The mixed hidden representation h_k is used as the input for the continuing forward pass,

$$h_n = f_{\theta,n}(h_{n-1}); \quad n \in [k+1, N] \tag{3}$$

DMix To perform distance-constrained interpolative Mixup, for a sample x_i, we calculate

*equal contribution

Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 242–244
November 11, 2021.

its similarity with every other sample $x \in X$ between their sentence embedding. As natural language exhibits hierarchical structure, embeddings are more expressive when represented in the hyperbolic space (Dhingra et al., 2018). We use hyperbolic distance $\mathcal{D}_h = 2\tan^{-1}(\|(-x_i) \oplus x\|)$ as a similarity measure. We sort the distances in decreasing order for x_i, and randomly select one sample x_j from top-τ samples, where τ is a hyperparameter, which we call threshold. Formally,

$$x_j \sim \text{top-}\tau([\mathcal{D}_h(x_i, x) \forall x \in X]) \quad (4)$$

3 Experiments and Results

We evaluate DMIX on sentence classification tasks:
Arabic Hate Speech Detection AHS is a binary classification task over 3950 Arabic tweets containing hate speech.
English SMS Spam Collection ESSC is a dataset with 5574 raw text messages classified as spam or not spam.
Turkish News Classification TTC-3600 contains 3600 Turkish news text across six news categories.
Gujarati Headline Classification GHC has 1632 Gujarati news headlines over three news categories.
Training Setup: Mixup is performed over a random layer sampled from all the layers of the model. The model was trained with a learning rate of 2e-5, with a training batch size of 8 and a weight decay of 0.01. All hyperparameters were selected based on validation F1-score.

3.1 Performance Comparison

Model	AHS	ESSC	TTC	GHC
mBERT	[illegible]6.20	98.30	2[illegible].54	64.88
+Input Mixup	[illegible]	98.6[illegible]	30.05	65.64
+Sentence Mixup	67.[illegible]	98.4[illegible]	30.88	65.60
+Mixup	67.78	9[illegible].90	30.71	66.41
mBERT+distance-constrained Mixup (Ours)				
- Euclidean	74.[illegible]2	86.87	30.89	65.88
- Cosine	77.50*	98.23*	31.60*	67.39*
- DMIX (Hyperbolic)	79.19*	99.30*	32.00*	69.67*

Table 1: Performance comparison in terms of F1 score of DMix with vanilla Mixup and distance-constrained Mixup methods using different similarity techniques (average of 10 runs). Improvements are shown with blue (↑) and poorer performance with red (↓). * shows significant ($p < 0.01$) improvement over Mixup.

We observe that distance-constrained Mixup outperforms vanilla Mixup ($p < 0.01$) across numerous tasks and distance based (dis)similarity formulation, validating that similarity-based sample selection improves model performance, likely owing to enhanced diversity or minimizing sparsification across tasks. Within distance-constrained Mixup, we observe that DMIX, the hyperbolic distance variant outperforms Euclidean distance and cosine similarity measures. This suggests that the hyperbolic space is more capable of capturing the complex hierarchical information present in sentence representations, leading to more pronounced comparisons and sample selection.

3.2 Threshold Variation Analysis

We perform an ablation study by varying the threshold τ for DMix and present it in Figure 1[1]. An increasing τ denotes a larger distribution space for sampling instances for Mixup, and a τ of 100% degenerating to vanilla Mixup. We observe an initial increase in the performance as we expand the sampling embedding space, and then it decreases, essentially decomposing into randomized Mixup. This suggests the existence of an optimum set of input samples for performing Mixup, and we conjecture it can be related to the sparsity in the embedding distribution of different languages.

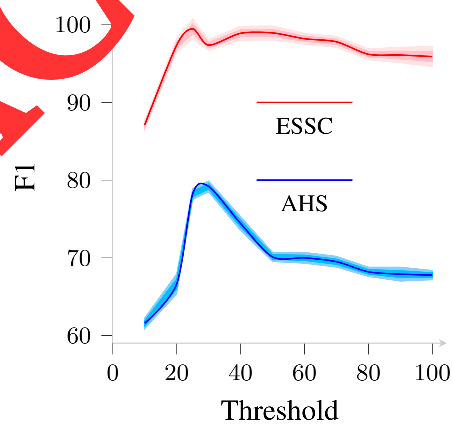

Figure 1: Change in performance in terms of F1 with varying threshold for DMIX. A threshold of 100% decomposes DMIX into vanilla Mixup.

4 Conclusion

We propose DMIX, an interpolative regularization based data augmentation technique sampling inputs based on their latent hyperbolic similarity. DMIX achieves state-of-the-art results over existing data augmentation approaches on datasets in four languages.We further analyze DMIX through ablations over different similarity threshold values across the languages. DMIX being data-, modality-, and model- agnostic, holds potential to be applied on text, speech, and vision tasks.

[1] We obtain similar results for TTC and GHC.

References

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. 2018. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.