MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification

ACL 2020

Jiaao Chen, Zichao Yang, Diyi Yang

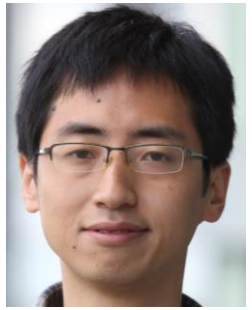
2022.01.17
AILAB 조환희

Authors

Jiaao Chen
Ph.D. School of Interactive Computing at Georgia Tech

Research Area
-Machine Learning, NLP

Recent Paper
-Simple Conversational Data Augmentation for Semi-supervised Abstractive Dialogue Summarization, 2021, EMNLP
-An Empirical Survey of Data Augmentation for Limited Data Learning in NLP, 2021



Zichao Yang
Ph.D. student in the Computer Science Department at CMU

Research Area
-machine learning, deep learning and their applications in computer vision, natural language processing

Recent Paper
-Dense-to-Sparse Gate for Mixture-of-Experts, ICLR 2022 (under review)
-Don't Take It Literally: An Edit-Invariant Sequence Loss for Text Generation, 2022

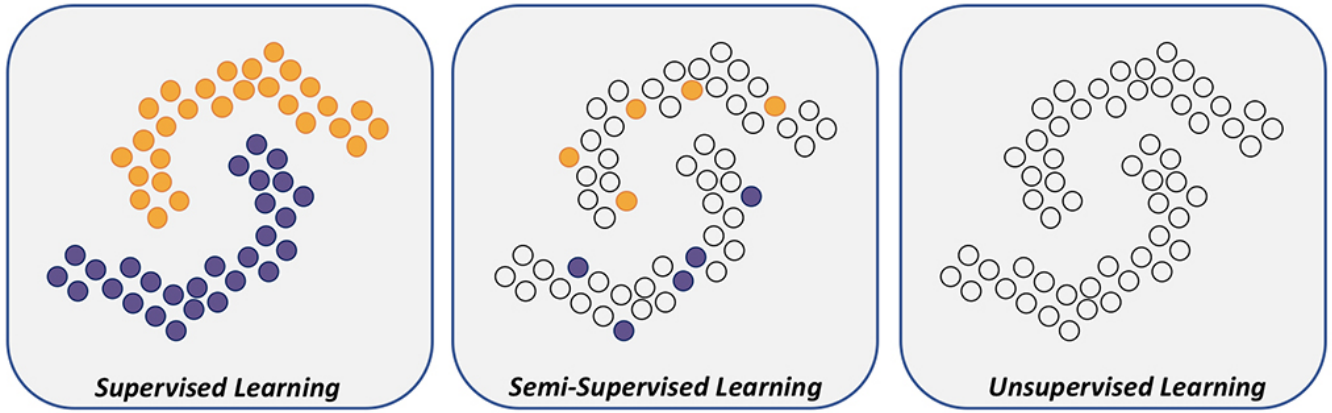
Motivation

이전 semi-supervised learning for text classification은
Labeled Data와 Unlabeled Data를 분리해서 학습

두 데이터 간의 차이를 줄이기 위해
New Augmentation Method 소개
: BERT layer hidden states mixup

Semi-Supervised Learning for text classification

Few labeled + Lots of unlabeled



Method

TMix

새로운 augmentation method

MixText

TMix를 적용한 semi-supervised model

Method: 1. TMix

Mixup

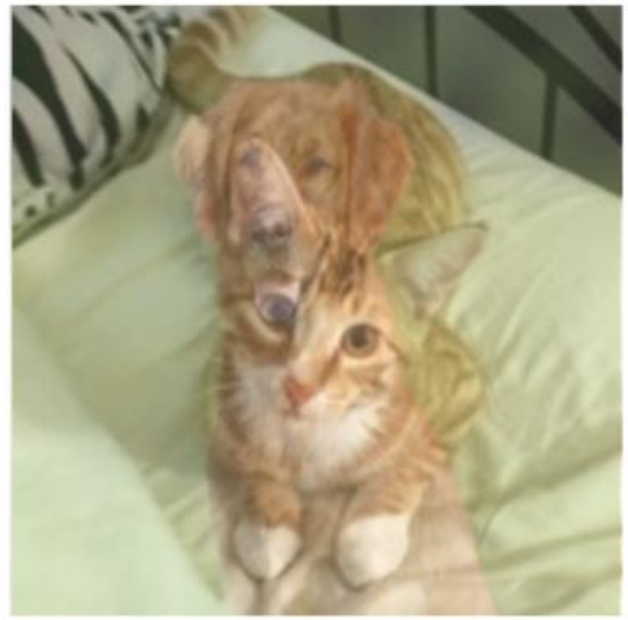
data간 mix

[1.0, 0.0]

cat dog

[0.0, 1.0]

cat dog



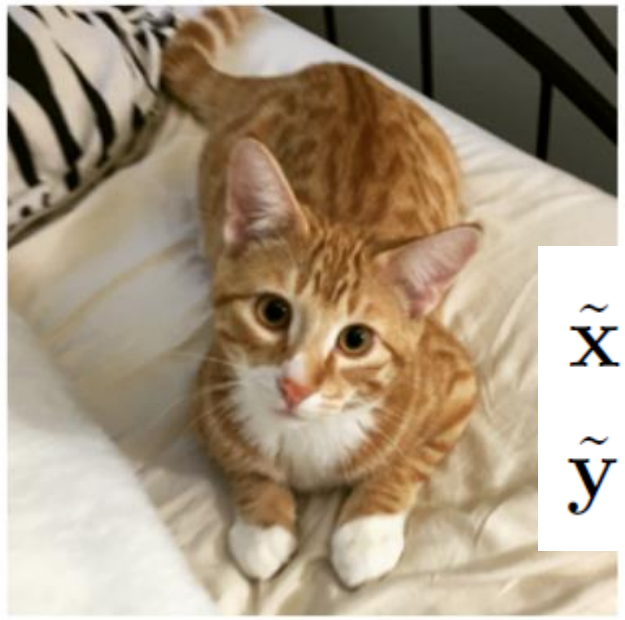
[0.7, 0.3]

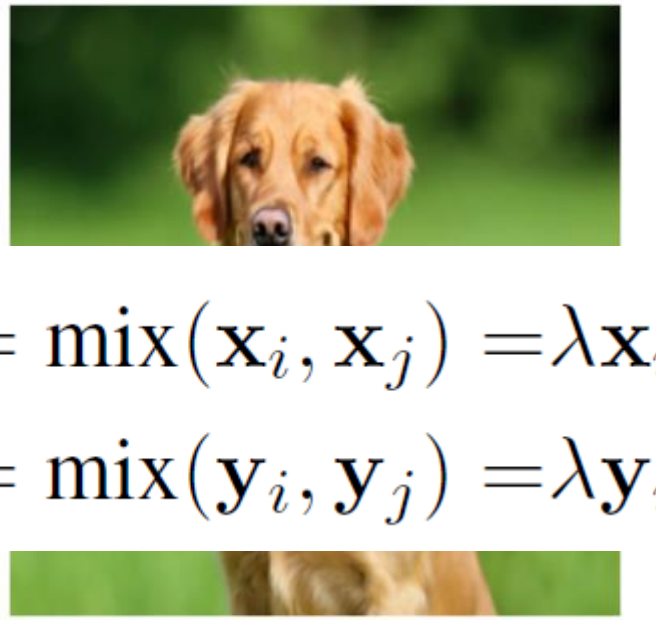
cat dog

Method: 1. TMix

Mixup

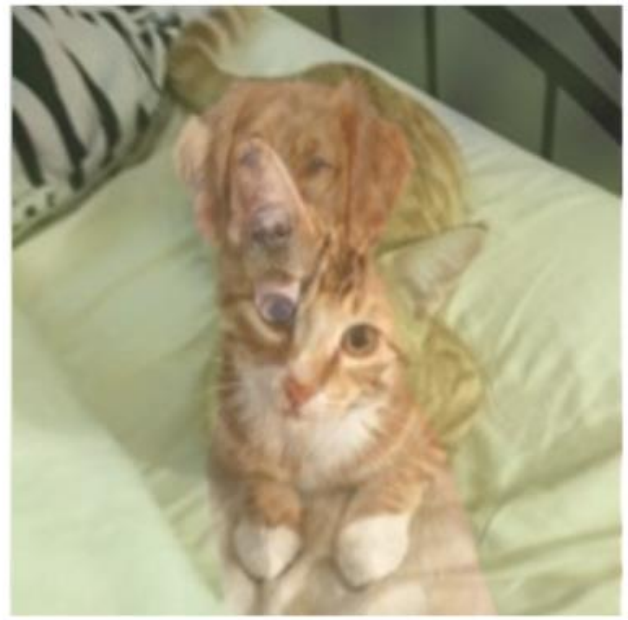
일종의 data augmentation





$$\tilde{\mathbf{x}} = \mathrm{mix}(\mathbf{x}_i, \mathbf{x}_j) = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j$$

$$\tilde{\mathbf{y}} = \mathrm{mix}(\mathbf{y}_i, \mathbf{y}_j) = \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j$$



[1.0, 0.0]
cat dog

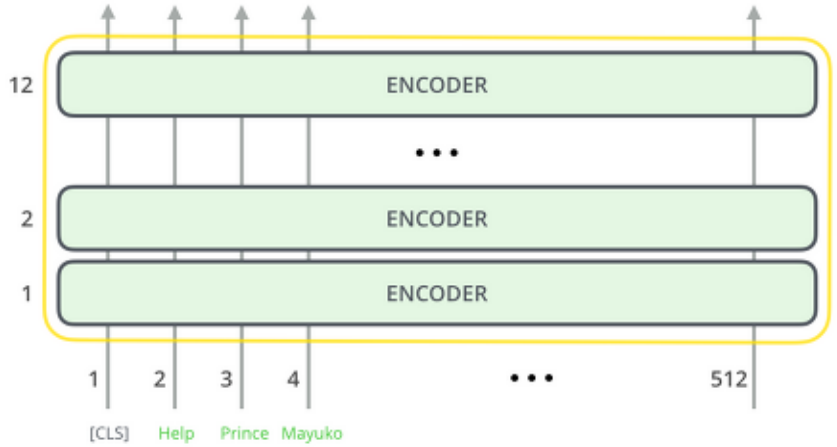
[0.0, 1.0]
cat dog

[0.7, 0.3]
cat dog

Method: 1. TMix

Text Data에서의 mixup 응용

BERT의 12개 hidden layer의 output mixup



BERT

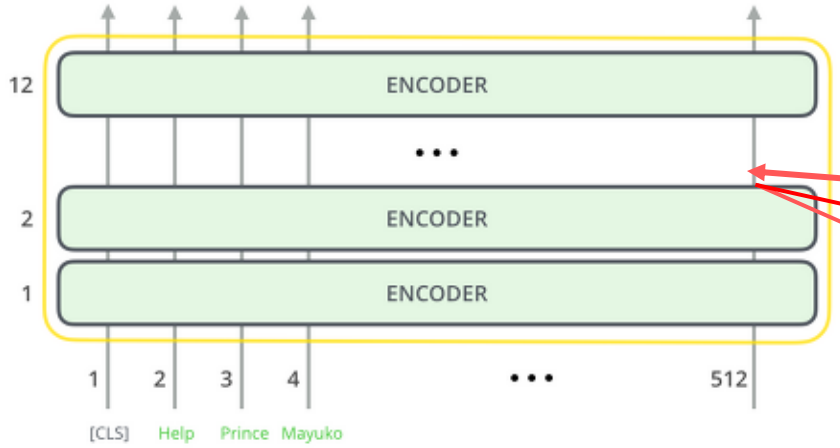
Method: 1. TMix

Text Data에서의 mixup 응용

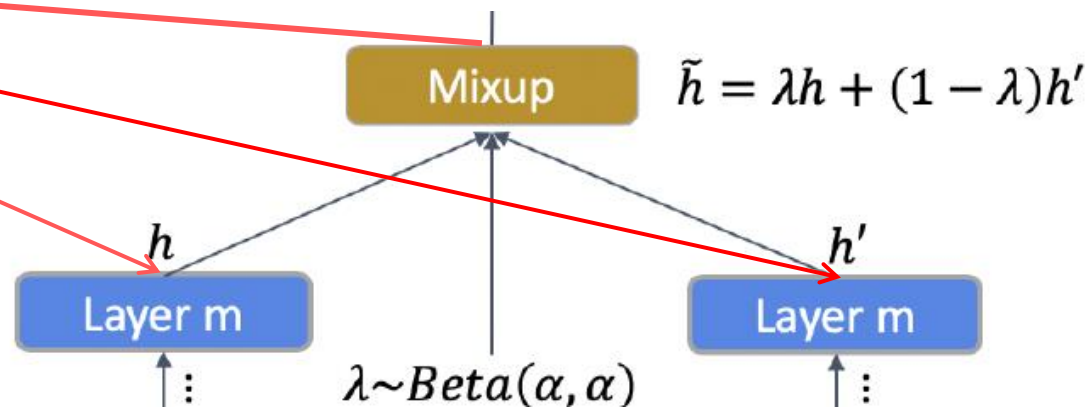
BERT의 12개 hidden layer의 output mixup

h : hidden layer
g : encoder

$$\tilde{\mathbf{h}}_m = \lambda \mathbf{h}_m^i + (1-\lambda)\mathbf{h}_m^j,$$
$$\tilde{\mathbf{h}}_l = g_l(\tilde{\mathbf{h}}_{l-1}; \boldsymbol{\theta}), l \in [m+1, L].$$



BERT



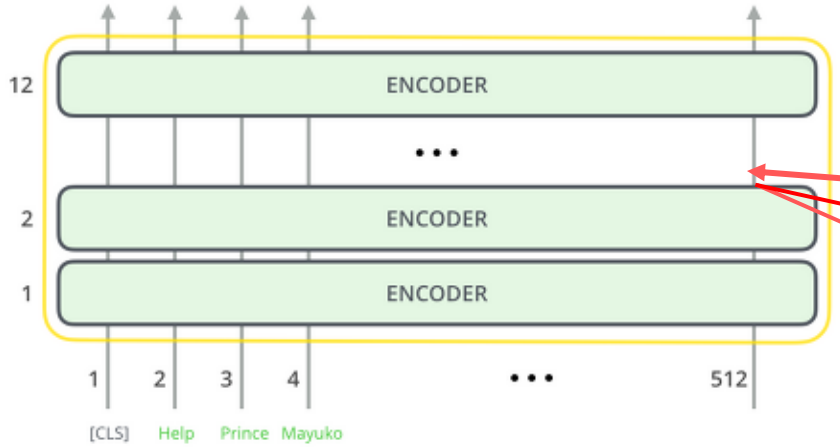
Method: 1. TMix

Text Data에서의 mixup 응용

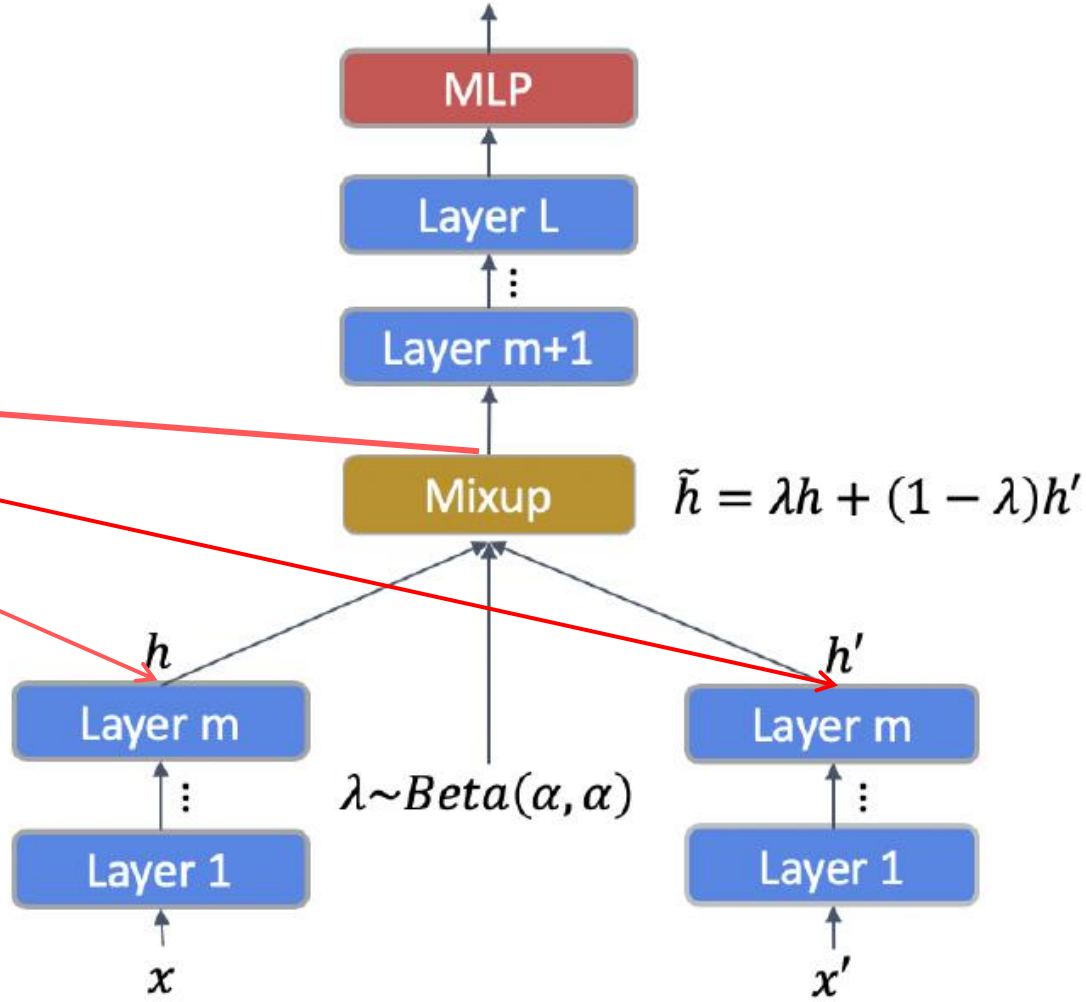
BERT의 12개 hidden layer의 output mixup

$$\text{TMix}(\mathbf{x}_i, \mathbf{x}_j; g(.;\boldsymbol{\theta}), \lambda, m) = \tilde{\mathbf{h}}_L$$

$$\tilde{y} = \lambda y + (1-\lambda)y'$$



BERT

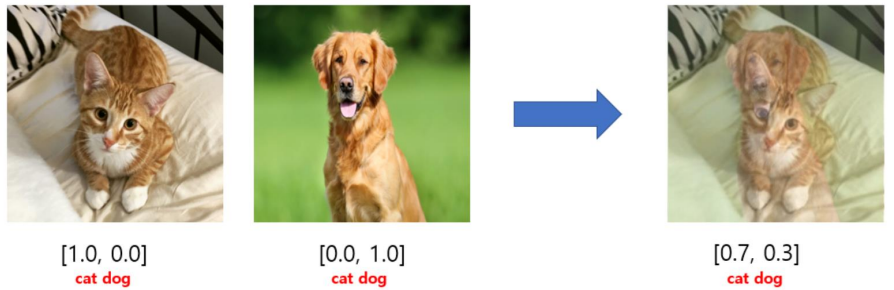


Method: 1. TMix

Text Data에서의 mixup 응용

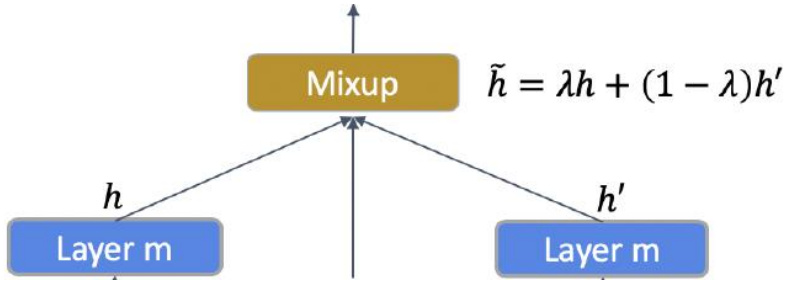
Raw data

$$\tilde{\mathbf{x}} = \text{mix}(\mathbf{x}_i, \mathbf{x}_j) = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j$$
$$\tilde{\mathbf{y}} = \text{mix}(\mathbf{y}_i, \mathbf{y}_j) = \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j$$



Layer output

$$\tilde{\mathbf{h}}_m = \lambda \mathbf{h}_m^i + (1-\lambda)\mathbf{h}_m^j,$$
$$\tilde{\mathbf{h}}_l = g_l(\tilde{\mathbf{h}}_{l-1}; \boldsymbol{\theta}), l \in [m+1, L].$$



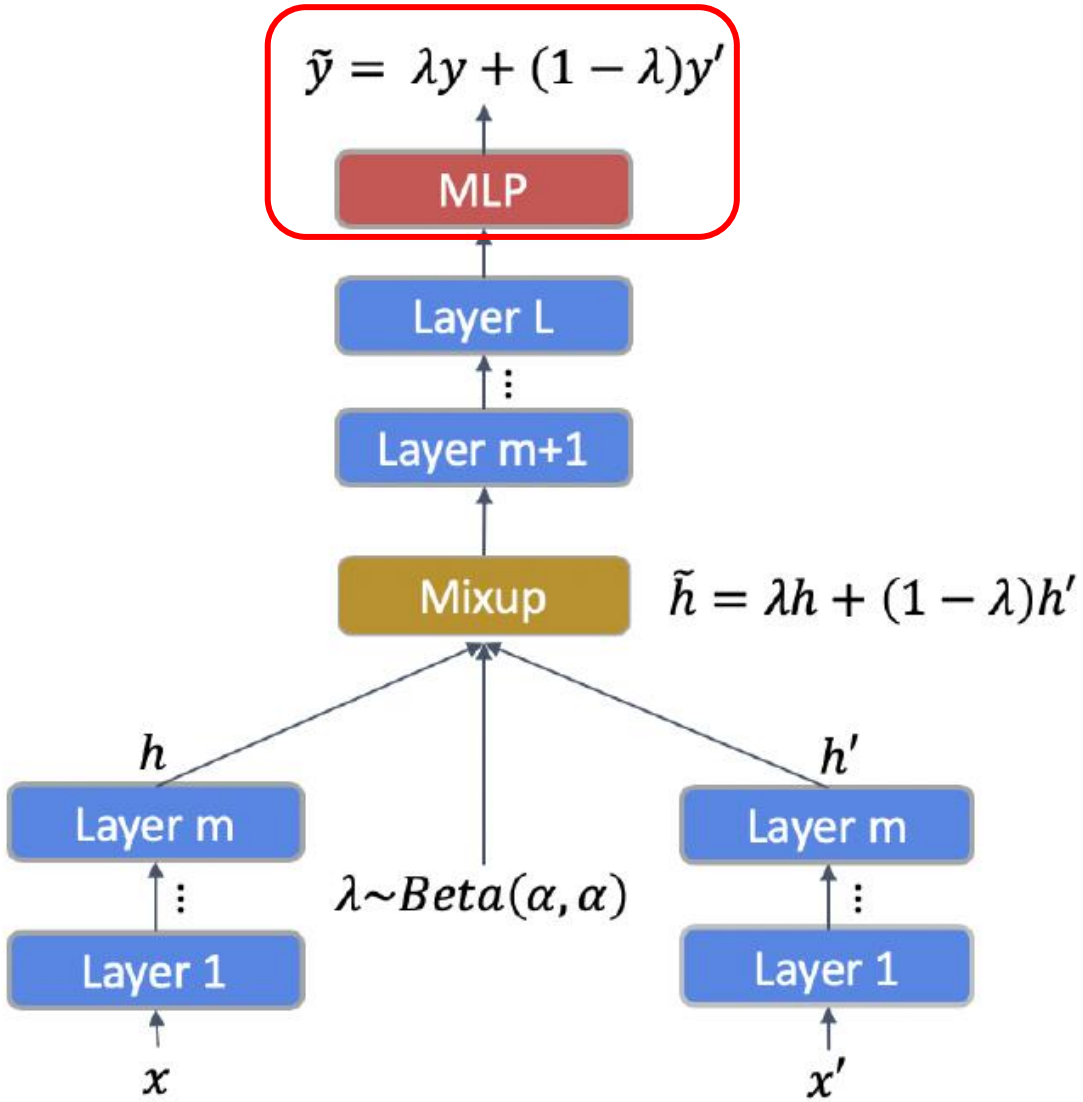
Method: 1. TMix

Text Classification

MLP : fully-connected 2 layers

Minimize kl-divergence

$$L_{\text{TMix}} = \text{KL}(\text{mix}(\mathbf{y}_i, \mathbf{y}_j) || p(\text{TMix}(\mathbf{x}_i, \mathbf{x}_j); \boldsymbol{\phi})$$



Method: 1. TMix

어떤 BERT layer에서 Mixup?

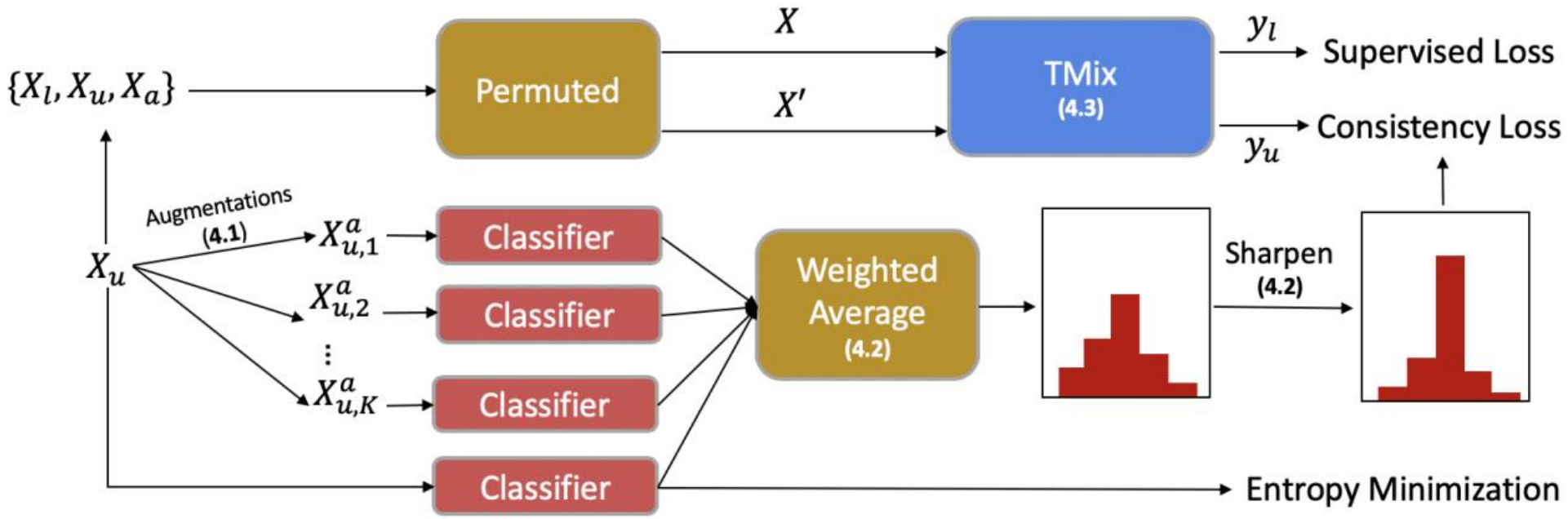
1. BERT는 layer마다 다른 정보를 배움
   What does BERT learn about the structure of language? , ACL, 2019

2. {7,9,12} layer 선택
   - syntactic and semantic information
   - sensitivity to word order
   - sensitivity to random replacement of a noun/verb

Batch 마다 mixup layer randomly sample

MixText : TMix on Semi-Supervised Learning

TMix를 하기 위해 unlabeled data의 label이 필요!

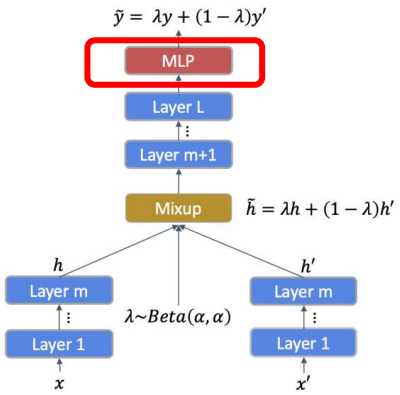


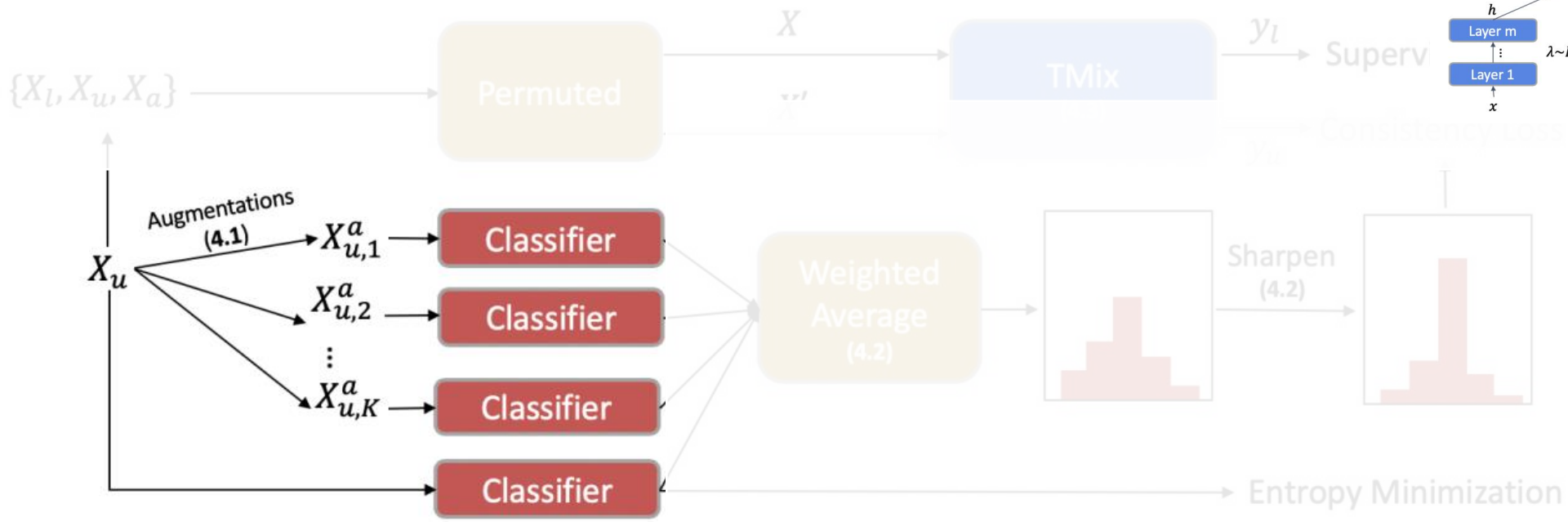
Method: 2. MixText

1. Label Guessing

Unlabeled data back translation으로 augmentation

- 영어->러시아어->영어



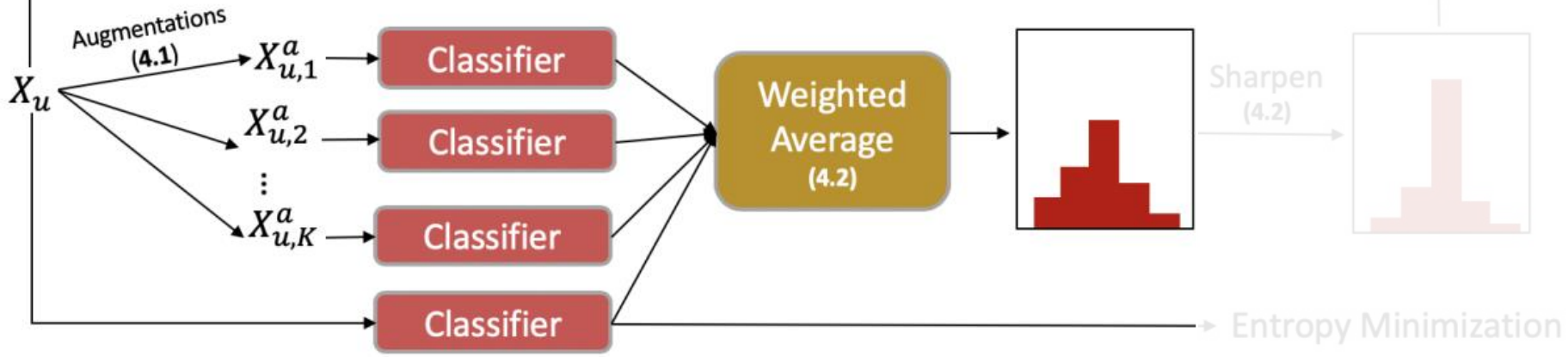


Method: 2. MixText

1. Label Guessing

consistent label guessing → Weighted average

$$\mathbf{y}_i^u = \frac{1}{w_{ori} + \sum_k w_k}(w_{ori} p(\mathbf{x}_i^u) + \sum_{k=1}^{K} w_k p(\mathbf{x}_{i,k}^a)))$$

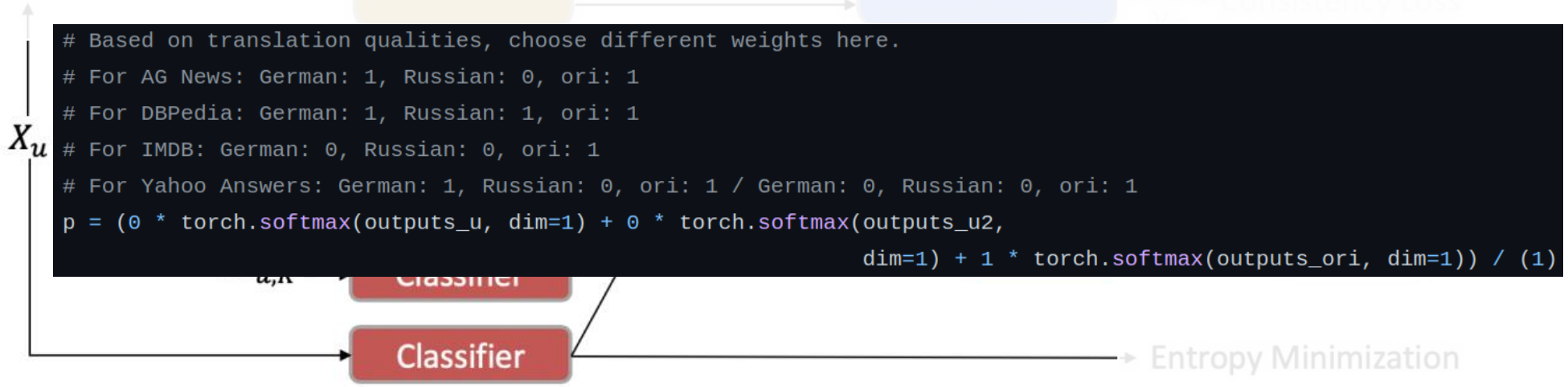


Method: 2. MixText

1. Label Guessing

Weight는 hyperparameter

$$\mathbf{y}_i^u = \frac{1}{w_{ori} + \sum_k w_k}\left(w_{ori} p(\mathbf{x}_i^u) + \sum_{k=1}^{K} w_k p(\mathbf{x}_{i,k}^a)\right)$$



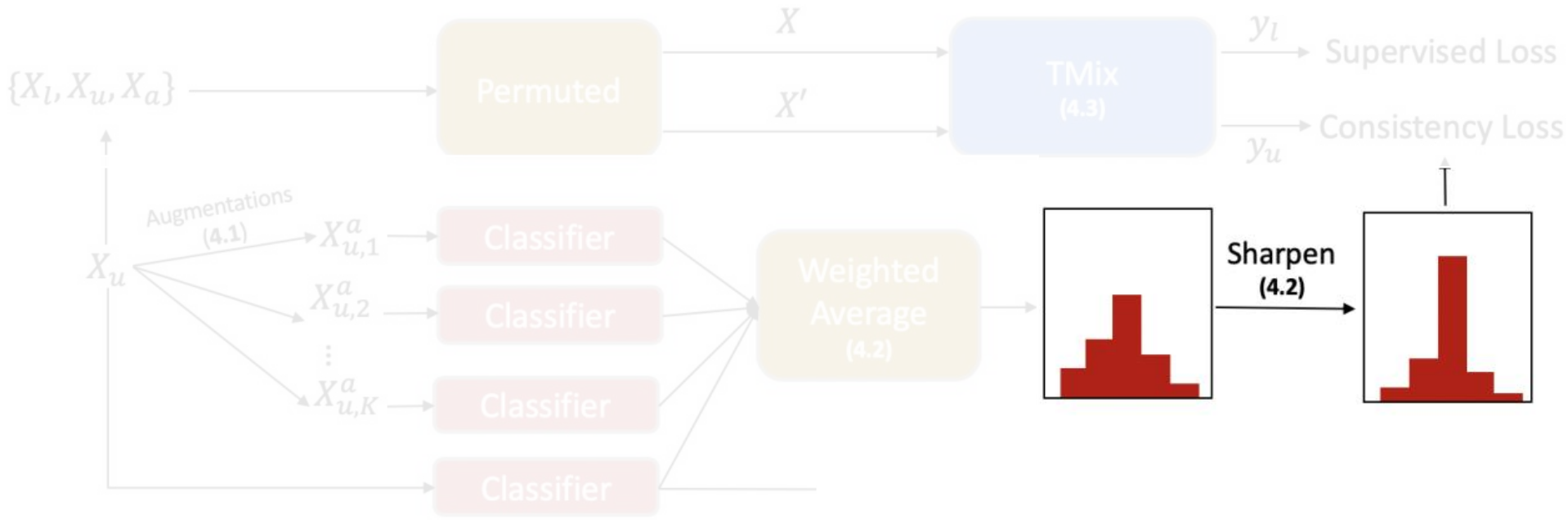

```
# Based on translation qualities, choose different weights here.
# For AG News: German: 1, Russian: 0, ori: 1
# For DBPedia: German: 1, Russian: 1, ori: 1
# For IMDB: German: 0, Russian: 0, ori: 1
# For Yahoo Answers: German: 1, Russian: 0, ori: 1 / German: 0, Russian: 0, ori: 1
p = (0 * torch.softmax(outputs_u, dim=1) + 0 * torch.softmax(outputs_u2,
                                                             dim=1) + 1 * torch.softmax(outputs_ori, dim=1)) / (1)
```

Method: 2. MixText

1. Label Guessing

Weighted average being uniform → Sharpening over predicted labels

$T \to 0$ Label become one-hot



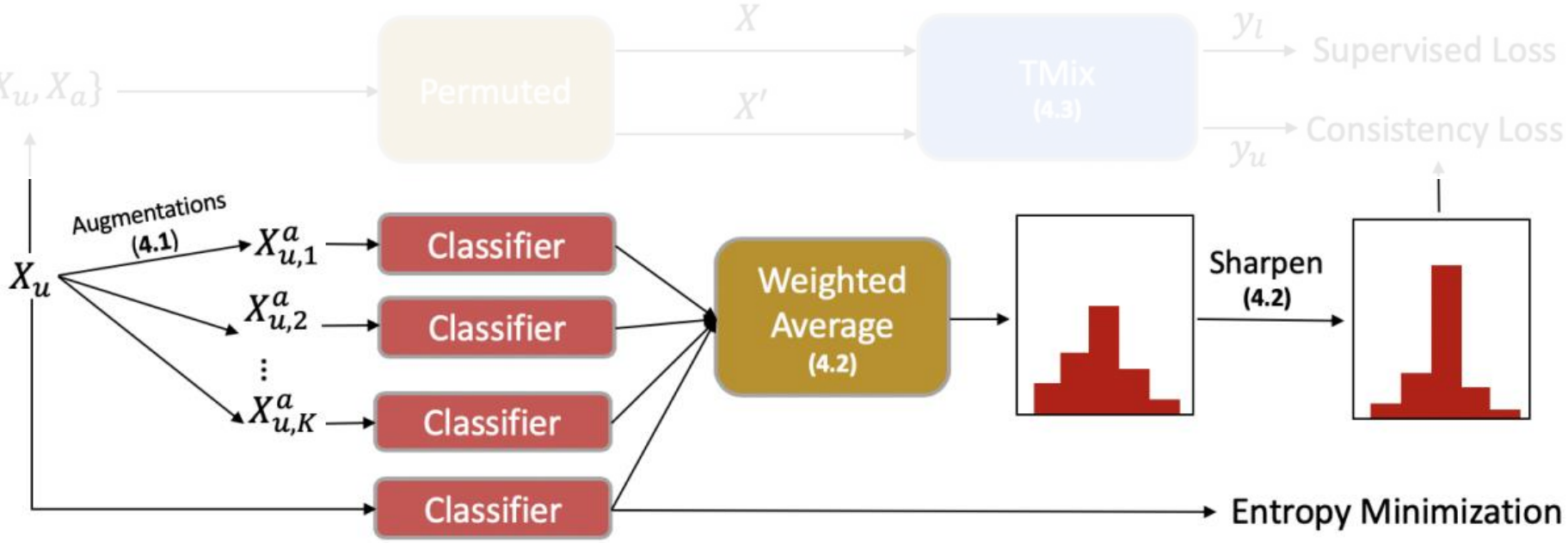
$$\text{Sharpen}(\mathbf{y}_i^u, T) = \frac{(\mathbf{y}_i^u)^{\frac{1}{T}}}{||(\mathbf{y}_i^u)^{\frac{1}{T}}||_1},$$

Method: 2. MixText

1. Label Guessing

Entropy minimization

확률 값과 어떠한 r과의 거리가 0이 되도록 함 ← confident한 확률 값을 얻기 위함

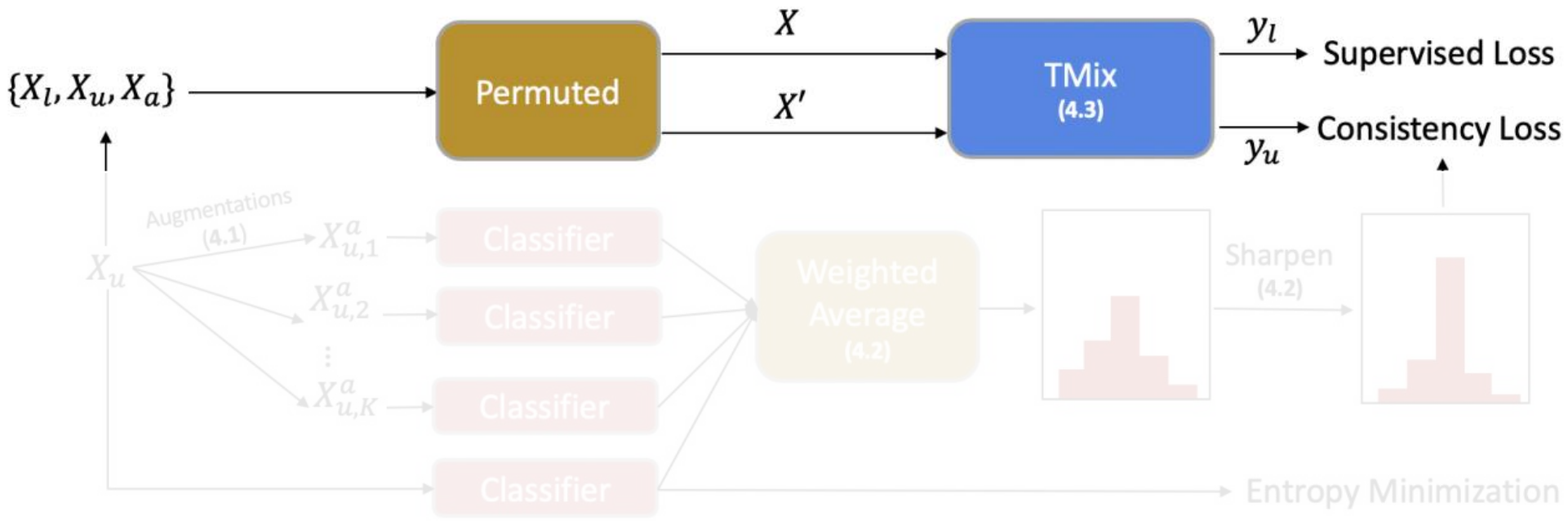


$$L_{\text{margin}} = \mathbb{E}_{\mathbf{x} \in \mathbf{x}_u} \max(0, \gamma - ||\mathbf{y}^u||_2^2),$$

Method: 2. MixText

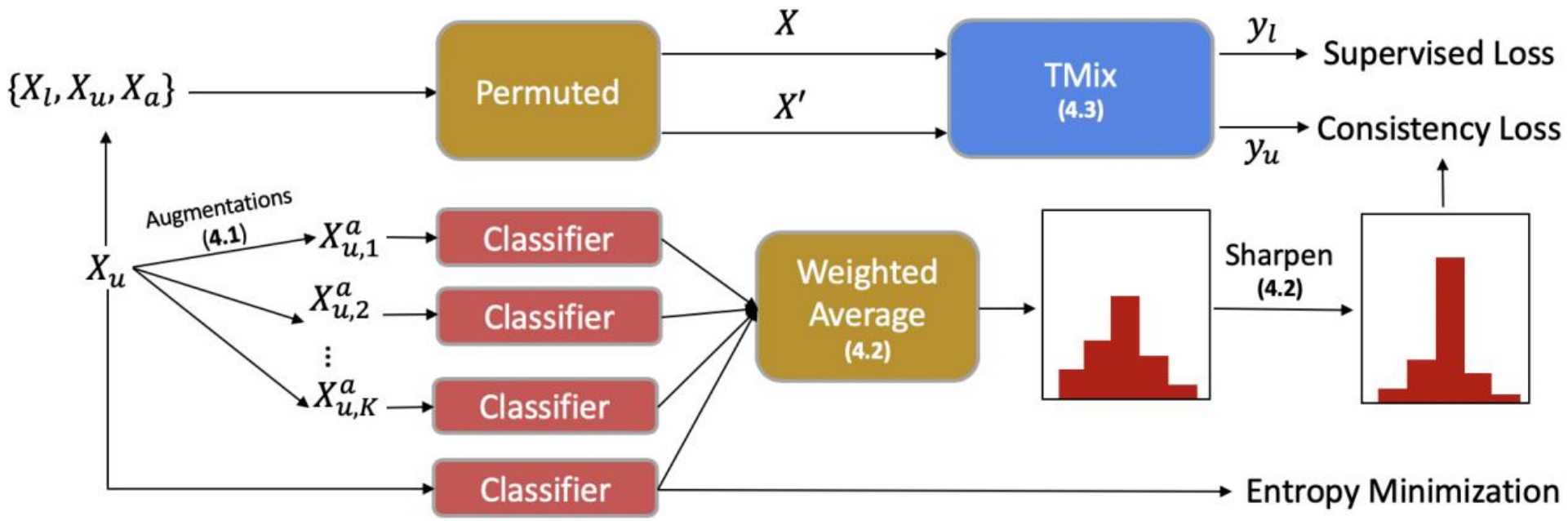
2. Tmix

Guessed label의 unlabeled data, augmentation data과 labeled data mixup



$$L_{\text{TMix}} = \mathbb{E}_{\mathbf{x},\mathbf{x}' \in \mathbf{X}} \text{KL}(\text{mix}(\mathbf{y}, \mathbf{y}') || p(\text{TMix}(\mathbf{x}, \mathbf{x}')).$$

MixText : TMix on Semi-Supervised Learning



$$L_{\text{MixText}} = L_{\text{TMix}} + \gamma_m L_{\text{margin}}.$$

Experiments

Datasets

Dataset	Label Type	Classes	Unlabeled	Dev	Test
AG News	News Topic	4	5000	2000	1900
DBpedia	Wikipeida Topic	14	5000	2000	5000
Yahoo! Answer	QA Topic	10	5000	5000	6000
IMDB	Review Sentiment	2	5000	2000	12500

Semi-supervised learning performance

Labeled data per class 10, 200, 2500 & Unlabeled data 5000

Datset	Model	10	200	2500	Dataset	Model	10	200	2500
AG News	VAMPIRE	-	83.9	86.2	DBpedia	VAMPIRE	-	-	-
	BERT	69.5	87.5	90.8		BERT	95.2	98.5	99.0
	TMix*	74.1	88.1	91.0		TMix*	96.8	98.7	99.0
	UDA	84.4	88.3	91.2		UDA	97.8	98.8	99.1
	MixText*	**88.4**	**89.2**	**91.5**		MixText*	**98.5**	**98.9**	**99.2**
Yahoo!	VAMPIRE	-	59.9	70.2	IMDB	VAMPIRE	-	82.2	85.8
	BERT	56.2	69.3	73.2		BERT	67.5	86.9	89.8
	TMix*	58.6	69.8	73.5		TMix*	69.3	87.4	90.3
	UDA	63.2	70.2	73.6		UDA	78.2	89.1	90.8
	MixText*	**67.6**	**71.3**	**74.1**		MixText*	**78.7**	**89.4**	**91.3**

Various number of Labeled Data

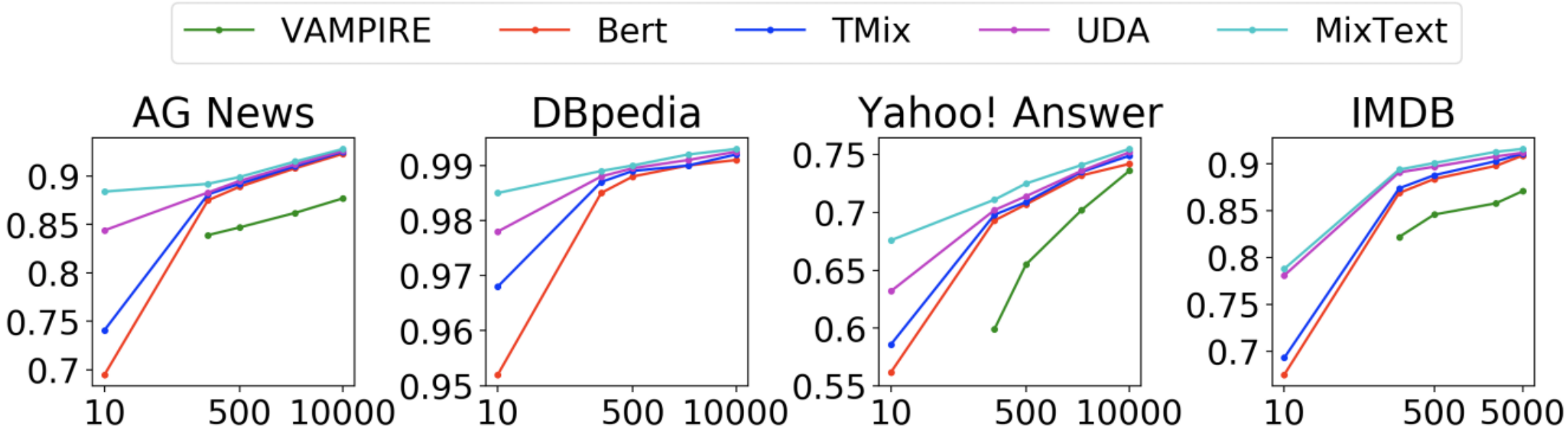
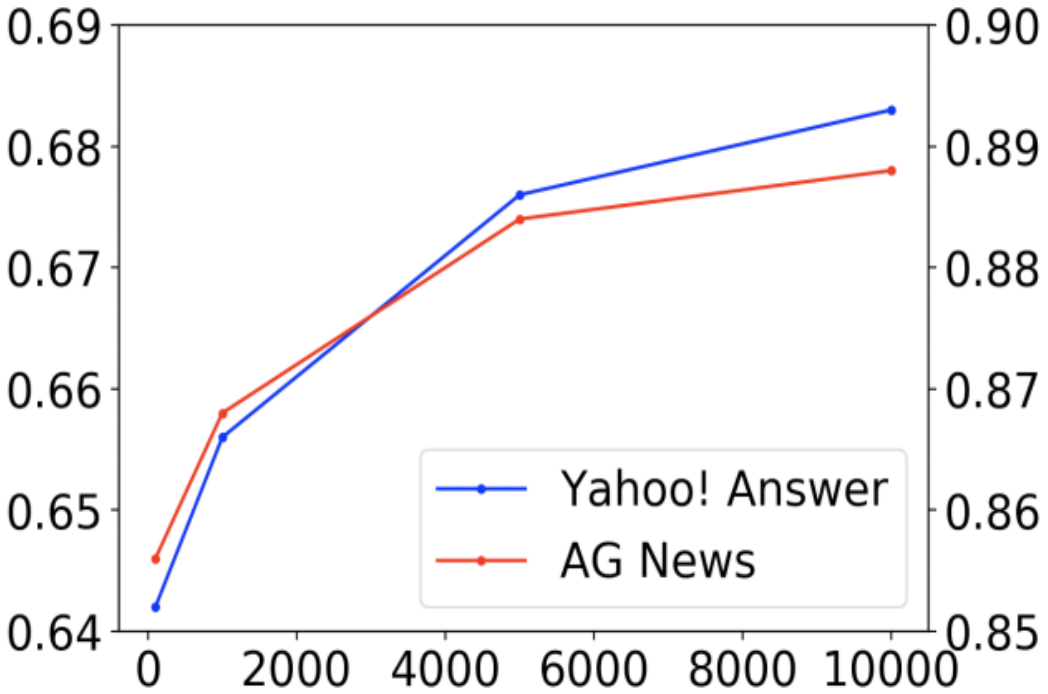


Figure 3: Performance (test accuracy (%)) on AG News, DBpedia, Yahoo! Answer and IMDB with 5000 unlabeled data and varying number of labeled data per class for each model.

Various number of Unlabeled Data



Unlabeled data가 많을 수록 성능이 좋아짐

모델이 Unlabeled data를 잘 활용하고 있다.

TMix Layer 선택

Mixup Layers Set	Accuracy(%)
∅	69.5
{0,1,2}	69.3
{3,4}	70.4
{6,7,9}	71.9
{7,9,12}	**74.1**
{6,7,9,12}	72.2
{3,4,6,7,9,12}	71.6

Table 3: Performance (test accuracy (%)) on AG News with 10 labeled data per class with different mixup layers set for TMix. ∅ means no mixup.

Ablation study

Model	Accuracy(%)
MixText	**67.6**
- weighted average	67.1
- TMix	63.5
- unlabeled data	58.6
- all	56.2

Table 4: Performance (test accuracy (%)) on Yahoo! Answer with 10 labeled data and 5000 unlabeled data per class after removing different parts of MixText.

Summary

1. TMix

    BERT hidden layer output을 mixup

2. MixText

    Semi-Supervised learning for text classification

    Tmix 응용한 모델

    back translation을 활용해 unlabeled data label guessing

Critiques

Strength

- 최초로 BERT hidden state에서 mixup 활용
- States-of-the-art

Weakness

- Back translation 품질에 의존하는 semi-supervised 모델

Question?

Thank You!