A new approach for fine-tuning sentence transformers for intent classification and out-of-scope detection tasks

Northeastern University Khoury College of Computer Sciences

University of Victoria

cerence

Tianyi Zhang[1,2,3], Atta Norouzian[3], Aanchan Mohan[1], Frederick Ducatelle[3]

1. Khoury College of Computer Science, Northeastern University, Vancouver, BC, Canada

2. Engineering and Computer Science, University of Victoria, BC, Canada 3. Cerence Inc, Burlington, MA, USA

Problem Context

What is Out-Of-Scope(OOS) rejection and why is difficult?

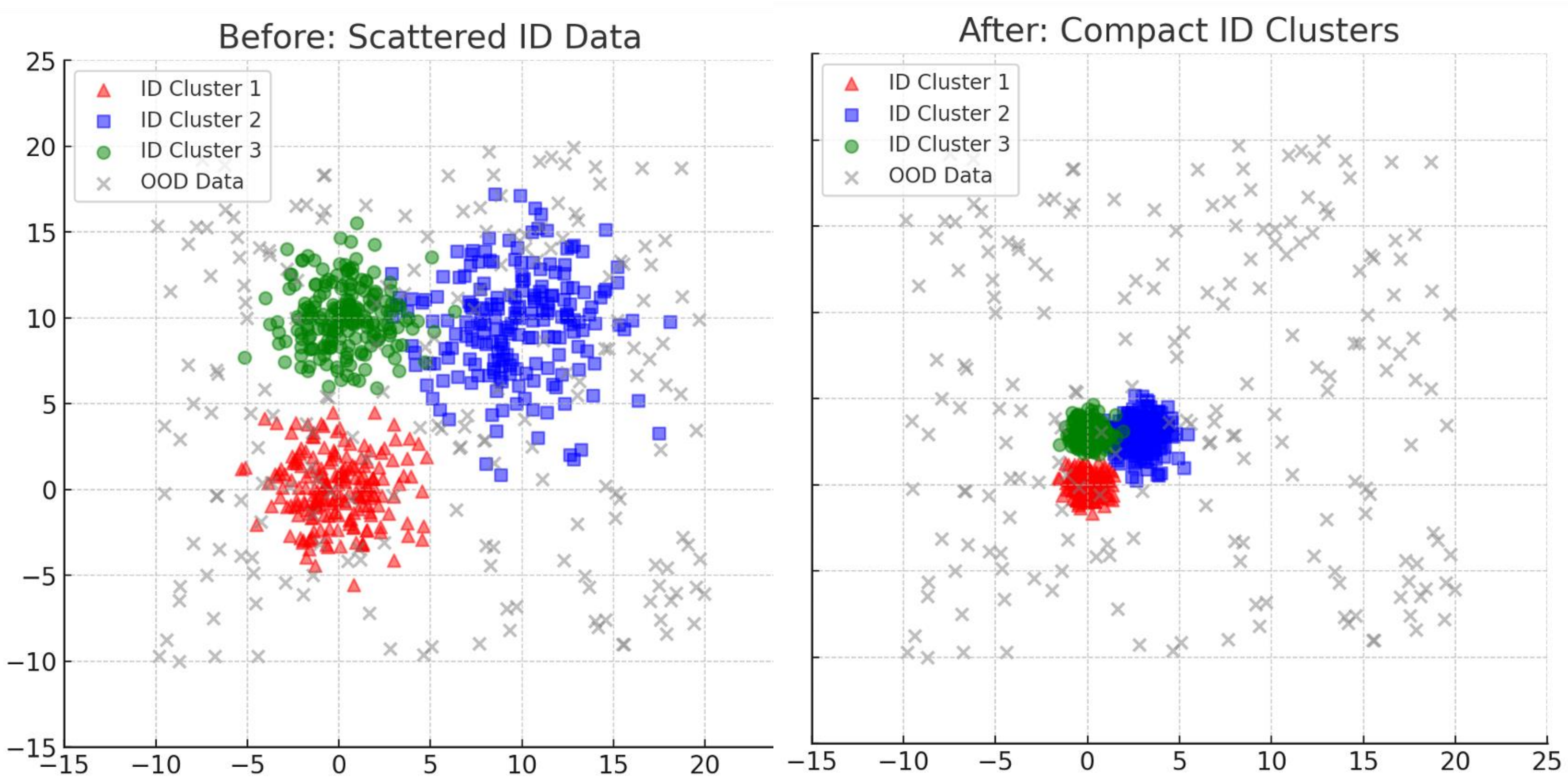
- Task of rejecting input samples outside of a set of limited domains
- Important in virtual assistant systems to handle unsupported queries or commands
- OOS has a very wide scope and cannot be efficiently represented in the training

Related Work

- Generate sentence embeddings using pretrained transformers
- Classify sample IS/OOS using parametric/non-parametric methods
- Fine-tuning encoder based on cross-entropy loss provides more suitable embeddings to distinguish intent classes
- However, fine-tuning without regularization makes forget some of the task-agnostic knowledge, leads to worse OOS detection
- Zhou et al adds a secondary loss function based on contrastive loss to increase intent class distance

Proposed Approach

Create embeddings that lead to same intent classification accuracy while projecting all in-domain samples to a small neighborhood:



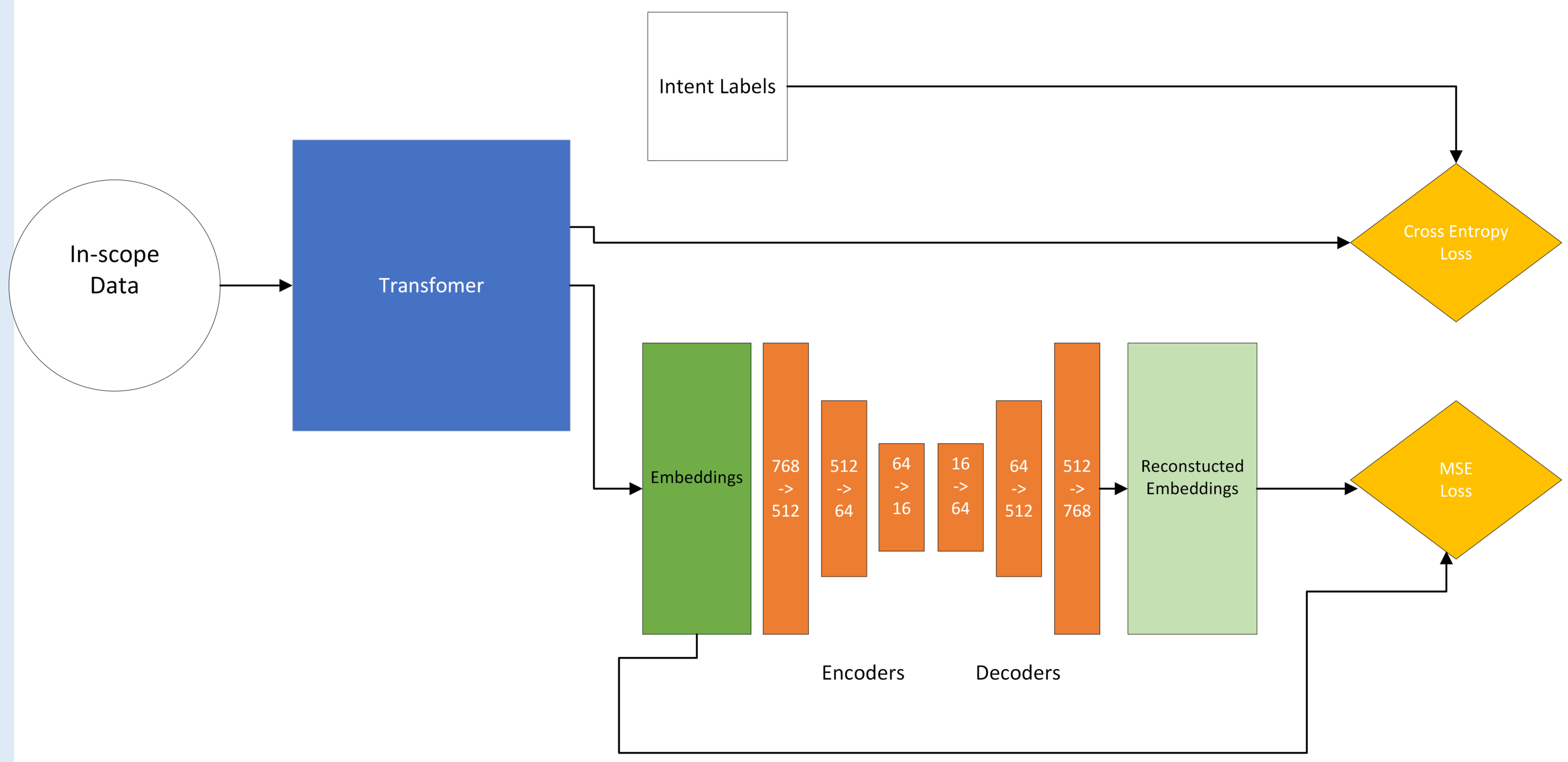
Our approach reduces the dispersion of the in-scope intent classes by regularizing fine-tuning with reconstruction loss obtained using an autoencoder.

- Start with pretrained transformer *bert-base-uncased*
- Add softmax layer with max pooling for classification
- Add secondary head with autoencoder layers
- Fine-tune model on in-scope data using joint loss:

$$\alpha * \text{CE loss} + (1-\alpha) * \text{MSE Loss}$$

- Both heads as well as softmox layer removed after fine-tuning

Model Architecture



- Embedding s^q is used for both intent classification and OOS detection
- Each in-scope intent Gaussian is represented by a centroid μ_j.
- For each query embedding s^q, we calculate Mahalanobis distance:

$$d_j(s^q) = \sqrt{(s^q - \mu_j)^T \Sigma^{-1} (s^q - \mu_j)}$$

- Pick minimum distance over all candidate centroids and assign intent
- If distance above certain global threshold τ, reject as OOS

Datasets and Results

- The trace of global covariance matrix from training embeddings was calculated to measure dispersion
- Area Under the Precision-Recall curve (AUPR) used as primary metric
- Suitable for imbalanced queries (high proportion in scope samples, positive class)

Dataset	CE	CE+AE
CLINC150	17.767	17.762
StackOverflow	16.854	16.026
MTOP	17.269	16.744

Dataset	#Train	#Test(is/oos)	Fine-tuning	AUPRoos	AUROC	Intent Classification Accuracy (%)
CLINC150	15,000	4,500 / 1,000	CE	0.916 ± 0.007	0.977 ± 0.001	95.8
			CE+AE	0.918 ± 0.004	0.978 ± 0.004	95.8
StackOverflow	79,048	16,940 / 14,617	CE	0.822 ± 0.053	0.881 ± 0.028	91.2
			CE+AE	**0.849** ± **0.050**	**0.893** ± **0.030**	90.9
MTOP	14,465	4,134 / 997	CE	0.869 ± 0.018	0.974 ± 0.004	97.0
			CE+AE	**0.899** ± **0.039**	**0.979** ± **0.009**	97.0
Car Assistant	600k	150k / 200k	CE	0.954 ± 0.005	0.959 ± 0.002	96.5
			CE+AE	**0.965** ± **0.004**	**0.966** ± **0.003**	96.6

Conclusion

- Reduced dispersion of in scope embeddings
- Similar classification accuracy compared to cross-entropy baseline
- Improved OOS detection across datasets