Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation

집현전 중급반

발표자 양수영

About

Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation

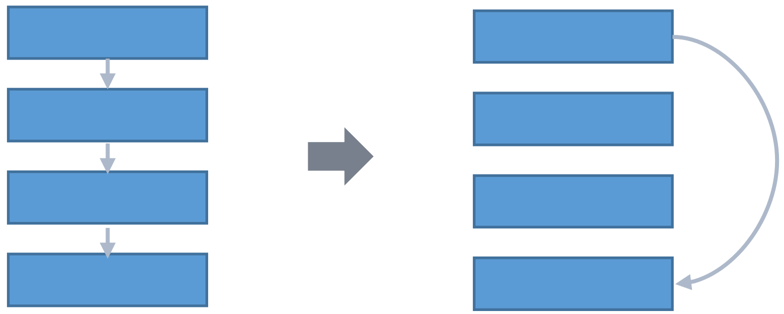
Nils Reimers and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

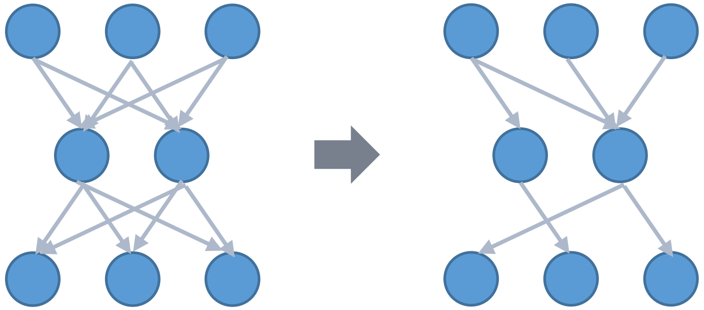
EMNLP 2020 Long paper

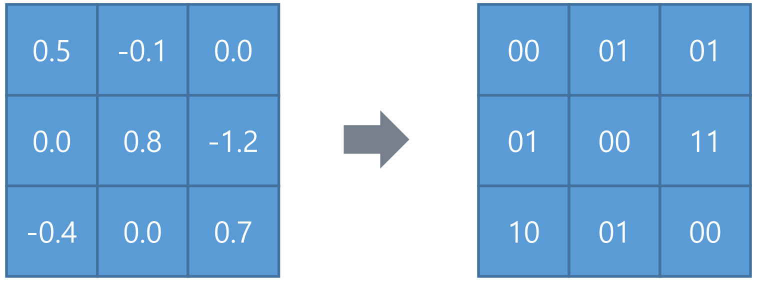
Volume: 2004.09813v2

Paper: https://arxiv.org/pdf/2004.09813.pdf

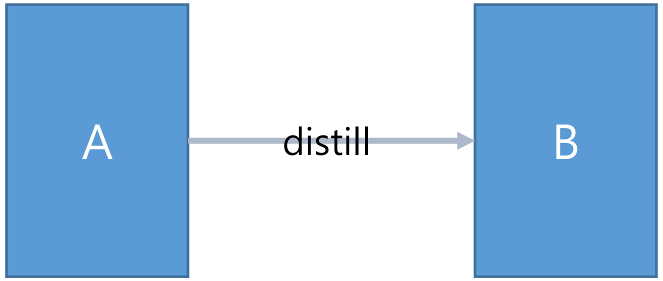
Deep Neural Network Lightweight


Compact Network Design


Pruning



Quantization



Knowledge Distillation

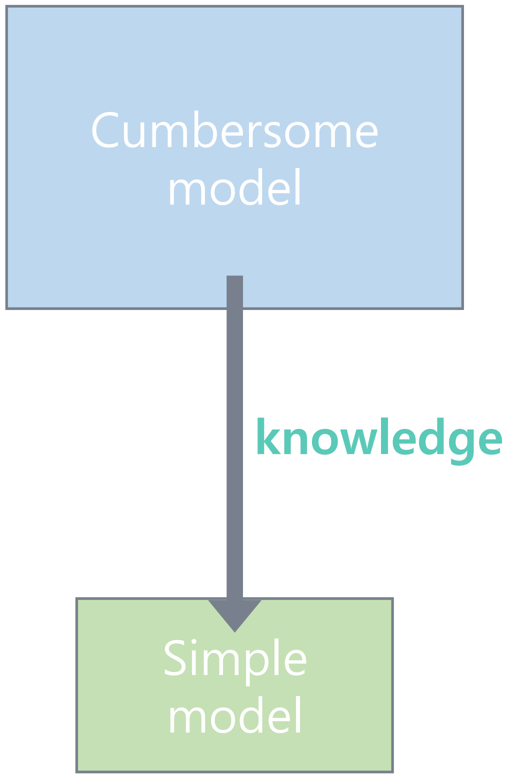
Knowledge Distillation

Proposer

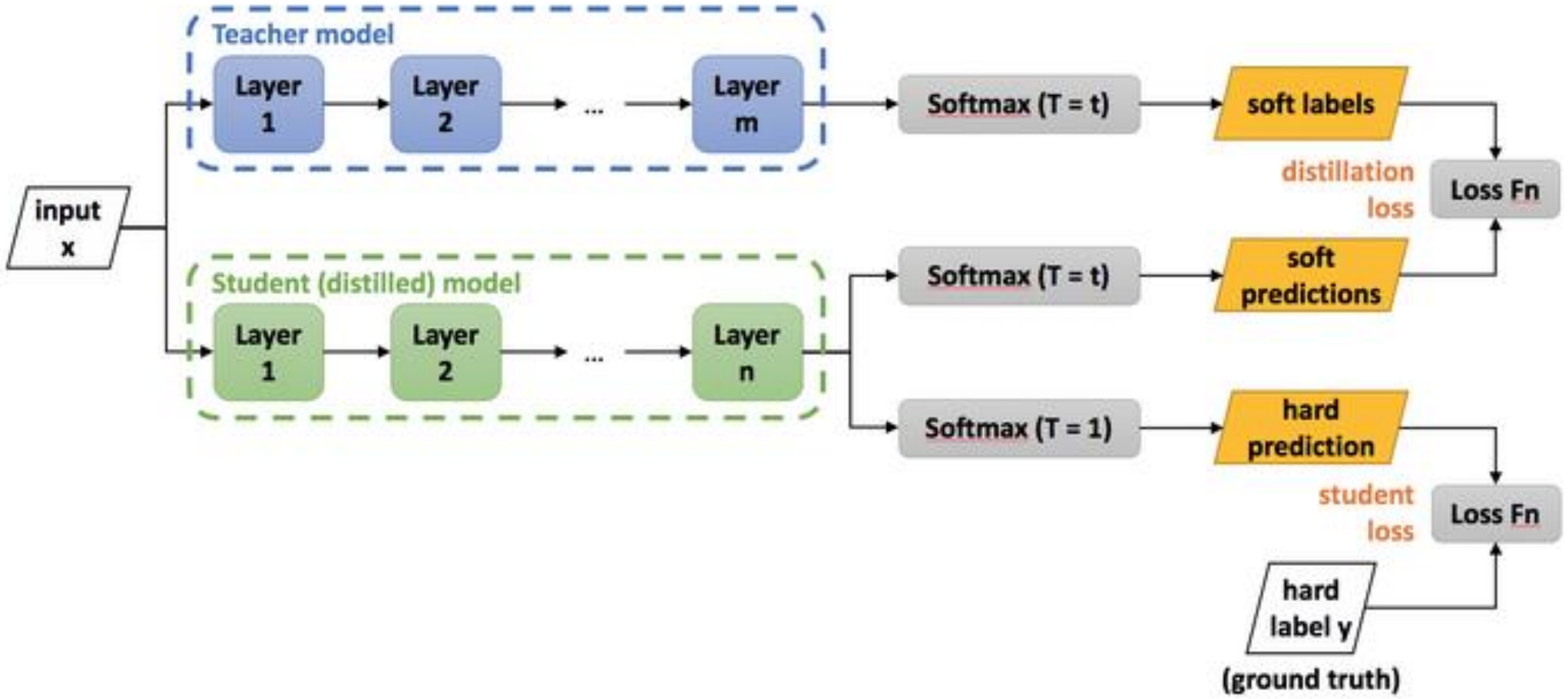
- Distilling the Knowledge in a Neural Network
- Geoffrey Hinton, Oriol Vinyals, Jeff Dean
- https://arxiv.org/pdf/1503.02531.pdf

Need

- Deployment to a large number of users has much more stringent requirements on latency and computational resources.

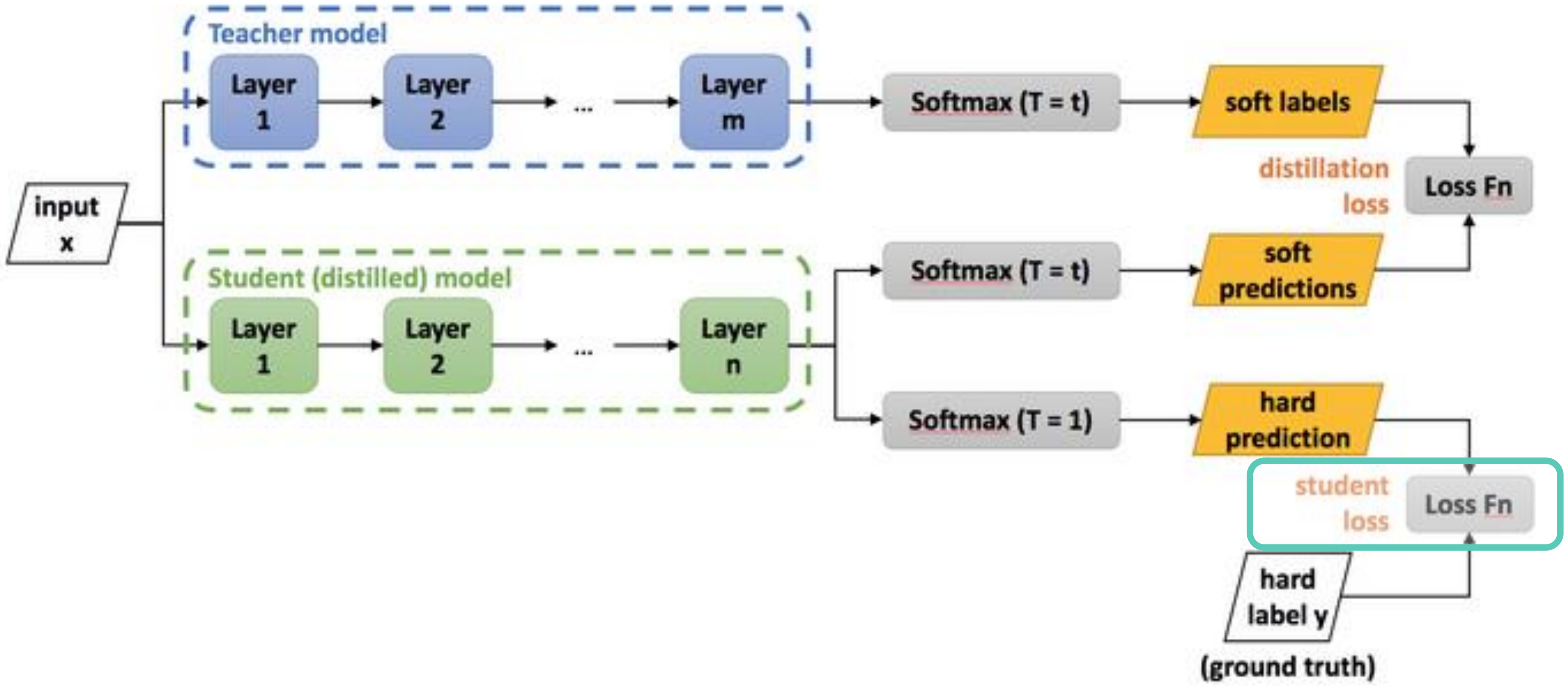


Knowledge Distillation



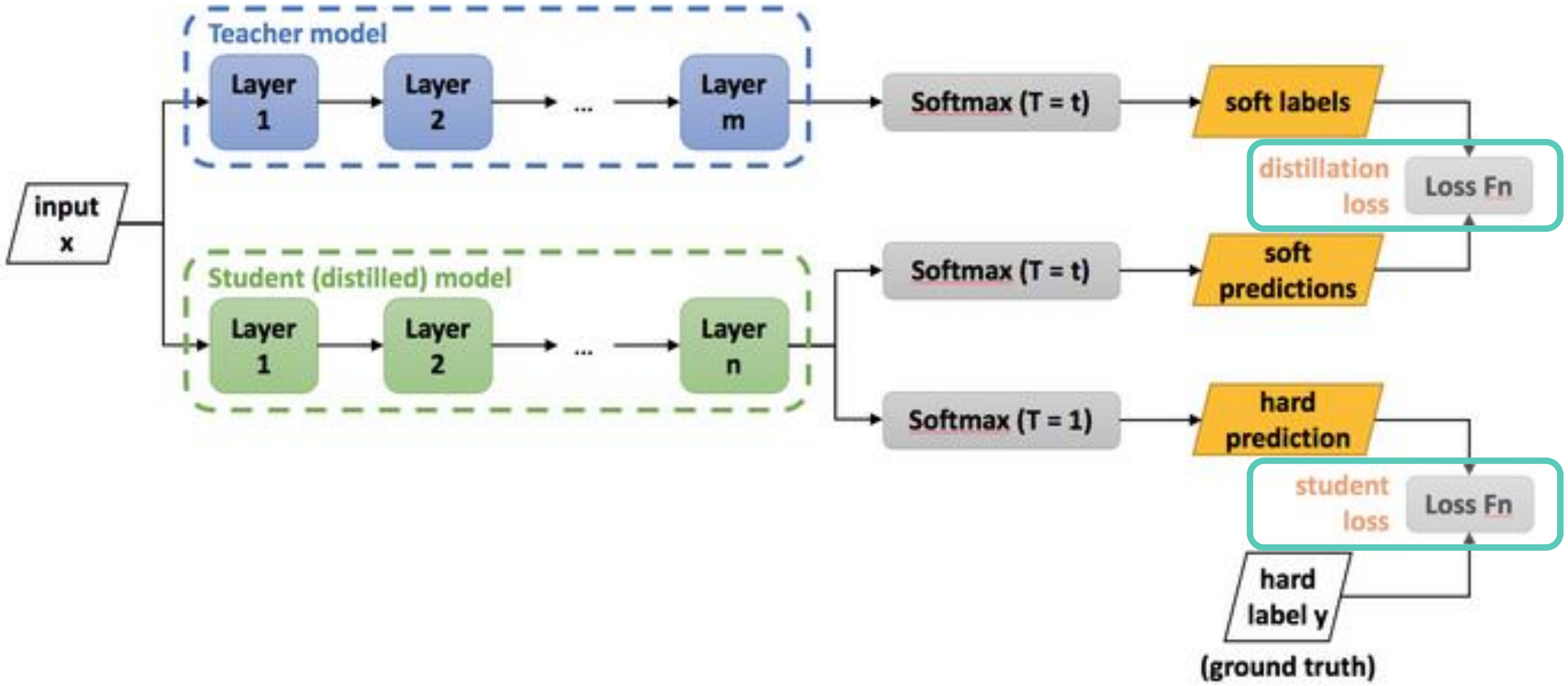
이미지 출처: https://nervanasystems.github.io/distiller/knowledge_distillation.html

Knowledge Distillation



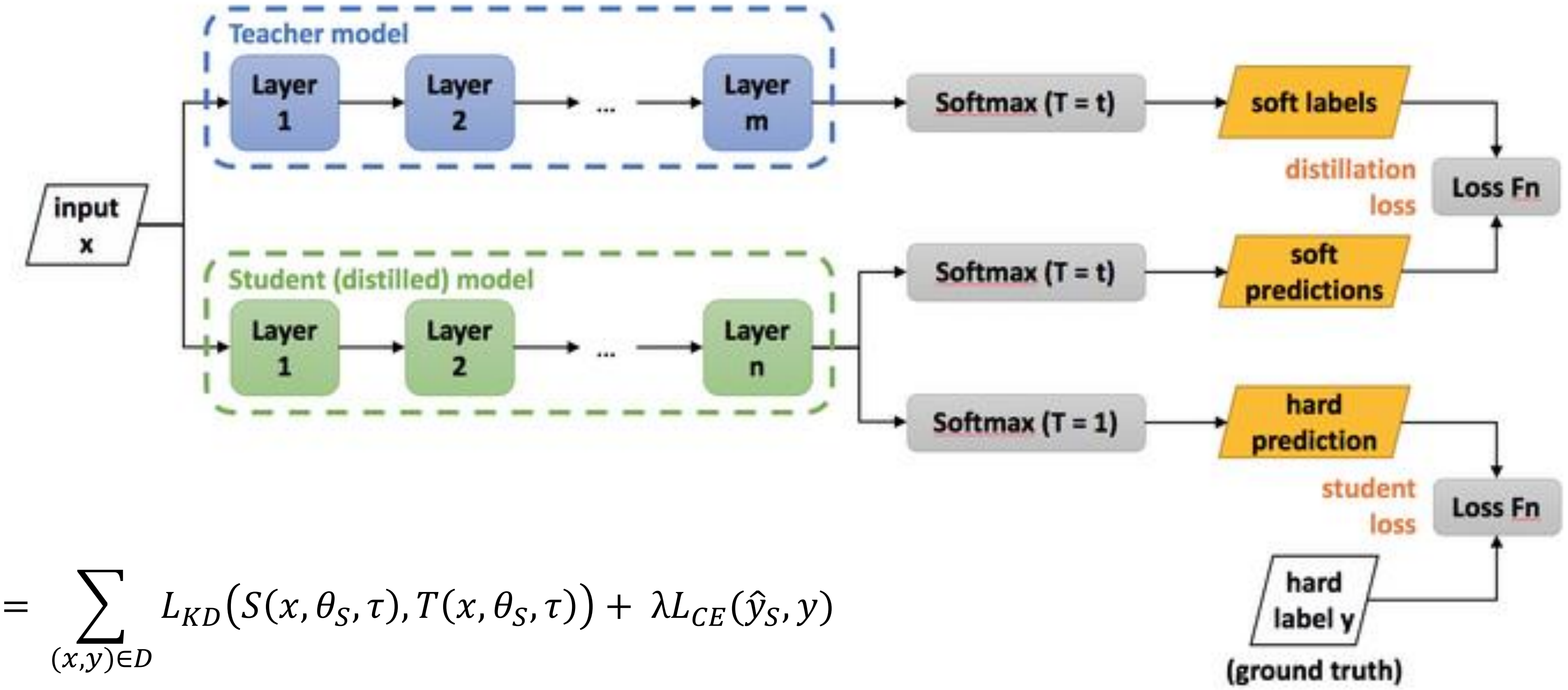
이미지 출처: https://nervanasystems.github.io/distiller/knowledge_distillation.html

Knowledge Distillation



이미지 출처: https://nervanasystems.github.io/distiller/knowledge_distillation.html

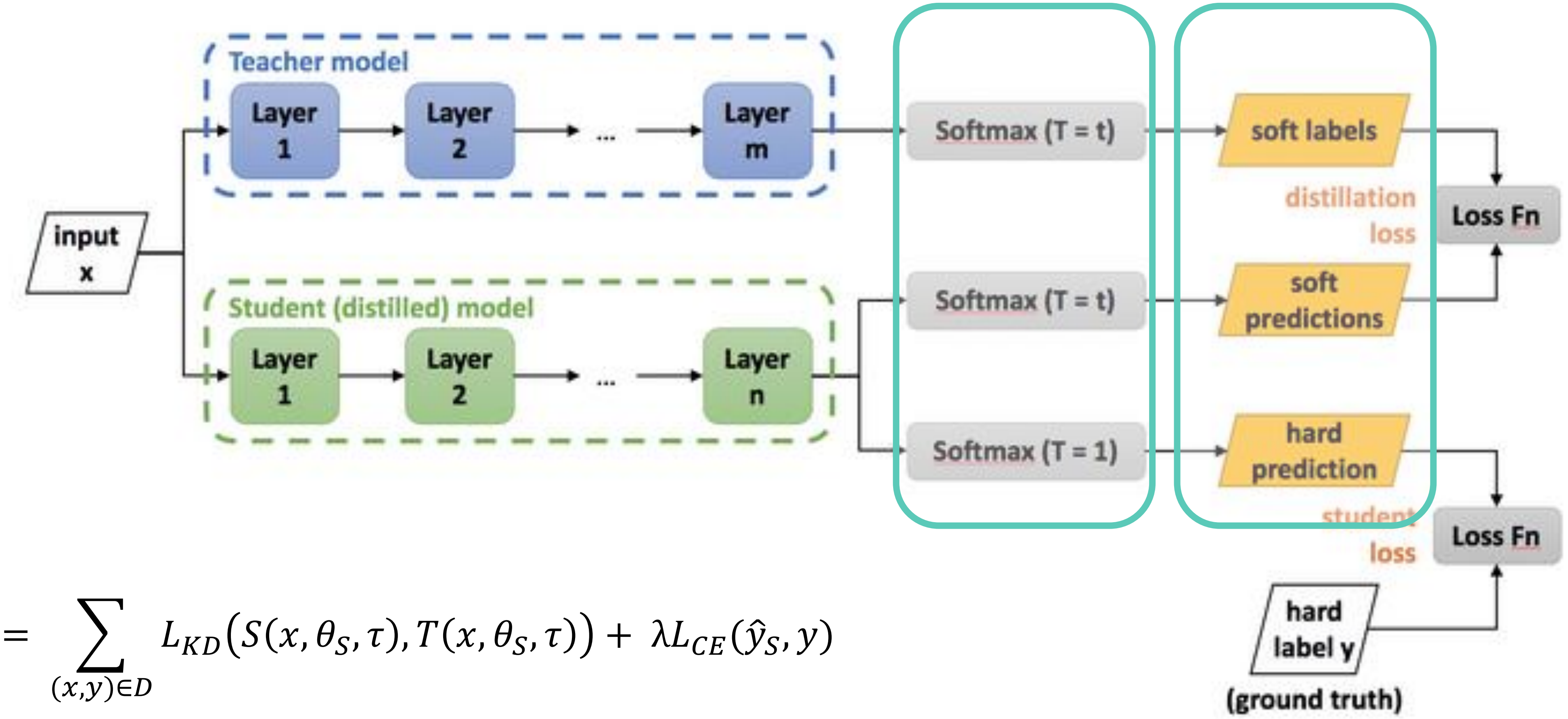
Knowledge Distillation



$$L = \sum_{(x,y)\in D} L_{KD}\big(S(x,\theta_S,\tau), T(x,\theta_S,\tau)\big) + \lambda L_{CE}(\hat{y}_S, y)$$

이미지 출처: https://nervanasystems.github.io/distiller/knowledge_distillation.html

Knowledge Distillation



$$L = \sum_{(x,y)\in D} L_{KD}\big(S(x,\theta_S,\tau), T(x,\theta_S,\tau)\big) + \lambda L_{CE}(\hat{y}_S, y)$$

이미지 출처: https://nervanasystems.github.io/distiller/knowledge_distillation.html

Knowledge Distillation

Hard label / Soft label

$$\begin{pmatrix} Bear \\ Cat \\ Dog \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad \begin{pmatrix} Bear \\ Cat \\ Dog \end{pmatrix} = \begin{pmatrix} 0.05 \\ 0.75 \\ 0.2 \end{pmatrix}$$

Softmax / Softer softmax

$$Softmax\begin{pmatrix} 1 \\ 2 \\ 9 \end{pmatrix} = \begin{pmatrix} 0.000335 \\ 0.000911 \\ 0.998754 \end{pmatrix} \qquad \underset{T=3}{Softmax}\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 0.059 \\ 0.083 \\ 0.857 \end{pmatrix}$$

$$q_i = \frac{\exp(z_i)}{\Sigma_j \exp(z_j)} \qquad q_i = \frac{\exp(z_i/T)}{\Sigma_j \exp(z_j/T)}$$

이미지 출처: https://light-tree.tistory.com/196

Introduction

Main Idea

- a translated sentence should be mapped to the same location in the vector space as the original sentence

Requirements

- Teacher Model M
- a set of parallel sentences $((s_1, t_1), \ldots, (s_n, t_n))$
- student model $\widehat{M}$ such that $\widehat{M}(s_i) \approx M(s_i)$ and $\widehat{M}(t_i) \approx M(s_i)$

Training

Use

- English SBERT model as teacher model
  : fine-tuned on English NLI and STS data

- XLM-RoBERTa model as student model
  : pre-trained 100 different languages

Meaning

- $\widehat{M} \leftarrow M$
  : the student model $\widehat{M}$ learns the representation of the teacher model M

Training

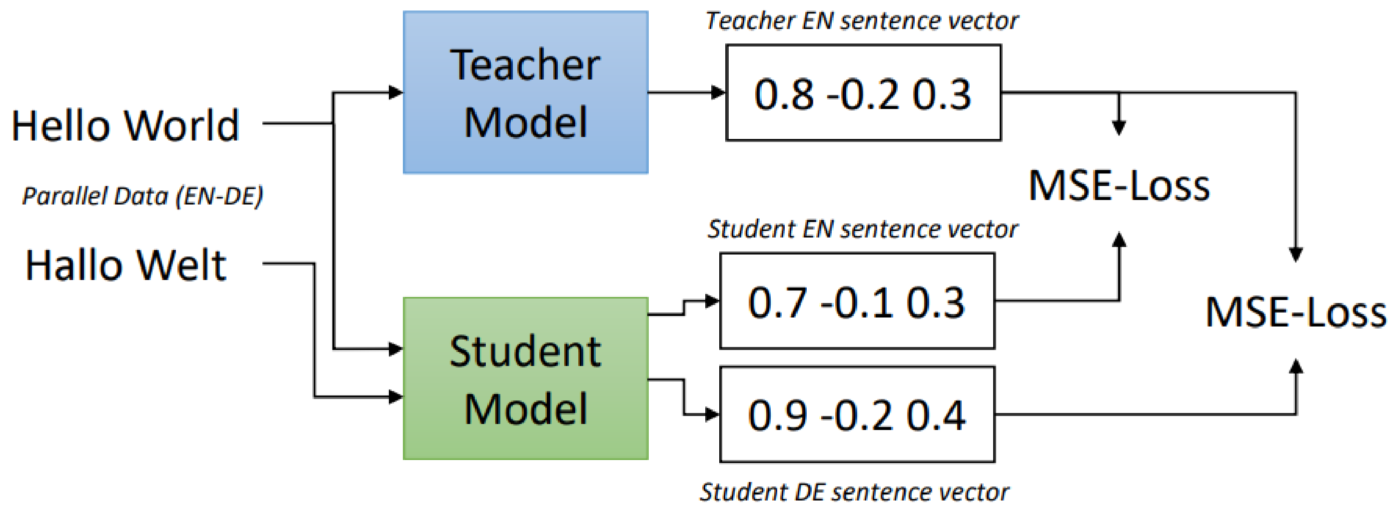


Figure 1: Given parallel data (e.g. English and German), train the student model such that the produced vectors for the English and German sentences are close to the teacher English sentence vector.

Training Data

Dataset

- GlobalVoices : news
- TED2020 : subtitles
- NewsCommentary : political and economic commentary
- WikiMatrix : parallel sentences from Wikipedia in different languages
- Tatoeba : a large database of example sentences and translations
- Europarl : the European Parliament website
- JW300 : magazines
- OpenSubtitles2018 : movie subtitles
- UNPC : United Nations documents

Bilingual Dictionaries

- MUSE
- Wikititles

Experiments

Multilingual Semantic Textual Similarity

BUCC: Bitext Retrieval

Tatoeba: Similarity Search

Multilingual Semantic Textual Similarity

Goal

- assign for a pair of sentences a score indicating their semantic similarity

Experiment

- compute cosine similarity
- compute the Spearman's rank correlation ρ between the computed score and the gold score

Multilingual Semantic Textual Similarity

Table 1

Model	EN-EN	ES-ES	AR-AR	Avg.
mBERT mean	54.4	56.7	50.9	54.0
XLM-R mean	50.7	51.8	25.7	42.7
mBERT-nli-stsb	80.2	83.9	65.3	76.5
XLM-R-nli-stsb	78.2	83.1	64.4	75.3
Knowledge Distillation				
mBERT ← SBERT-nli-stsb	82.5	83.0	78.8	81.4
DistilmBERT ← SBERT-nli-stsb	82.1	84.0	77.7	81.2
XLM-R ← SBERT-nli-stsb	82.5	83.5	79.9	82.0
XLM-R ← SBERT-paraphrases	88.8	86.3	79.6	**84.6**
Other Systems				
LASER	77.6	79.7	68.9	75.4
mUSE	86.4	86.9	76.4	83.2
LaBSE	79.4	80.8	69.1	76.4

Table 2

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
mBERT mean	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2
XLM-R mean	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6
Knowledge Distillation								
mBERT ← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6
DistilmBERT ← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0
XLM-R ← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	77.9
XLM-R ← SBERT-paraphrases	82.3	84.0	80.9	83.1	84.9	86.3	84.5	**83.7**
Other Systems								
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
mUSE	79.3	82.1	75.5	79.6	82.6	84.5	84.1	81.1
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5

Multilingual Semantic Textual Similarity

Table 1

Model	EN-EN	ES-ES	AR-AR	Avg.
mBERT mean	54.4	56.7	50.9	54.0
XLM-R mean	50.7	51.8	25.7	42.7
mBERT-nli-stsb	80.2	83.9	65.3	76.5
XLM-R-nli-stsb	78.2	83.1	64.4	75.3
Knowledge Distillation				
mBERT ← SBERT-nli-stsb	82.5	83.0	78.8	81.4
DistilmBERT ← SBERT-nli-stsb	82.1	84.0	77.7	81.2
XLM-R ← SBERT-nli-stsb	82.5	83.5	79.9	82.0
XLM-R ← SBERT-paraphrases	88.8	86.3	79.6	**84.6**
Other Systems				
LASER	77.6	79.7	68.9	75.4
mUSE	86.4	86.9	76.4	83.2
LaBSE	79.4	80.8	69.1	76.4

Table 2

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
mBERT mean	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2
XLM-R mean	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6
Knowledge Distillation								
mBERT ← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6
DistilmBERT ← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0
XLM-R ← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	77.9
XLM-R ← SBERT-paraphrases	82.3	84.0	80.9	83.1	84.9	86.3	84.5	**83.7**
Other Systems								
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
mUSE	79.3	82.1	75.5	79.6	82.6	84.5	84.1	81.1
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5

Multilingual Semantic Textual Similarity

Table 1

Model	EN-EN	ES-ES	AR-AR	Avg.
mBERT mean	54.4	56.7	50.9	54.0
XLM-R mean	50.7	51.8	25.7	42.7
mBERT-nli-stsb	80.2	83.9	65.3	76.5
XLM-R-nli-stsb	78.2	83.1	64.4	75.3
Knowledge Distillation				
mBERT ← SBERT-nli-stsb	82.5	83.0	78.8	81.4
DistilmBERT ← SBERT-nli-stsb	82.1	84.0	77.7	81.2
XLM-R ← SBERT-nli-stsb	82.5	83.5	79.9	82.0
XLM-R ← SBERT-paraphrases	88.8	86.3	79.6	**84.6**
Other Systems				
LASER	77.6	79.7	68.9	75.4
mUSE	86.4	86.9	76.4	83.2
LaBSE	79.4	80.8	69.1	76.4

Table 2

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
mBERT mean	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2
XLM-R mean	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6
Knowledge Distillation								
mBERT ← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6
DistilmBERT ← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0
XLM-R ← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	77.9
XLM-R ← SBERT-paraphrases	82.3	84.0	80.9	83.1	84.9	86.3	84.5	**83.7**
Other Systems								
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
mUSE	79.3	82.1	75.5	79.6	82.6	84.5	84.1	81.1
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5

Multilingual Semantic Textual Similarity

Table 1

Model	EN-EN	ES-ES	AR-AR	Avg.
mBERT mean	54.4	56.7	50.9	54.0
XLM-R mean	50.7	51.8	25.7	42.7
mBERT-nli-stsb	80.2	83.9	65.3	76.5
XLM-R-nli-stsb	78.2	83.1	64.4	75.3
Knowledge Distillation				
mBERT ← SBERT-nli-stsb	82.5	83.0	78.8	81.4
DistilmBERT ← SBERT-nli-stsb	82.1	84.0	77.7	81.2
XLM-R ← SBERT-nli-stsb	82.5	83.5	79.9	82.0
XLM-R ← SBERT-paraphrases	88.8	86.3	79.6	**84.6**
Other Systems				
LASER	77.6	79.7	68.9	75.4
mUSE	86.4	86.9	76.4	83.2
LaBSE	79.4	80.8	69.1	76.4

Table 2

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
mBERT mean	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2
XLM-R mean	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6
Knowledge Distillation								
mBERT ← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6
DistilmBERT ← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0
XLM-R ← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	77.9
XLM-R ← SBERT-paraphrases	82.3	84.0	80.9	83.1	84.9	86.3	84.5	**83.7**
Other Systems								
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
mUSE	79.3	82.1	75.5	79.6	82.6	84.5	84.1	81.1
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5

Multilingual Semantic Textual Similarity

Table 1

Model	EN-EN	ES-ES	AR-AR	Avg.
mBERT mean	54.4	56.7	50.9	54.0
XLM-R mean	50.7	51.8	25.7	42.7
mBERT-nli-stsb	80.2	83.9	65.3	76.5
XLM-R-nli-stsb	78.2	83.1	64.4	75.3
Knowledge Distillation				
mBERT ← SBERT-nli-stsb	82.5	83.0	78.8	81.4
DistilmBERT ← SBERT-nli-stsb	82.1	84.0	77.7	81.2
XLM-R ← SBERT-nli-stsb	82.5	83.5	79.9	82.0
XLM-R ← SBERT-paraphrases	88.8	86.3	79.6	**84.6**
Other Systems				
LASER	77.6	79.7	68.9	75.4
mUSE	86.4	86.9	76.4	83.2
LaBSE	79.4	80.8	69.1	76.4

Table 2

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
mBERT mean	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2
XLM-R mean	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6
Knowledge Distillation								
mBERT ← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6
DistilmBERT ← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0
XLM-R ← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	77.9
XLM-R ← SBERT-paraphrases	82.3	84.0	80.9	83.1	84.9	86.3	84.5	**83.7**
Other Systems								
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
mUSE	79.3	82.1	75.5	79.6	82.6	84.5	84.1	81.1
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5

Multilingual Semantic Textual Similarity

Table 1

Model	EN-EN	ES-ES	AR-AR	Avg.
mBERT mean	54.4	56.7	50.9	54.0
XLM-R mean	50.7	51.8	25.7	42.7
mBERT-nli-stsb	80.2	83.9	65.3	76.5
XLM-R-nli-stsb	78.2	83.1	64.4	75.3
Knowledge Distillation				
mBERT ← SBERT-nli-stsb	82.5	83.0	78.8	81.4
DistilmBERT ← SBERT-nli-stsb	82.1	84.0	77.7	81.2
XLM-R ← SBERT-nli-stsb	82.5	83.5	79.9	82.0
XLM-R ← SBERT-paraphrases	88.8	86.3	79.6	**84.6**
Other Systems				
LASER	77.6	79.7	68.9	75.4
mUSE	86.4	86.9	76.4	83.2
LaBSE	79.4	80.8	69.1	76.4

Table 2

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
mBERT mean	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2
XLM-R mean	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6
Knowledge Distillation								
mBERT ← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6
DistilmBERT ← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0
XLM-R ← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	77.9
XLM-R ← SBERT-paraphrases	82.3	84.0	80.9	83.1	84.9	86.3	84.5	**83.7**
Other Systems								
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
mUSE	79.3	82.1	75.5	79.6	82.6	84.5	84.1	81.1
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5

BUCC: Bitext Retrieval

Goal

- identify sentence pairs that are translations in two corpora in different languages

Experiments

- the BUCC bitext retrieval code from LASER with the scoring function
- x, y : sentence embeddings
- NNk(x) : k nearest neighbors
- margin(a, b) = a/b.

$$score(x, y) = margin(\cos(x, y), \sum_{z \in NNk(x)} \frac{\cos(x,z)}{2k} + \sum_{z \in NNk(y)} \frac{\cos(y,z)}{2k})$$

BUCC: Bitext Retrieval

Model	DE-EN	FR-EN	RU-EN	ZH-EN	Avg.
mBERT mean	44.1	47.2	38.0	37.4	41.7
XLM-R mean	5.2	6.6	22.1	12.4	11.6
mBERT-nli-stsb	38.9	39.5	26.4	30.2	33.7
XLM-R-nli-stsb	44.0	51.0	51.5	44.0	47.6
Knowledge Distillation					
XLM-R ← SBERT-nli-stsb	86.8	84.4	86.3	85.1	85.7
XLM-R ← SBERT-paraphrase	90.8	87.1	88.6	87.8	88.6
Other systems					
mUSE	88.5	86.3	89.1	86.9	87.7
LASER	95.4	92.4	92.3	91.7	93.0
LaBSE	95.9	92.5	92.4	93.0	93.5

Table 3: F_1 score on the BUCC bitext mining task.

BUCC: Bitext Retrieval

Model	DE-EN	FR-EN	RU-EN	ZH-EN	Avg.
mBERT mean	44.1	47.2	38.0	37.4	41.7
XLM-R mean	5.2	6.6	22.1	12.4	11.6
mBERT-nli-stsb	38.9	39.5	26.4	30.2	33.7
XLM-R-nli-stsb	44.0	51.0	51.5	44.0	47.6
Knowledge Distillation					
XLM-R $\leftarrow$ SBERT-nli-stsb	86.8	84.4	86.3	85.1	85.7
XLM-R $\leftarrow$ SBERT-paraphrase	90.8	87.1	88.6	87.8	88.6
Other systems					
mUSE	88.5	86.3	89.1	86.9	87.7
LASER	95.4	92.4	92.3	91.7	93.0
LaBSE	95.9	92.5	92.4	93.0	93.5

Table 3: F_1 score on the BUCC bitext mining task.

BUCC: Bitext Retrieval

Model	DE-EN	FR-EN	RU-EN	ZH-EN	Avg.
mBERT mean	44.1	47.2	38.0	37.4	41.7
XLM-R mean	5.2	6.6	22.1	12.4	11.6
mBERT-nli-stsb	38.9	39.5	26.4	30.2	33.7
XLM-R-nli-stsb	44.0	51.0	51.5	44.0	47.6
Knowledge Distillation					
XLM-R ← SBERT-nli-stsb	86.8	84.4	86.3	85.1	85.7
XLM-R ← SBERT-paraphrase	90.8	87.1	88.6	87.8	88.6
Other systems					
mUSE	88.5	86.3	89.1	86.9	87.7
LASER	95.4	92.4	92.3	91.7	93.0
LaBSE	95.9	92.5	92.4	93.0	93.5

Table 3: F_1 score on the BUCC bitext mining task.

BUCC: Bitext Retrieval

Sentence1

- Olympischen Jugend-Sommerspiele fanden vom 16. bis 28. August 2014 in Nanjing (China) statt.
  : 하계 청소년 올림픽이 2014년 8월 16일부터 28일까지 중국 난징에서 개최되었습니다.

Sentence2

- China hosted the 2014 Youth Olympic Games.
  : 중국은 2014년 청소년 올림픽을 개최하였습니다.

Tatoeba: Similarity Search

Goal

- lower resource languages can be especially challenging to get well-aligned sentence embeddings

Experiments

- Tatoeba test setup from LASER
- finding for all sentences the most similar sentence in the other language using cosine similarity
- computed for both directions

Tatoeba: Similarity Search

Language and Size

- KA : Georgian (296K)
- SW : Swahili (173K)
- TL : Tagalog (36K)
- TT : Tatar (119K)

Model	KA	SW	TL	TT
LASER				
en $\rightarrow$ xx	39.7	54.4	52.6	28.0
xx $\rightarrow$ en	32.2	60.8	48.5	34.3
XLM-R $\leftarrow$ SBERT-nli-stsb				
en $\rightarrow$ xx	73.1	85.4	86.2	54.5
xx $\rightarrow$ en	71.7	86.7	84.0	52.3

Table 4: Accuracy on the Tatoeba test set in both directions (en to target language and vice versa).

Evaluation of Training Datasets

Dataset	#DE	EN-DE	#AR	EN-AR
XLM-R mean	-	21.3	-	17.4
XLM-R-nli-stsb	-	59.5	-	44.0
MUSE Dict	101k	75.8	27k	68.8
Wikititles Dict	545k	71.4	748k	67.9
MUSE + Wikititles	646k	76.0	775k	69.1
GlobalVoices	37k	78.1	29k	68.6
TED2020	483k	80.4	774k	78.0
NewsCommentary	118k	77.7	7k	57.4
WikiMatrix	276k	79.4	385k	75.4
Tatoeba	303k	79.5	27k	76.7
Europarl	736k	78.7	-	-
JW300	1,399k	80.0	382k	74.0
UNPC	-	-	8M	66.1
OpenSubtitles	21M	79.8	28M	78.8
All datasets	25M	81.4	38M	79.0

Table 5: Data set sizes for the EN-DE / EN-AR sections. Performance (Spearman rank correlation) of XLM-R ← SBERT-nli-stsb on the STS 2017 dataset.

Dataset size	EN-DE	EN-AR
XLM-R mean	21.3	17.4
XLM-R-nli-stsb	59.5	44.0
1k	71.5	48.4
5k	74.5	59.6
10k	77.0	69.5
25k	80.0	70.2
Full TED2020	80.4	78.0

Table 6: Performance on STS 2017 dataset when trained with reduced TED2020 dataset sizes.

Evaluation of Training Datasets

Dataset	#DE	EN-DE	#AR	EN-AR
XLM-R mean	-	21.3	-	17.4
XLM-R-nli-stsb	-	59.5	-	44.0
MUSE Dict	101k	75.8	27k	68.8
Wikititles Dict	545k	71.4	748k	67.9
MUSE + Wikititles	646k	76.0	775k	69.1
GlobalVoices	37k	78.1	29k	68.6
TED2020	483k	80.4	774k	78.0
NewsCommentary	118k	77.7	7k	57.4
WikiMatrix	276k	79.4	385k	75.4
Tatoeba	303k	79.5	27k	76.7
Europarl	736k	78.7	-	-
JW300	1,399k	80.0	382k	74.0
UNPC	-	-	8M	66.1
OpenSubtitles	21M	79.8	28M	78.8
All datasets	25M	81.4	38M	79.0

Table 5: Data set sizes for the EN-DE / EN-AR sections. Performance (Spearman rank correlation) of XLM-R ← SBERT-nli-stsb on the STS 2017 dataset.

Dataset size	EN-DE	EN-AR
XLM-R mean	21.3	17.4
XLM-R-nli-stsb	59.5	44.0
1k	71.5	48.4
5k	74.5	59.6
10k	77.0	69.5
25k	80.0	70.2
Full TED2020	80.4	78.0

Table 6: Performance on STS 2017 dataset when trained with reduced TED2020 dataset sizes.

Evaluation of Training Datasets

Dataset	#DE	EN-DE	#AR	EN-AR
XLM-R mean	-	21.3	-	17.4
XLM-R-nli-stsb	-	59.5	-	44.0
MUSE Dict	101k	75.8	27k	68.8
Wikititles Dict	545k	71.4	748k	67.9
MUSE + Wikititles	646k	76.0	775k	69.1
GlobalVoices	37k	78.1	29k	68.6
TED2020	483k	80.4	774k	78.0
NewsCommentary	118k	77.7	7k	57.4
WikiMatrix	276k	79.4	385k	75.4
Tatoeba	303k	79.5	27k	76.7
Europarl	736k	78.7	-	-
JW300	1,399k	80.0	382k	74.0
UNPC	-	-	8M	66.1
OpenSubtitles	21M	79.8	28M	78.8
All datasets	25M	81.4	38M	79.0

Table 5: Data set sizes for the EN-DE / EN-AR sections. Performance (Spearman rank correlation) of XLM-R ← SBERT-nli-stsb on the STS 2017 dataset.

Dataset size	EN-DE	EN-AR
XLM-R mean	21.3	17.4
XLM-R-nli-stsb	59.5	44.0
1k	71.5	48.4
5k	74.5	59.6
10k	77.0	69.5
25k	80.0	70.2
Full TED2020	80.4	78.0

Table 6: Performance on STS 2017 dataset when trained with reduced TED2020 dataset sizes.

Evaluation of Training Datasets

Dataset	#DE	EN-DE	#AR	EN-AR
XLM-R mean	-	21.3	-	17.4
XLM-R-nli-stsb	-	59.5	-	44.0
MUSE Dict	101k	75.8	27k	68.8
Wikititles Dict	545k	71.4	748k	67.9
MUSE + Wikititles	646k	76.0	775k	69.1
GlobalVoices	37k	78.1	29k	68.6
TED2020	483k	80.4	774k	78.0
NewsCommentary	118k	77.7	7k	57.4
WikiMatrix	276k	79.4	385k	75.4
Tatoeba	303k	79.5	27k	76.7
Europarl	736k	78.7	-	-
JW300	1,399k	80.0	382k	74.0
UNPC	-	-	8M	66.1
OpenSubtitles	21M	79.8	28M	78.8
All datasets	25M	81.4	38M	79.0

Table 5: Data set sizes for the EN-DE / EN-AR sections. Performance (Spearman rank correlation) of XLM-R ← SBERT-nli-stsb on the STS 2017 dataset.

Dataset size	EN-DE	EN-AR
XLM-R mean	21.3	17.4
XLM-R-nli-stsb	59.5	44.0
1k	71.5	48.4
5k	74.5	59.6
10k	77.0	69.5
25k	80.0	70.2
Full TED2020	80.4	78.0

Table 6: Performance on STS 2017 dataset when trained with reduced TED2020 dataset sizes.

Target Language Training

Goal

- evaluate whether it is better to transfer an English model to a certain target language or if training from-scratch on suitable datasets in the target language yields better results

Experiments

- Kor NLI, Kor STS
- fine-tuned Korean RoBERTa and XLM-R on these datasets using the SBERT framework
- tuned XLM-R using multilingual knowledge distillation

Target Language Training

Model	KO-KO
LASER	68.44
mUSE	76.32
Trained on KorNLI & KorSTS	
Korean RoBERTa-base	80.29
Korean RoBERTa-large	80.49
XLM-R	79.19
XLM-R-large	81.84
Multiling. Knowledge Distillation	
XLM-R ← SBERT-nli-stsb	81.47
XLM-R-large ← SBERT-large-nli-stsb	83.00

Table 7: Spearman rank correlation on Korean STS-benchmark test-set (Ham et al., 2020).

Target Language Training

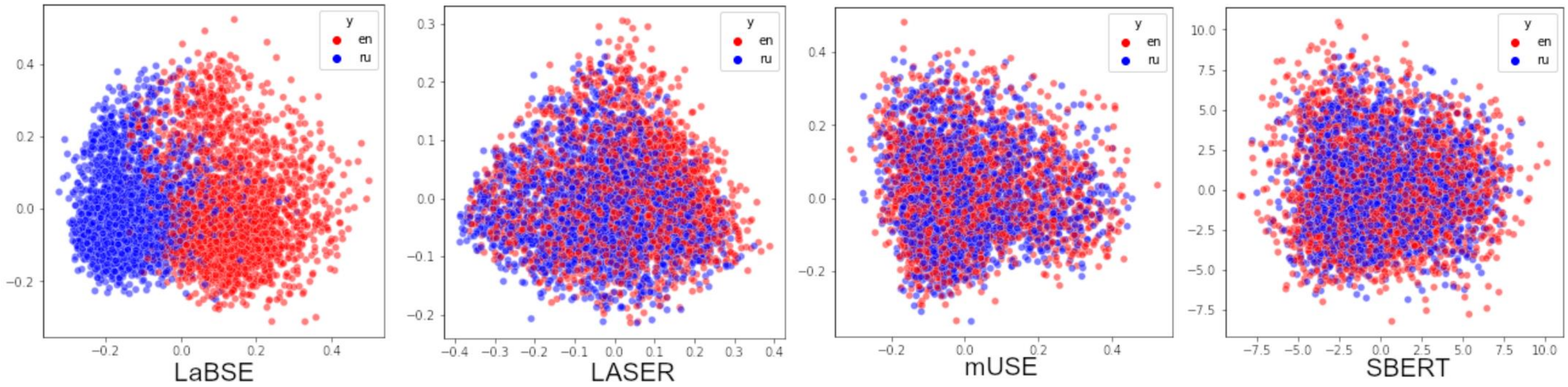
Model	KO-KO
LASER	68.44
mUSE	76.32
Trained on KorNLI & KorSTS	
Korean RoBERTa-base	80.29
Korean RoBERTa-large	80.49
XLM-R	79.19
XLM-R-large	81.84
Multiling. Knowledge Distillation	
XLM-R ← SBERT-nli-stsb	81.47
XLM-R-large ← SBERT-large-nli-stsb	83.00

Table 7: Spearman rank correlation on Korean STS-benchmark test-set (Ham et al., 2020).

Language Bias

Language Bias

- a model prefers one language or language pair over others



Language Bias

Model	Expected Score	Actual Score	Difference
LASER	69.5	68.6	-0.92
mUSE	81.7	81.6	-0.19
LaBSE	74.4	73.1	-1.29
XLM-R ← SBERT-paraphrases	84.0	83.9	-0.11

Table 8: Spearman rank correlation for the multilingual STS dataset. Expected score is the average over the performance on the individual sets (Table 1 & 2). Actual score is the correlation for one joined set of sentence pairs. Models without language bias would score on the joined set similar to the average over the individual sets. The difference shows the negative impact from the language bias.

Conclusion

Contribution

- Make monolingual sentence embeddings multilingual with aligned vector spaces between the languages

- Simplifies the training procedure compared to previous approaches

- Minimizes the potential language bias of the resulting model

Reference

- https://www.researchgate.net/publication/343574829_dib_leoning_model-ui_gyeonglyanghwa_gisul_donghyang

- https://baeseongsu.github.io/posts/knowledge-distillation/

- https://light-tree.tistory.com/196

THANK YOU ☺