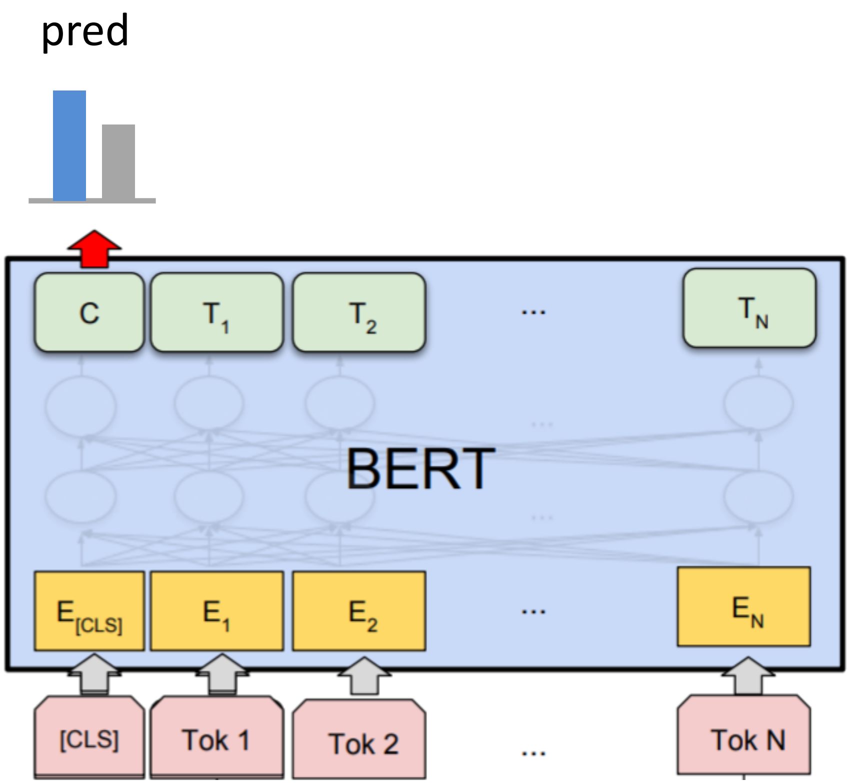
Contrastive Out-of-Distribution Detection for Pretrained Transformers

Wenxuan Zhou[1], Fangyu Liu[2], Muhao Chen[1]
University of Southern California[1], University of Cambridge[2]

Classification by Pretrained Transformers

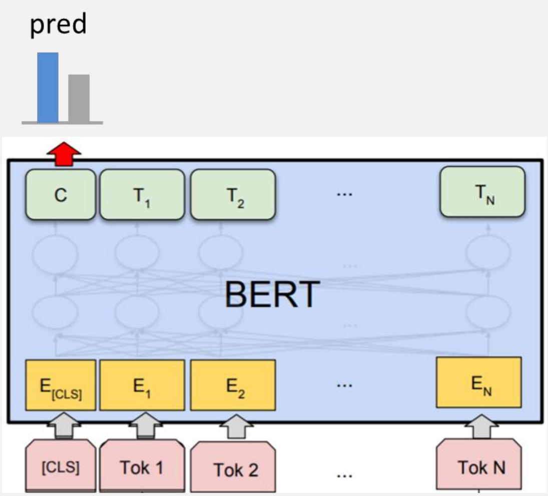


Source: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Out-of-Distribution (OOD) Instances

In real-world applications, instances from unknown classes may be present, in which case we need to identify and reject them.

Binary sentiment classifier



I like every minute of this movie.

Positive

We watched this movie last night.

Unknown
Reject this instance

Task Definition

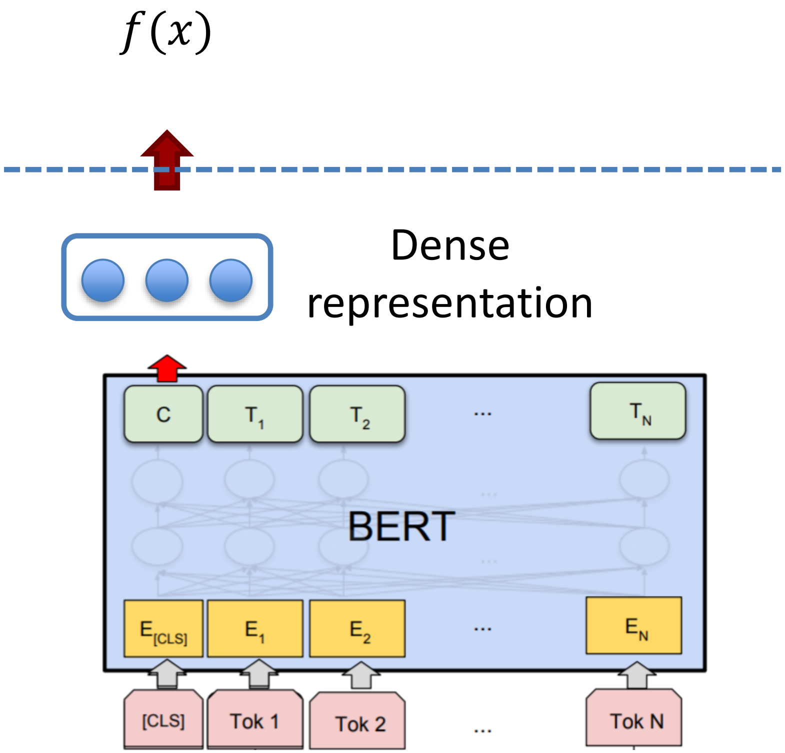
(OOD Definition) OOD instances are instances (x, y) sampled from a different distribution to the distribution of training data $P(X_{train}, Y_{train})$, where X_{train} and Y_{train} are the training corpus and the training label set.

- Non-semantic shift: $x \notin X_{train}, y \in Y_{train}$, e.g., a product review for a sentiment classifier trained on movie reviews.

- **Semantic shift (our focus)**: $y \notin Y_{train}$, unknown classes.

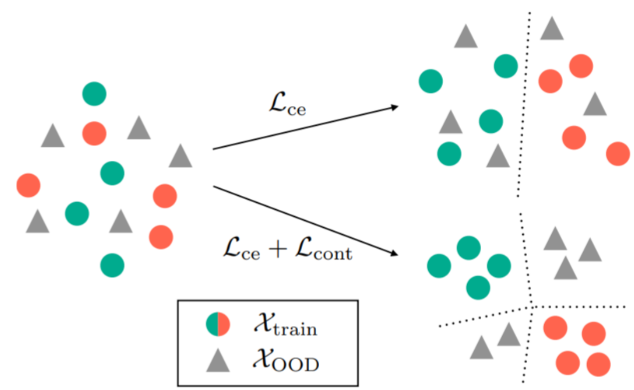
Our goal:
1. Get an OOD scoring function that returns a high score for OOD.
2. Maintain classification performance on the main task.
3. Unsupervised. Only use in-distribution data in training.

Overview



Scoring functions:
Transform dense representations to OOD scores.

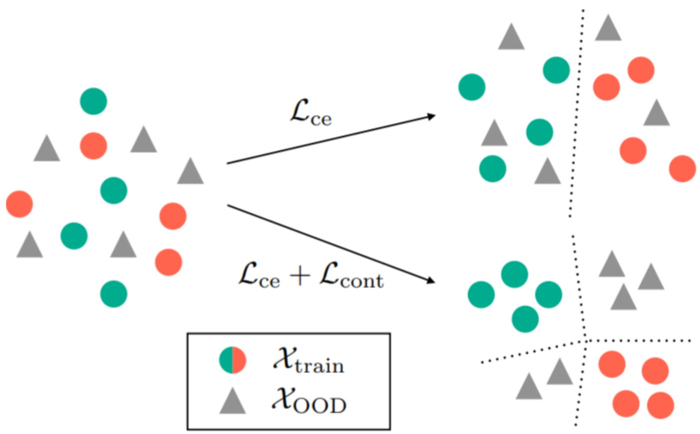
Representation Learning:
Learn a representation space where in-distribution and OOD data are well separated.



Contrastive Representation Learning

Motivation: for a random training instance x, instances from the same class can be seen as "in-distribution", while instances from other classes can be seen as "OOD".

Increase inter-class discrepancy ⇒ Better OOD detection



L_{cont}: Embedding instances of the same class closer and separating different classes.

Contrastive Representation Learning (Cont.)

Supervised Contrastive Loss:

$$\mathcal{L}_{\text{scl}} = \sum_{i=1}^{M} \frac{-1}{M|P(i)|} \sum_{p \in P(i)} \log \frac{e^{\boldsymbol{z}_i^\intercal \boldsymbol{z}_p/\tau}}{\sum_{a \in A(i)} e^{\boldsymbol{z}_i^\intercal \boldsymbol{z}_a/\tau}}$$

Cosine similarity

Optimize the ratio, easily saturate

Margin-based Contrastive Loss:

$$\mathcal{L}_{\text{pos}} = \sum_{i=1}^{M} \frac{1}{|P(i)|} \sum_{p \in P(i)} \|\boldsymbol{h}_i - \boldsymbol{h}_p\|^2,$$

L2 distance

$$\mathcal{L}_{\text{neg}} = \sum_{i=1}^{M} \frac{1}{|N(i)|} \sum_{n \in N(i)} (\xi - \|\boldsymbol{h}_i - \boldsymbol{h}_n\|^2)_+,$$

$$\mathcal{L}_{\text{margin}} = \frac{1}{dM} \left(\mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}}\right).$$

Optimize the margin, as compact as possible

Scoring Functions

Maximum Softmax Probability (baseline):

$$1 - \max_{j=1}^{C} \boldsymbol{p}_j$$

Mahalanobis Distance:

Fit the representation space with a multivariate Gaussian distribution. Use the probability density function as the OOD score.

$$\begin{gathered}
\boldsymbol{\mu}_j = \mathbb{E}_{y_i=j}\left[\boldsymbol{h}_i\right], j = 1, \ldots, C, \\
\boldsymbol{\Sigma} = \mathbb{E}\left[\left(\boldsymbol{h}_i - \boldsymbol{\mu}_{y_i}\right)\left(\boldsymbol{h}_i - \boldsymbol{\mu}_{y_i}\right)^{\intercal}\right], \\
g = -\min_{j=1}^{C}(\boldsymbol{h} - \boldsymbol{\mu}_j)^{\intercal}\boldsymbol{\Sigma}^{+}(\boldsymbol{h} - \boldsymbol{\mu}_j)
\end{gathered}$$

We tried other 2 scoring functions. For space limitation we don't put them here.

Experiments (Main)

Use different tasks as in-distribution and OOD data.

Tasks: sentiment analysis (SST2, IMDB), topic classification (20 Newsgroups), Question classification (TREC-10)
Additional OOD datasets: RTE, MNLI, WMT16, Multi30K

AUROC ↑ / FAR95 ↓		Avg	SST2	IMDB	TREC-10	20NG
w/o $\mathcal{L}_{\text{cont}}$	MSP	94.1 / 35.0	88.9 / 61.3	94.7 / 40.6	98.1 / 7.6	94.6 / 30.5
	Energy	94.0 / 34.7	87.7 / 63.2	93.9 / 49.5	98.0 / 10.4	96.5 / 15.8
	Maha	98.5 / 7.3	96.9 / 18.3	99.8 / 0.7	99.0 / 2.7	98.3 / 7.3
	Cosine	98.2 / 9.7	96.2 / 23.6	99.4 / 2.1	99.2 / 2.3	97.8 / 10.7
w/ $\mathcal{L}_{\text{scl}}$	$\mathcal{L}_{\text{scl}}$ + MSP	90.4 / 46.3	89.7 / 59.9	93.5 / 48.6	90.2 / 36.4	88.1 / 39.2
	$\mathcal{L}_{\text{scl}}$ + Energy	90.5 / 43.5	88.5 / 64.7	92.8 / 50.4	90.3 / 32.2	90.2 / 26.8
	$\mathcal{L}_{\text{scl}}$ + Maha	98.3 / 10.5	96.4 / 26.6	99.6 / 2.0	99.2 / 1.9	97.9 / 11.6
	$\mathcal{L}_{\text{scl}}$ + Cosine	97.7 / 13.0	95.9 / 28.2	99.2 / 4.2	99.0 / 2.4	96.8 / 17.0
w/ $\mathcal{L}_{\text{margin}}$	$\mathcal{L}_{\text{margin}}$ + MSP	93.0 / 33.7	89.7 / 49.2	93.9 / 46.3	97.6 / 6.5	90.9 / 32.6
	$\mathcal{L}_{\text{margin}}$ + Energy	93.9 / 31.0	89.6 / 48.8	93.4 / 52.1	98.4 / 4.6	94.1 / 18.6
	$\mathcal{L}_{\text{margin}}$ + Maha	99.5 / 1.7	99.9 / 0.6	100 / 0	99.3 / 0.4	98.9 / 6.0
	$\mathcal{L}_{\text{margin}}$ + Cosine	99.0 / 3.8	99.6 / 1.7	99.9 / 0.2	99.0 / 1.5	97.4 / 11.8

L_{margin} + Maha achieves nearly perfect OOD detection performance.

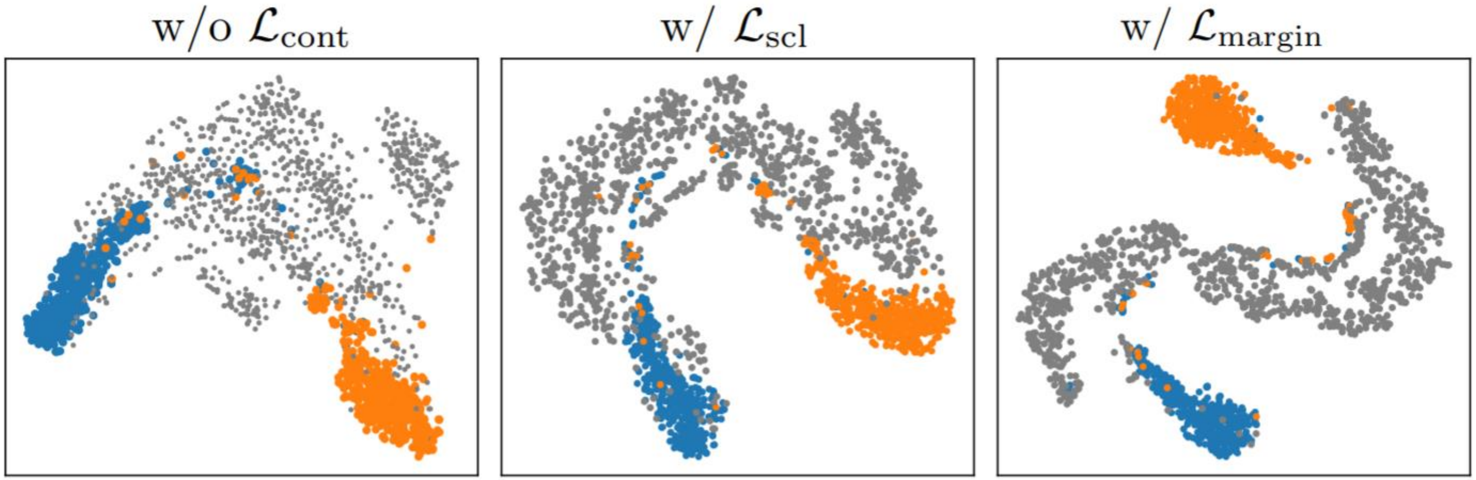
Experiments (Novel class detection)

Given a multi-class dataset, randomly reserve one class as OOD and treat others as in-distribution.

AUROC ↑ / FAR95 ↓	TREC-10	20NG
MSP	73.7 / 56.5	76.4 / 80.7
Maha	75.5 / **56.1**	77.2 / 74.1
$\mathcal{L}_{\text{margin}}$ + MSP	64.1 / 66.4	74.6 / 82.0
$\mathcal{L}_{\text{margin}}$ + Maha	**76.6** / 61.3	**78.5 / 72.7**

L_{margin} + Maha generally achieves better performance, but the gain is smaller.

Visualization



Orange: positive, blue: negative, grey: OOD

L_{margin} produces more compact representations.

Conclusion

1. We propose a margin-based contrastive objective for learning compact representations, which, in combination with the Mahalanobis distance, achieves near-perfect OOD detection on various tasks and datasets.
2. We propose novel class detection as the future challenge for OOD detection.
3. Future work includes extending our framework to more complex problems such as QA and IE.