Embedding Space Interpolation Beyond Mini-Batch, Beyond Pairs and Beyond Examples

Shashanka Venkataramanan [1], Ewa Kijak [1], Laurent Amsaleg [1], and Yannis Avrithis [2]

[1] Inria, Univ Rennes, CNRS, IRISA; [2] IARAI

Mixup improves generalization

- Data Augmentation technique that **interpolates** between pairs of examples (input/feature) and its labels.
- **Flattens** class representations, reduces overconfident incorrect predictions and **smoothens** decision boundaries.

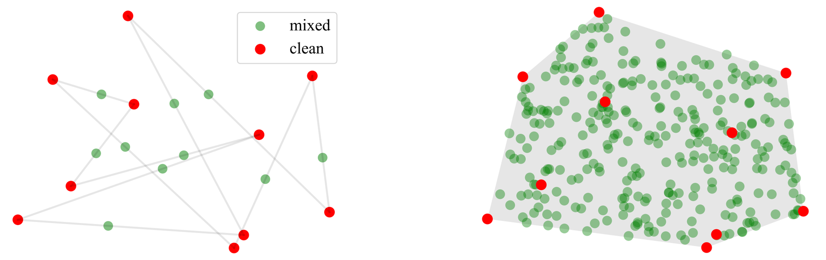
Empirical Risk Minimization to Mixup

- The **expected risk** is defined as an integral over the underlying continuous data distribution.
- Since that distribution is unknown, the integral is approximated by a finite sum, i.e., the **empirical risk**.
- A better approximation is the **vicinal risk** – augmented examples are sampled from a distribution in the vicinity of each training example: increasing the number of loss terms per training example.
- Input Mixup is inspired by vicinal risk, but for a mini-batch of size b, it generates **only b mixed examples** and thus incur **b loss term**.

Better approximation of expected risk integral

Data augmentation should increase the data seen by the model. We propose **MultiMix**, which:

- Increases the number n of **generated mixed examples** beyond the mini-batch size b.
- Increases the number m of **examples being interpolated** from $m = 2$ (pairs) to $m = b$.
- Performs interpolation in the **embedding space** rather than input space.



(a) mixup (b) MultiMix (ours)

MultiMix

Preliminaries

- For a mini-batch of b examples, $X = (x_1, \dots, x_b) \in \mathbb{R}^{D\times b}$ be the inputs, $Y = (y_1, \dots, y_b) \in \mathbb{R}^{c\times b}$ the targets; c is the total number of classes.
- $f_\theta : \mathcal{X} \to \mathbb{R}^d$ is an encoder that maps the input x to an embedding $z = f_\theta(x)$; d is the dimension of the embedding.

Mixup

- **Manifold mixup** [1] interpolates the embeddings (Z) and targets (Y) by forming a convex combination of the pairs with interpolation factor $\lambda \in [0, 1]$:

$$\widetilde{Z} = Z(\lambda I + (1-\lambda)\Pi) \tag{1}$$

$$\widetilde{Y} = Y(\lambda I + (1-\lambda)\Pi) \tag{2}$$

$\lambda \sim \mathrm{Beta}(\alpha, \alpha)$, I is the identity matrix, $\Pi \in \mathbb{R}^{b\times b}$ is a permutation matrix.

- The number of generated examples per mini-batch is $n = b$, and each is obtained by interpolating $m = 2$ examples.
- The total number of loss terms per mini-batch is again b.

MultiMix

- We draw interpolation vectors $\lambda_k \sim \mathrm{Dir}(\alpha)$ for $k = 1, \dots, n$. $\mathrm{Dir}(\alpha)$ is the symmetric **Dirichlet distribution**, $\lambda_k \in \Delta^{m-1}$ *i.e.* $\lambda_k \geq 0$ and $\mathbf{1}_m^\top \lambda_k = 1$.
- We interpolate embeddings and targets by taking n convex combinations over **all m examples**

$$\widetilde{Z} = Z\Lambda \tag{3}$$

$$\widetilde{Y} = Y\Lambda, \tag{4}$$

where $\Lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^{b\times n}$.

Generalizing manifold mixup

- from b to an arbitrary number $n \gg b$ of generated examples.
- from pairs ($m = 2$) to a tuple of length $m = b$, containing the entire mini-batch.
- **m-term convex combination** *vs.* 2-term, **Dirichlet** *vs.* Beta distribution.
- from fixed λ across the mini-batch to a **different λ_k** for each generated example.

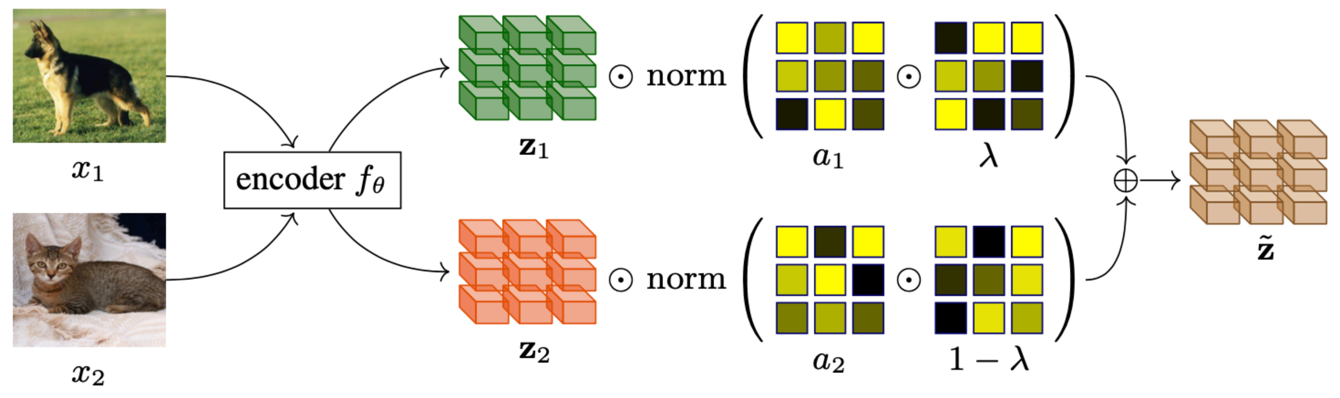
Dense MultiMix

Preliminaries

- Each embedding $\mathbf{z}_i = f_\theta(x_i) = (z_i^1, \dots, z_i^r) \in \mathbb{R}^{d\times r}$ for $i = 1, \dots, b$ consists of features $z_i^j \in \mathbb{R}^d$ for spatial position $j = 1, \dots, r$.
- We group features by position in matrices $Z^1, \dots, Z^r$, where $Z^j = (z_1^j, \dots, z_b^j) \in \mathbb{R}^{d\times b}$ for $j = 1, \dots, r$.

Dense MultiMix

- Common way to increase the number of loss terms – **Dense operations**.
- Densely interpolate features at each spatial location: generate **r interpolated features** and $nr > n$ per mini-batch.



Using attention as pseudo-labels

- Attention map gives a **level of confidence**, selects reliable spatial location to locate the target.
- Let $a_i = (a_i^1, \dots, a_i^r) \in \mathbb{R}^r$ be the attention map of embedding $\mathbf{z}_i$ for $i = 1, \dots, b$ obtained **using CAM**.
- We group attention by position in vectors $a^1, \dots, a^r$, where $a^j = (a_1^j, \dots, a_b^j) \in \mathbb{R}^b$ for $j = 1, \dots, r$.

Interpolation

- For each spatial position $j = 1, \dots, r$, we draw $\lambda_k^j \sim \mathrm{Dir}(\alpha)$ for $k = 1, \dots, n$ and define $\Lambda^j = (\lambda_1^j, \dots, \lambda_n^j) \in \mathbb{R}^{m\times n}$.
- We re-weight Λ using attention and normalize it as:

$$M^j = \mathrm{diag}(a^j)\Lambda^j \tag{5}$$

$$\hat{M}^j = M^j \, \mathrm{diag}(\mathbf{1}_m^\top M^j)^{-1} \tag{6}$$

- We interpolate embeddings and targets by taking n convex combinations over m examples:

$$\widetilde{Z}^j = Z^j \hat{M}^j \tag{7}$$

$$\widetilde{Y}^j = Y \hat{M}^j. \tag{8}$$

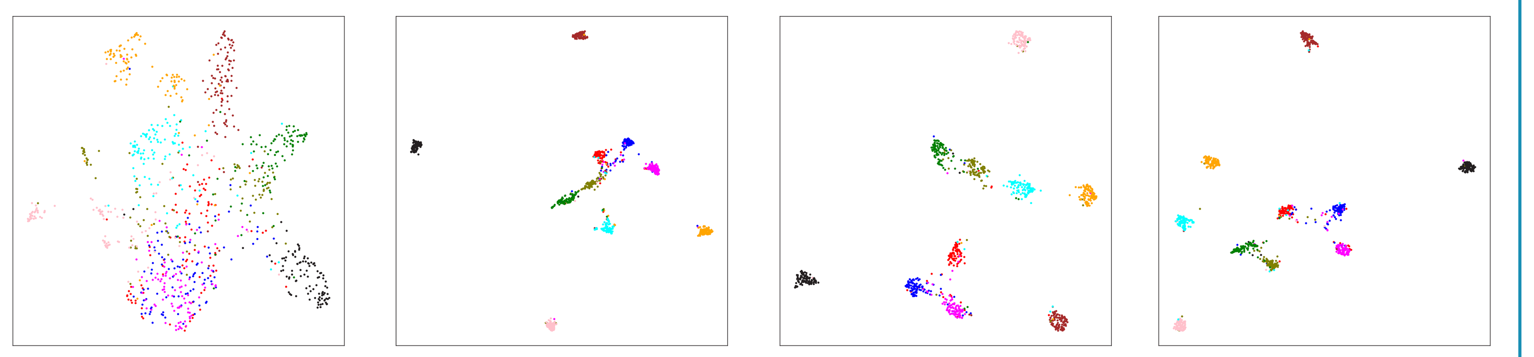
Experimental Results

Out-of-distribution Detection

Dataset	LSUN (crop)			iSUN			TI (crop)		
Metric	AuROC	AuPR (ID)	AuPR (OOD)	AuROC	AuPR (ID)	AuPR (OOD)	AuROC	AuPR (ID)	AuPR (OOD)
Baseline	47.1	54.5	45.6	72.3	74.5	69.2	64.8	67.8	60.6
Input mixup	59.3	61.4	55.2	63.0	60.2	63.4	62.8	63.0	62.1
Cutmix	63.1	61.9	63.4	76.3	81.0	77.7	84.3	87.1	80.6
Manifold mixup	60.3	57.8	59.5	73.1	80.7	76.0	69.9	69.3	70.5
AugMix	73.2	80.8	72.6	78.7	81.1	74.1	83.9	84.6	78.6
SaliencyMix	79.7	82.2	64.4	76.9	78.3	79.8	83.7	87.0	82.0
StyleMix	64.2	70.9	63.9	68.4	67.6	60.3	73.9	71.5	78.4
AlignMixup	<u>79.9</u>	<u>84.1</u>	<u>75.1</u>	<u>83.2</u>	<u>84.1</u>	<u>80.3</u>	<u>85.0</u>	<u>87.8</u>	<u>85.0</u>
ζ-Mixup	73.2	80.8	73.1	82.3	82.2	79.4	84.3	82.2	77.2
MultiMix (ours)	82.6	85.2	77.6	85.1	87.8	83.1	86.6	89.0	**88.2**
Dense MultiMix (ours)	**84.3**	**85.9**	**78.0**	**85.4**	**88.0**	**84.6**	89.0	**90.8**	88.0
Gain	+4.4	+1.8	+2.9	+2.2	+3.9	+4.3	+4.0	+3.0	+3.2

Out-of-distribution detection using R-18. ID: In-distribution, OOD: Out-of-distribution. Evaluation metric - AuROC, AuPR (ID) and AuPR (OOD): higher is better. underline: best baseline. Gain: increase in performance. TI: TinyImagenet.

Analysis of Embedding space



Baseline AlignMixup MultiMix (ours) Dense MultiMix (ours)

Metric	Alignment	Uniformity
Baseline	3.02	-1.94
AlignMixup [3]	2.04	-2.38
MultiMix (ours)	1.27	-4.77
Dense MultiMix (ours)	**0.92**	**-5.68**

References

[1] Verma et al. Manifold mixup: Better representations by interpolating hidden states *ICML*, 2019.
[2] Zhang et al. mixup: Beyond empirical risk minimization *ICLR*, 2018.
[3] Venkataramanan et al. AlignMixup: Improving Representations By Interpolating Aligned Features *CVPR*, 2022.

Acknowledgement

This work was partially supported by the ANR-19-CE23-0028 MEERQAT and was performed using the HPC resources from GENCI-IDRIS Grant 2021 AD011012528.