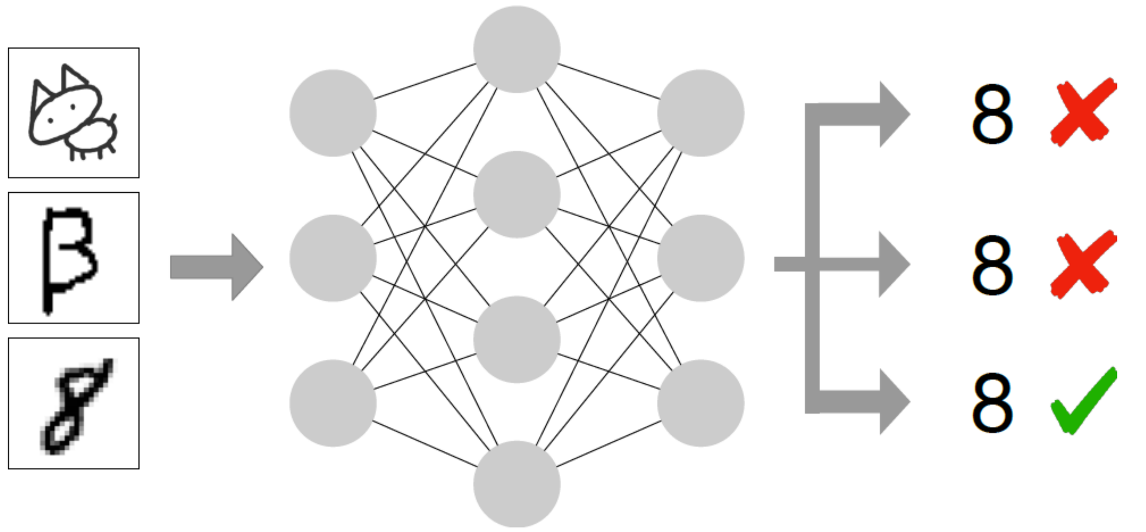
Skip The Question You Don't Know: An Embedding Space Approach

Kaiyuan Chen and Jinghao Zhao
Department of Computer Science
University of California, Los Angeles

P337

Background

- We consider the following scenario: when we meet a multiple-choice question that has low expectation to answer it correctly, then we just don't answer it to avoid extra costs
- Machine learning models cannot answer "I don't know"



MNIST Perceptron predicts Class "8" in any case

- Detecting such anomalous examples is called "Out-of-Distribution"(OOD) Detection
- Current Related Works involves evaluating based on softmax activation temperature and autoencoder
- We try to answer a question: *could we build an end-to-end model that jointly performs out-of-distribution detection and classification?*

Problem Formulation

Consider a dataset distribution X^1 with n_1 examples and their associated labels $\{(x^1_1, y^1_1)...(x^1_n, y^1_{n_1})\}$ and an anomaly distribution X^2 with $n_2 << n_1$ examples $\{(x^2_1, *)...(x^2_{n_2}, *)\}$. We mark labels of anomaly distribution as $*$ since we don't want our model to assign any label to these data. Then for all $\epsilon, \delta \in [0, 1]$, a successful OOD algorithm A with its classifier C trained with $X^1 \cup X^2$ should have at least $1 - \delta$ probability to identify x such that

$$\mathbb{E}(\ell(C(x), y)) > \epsilon$$

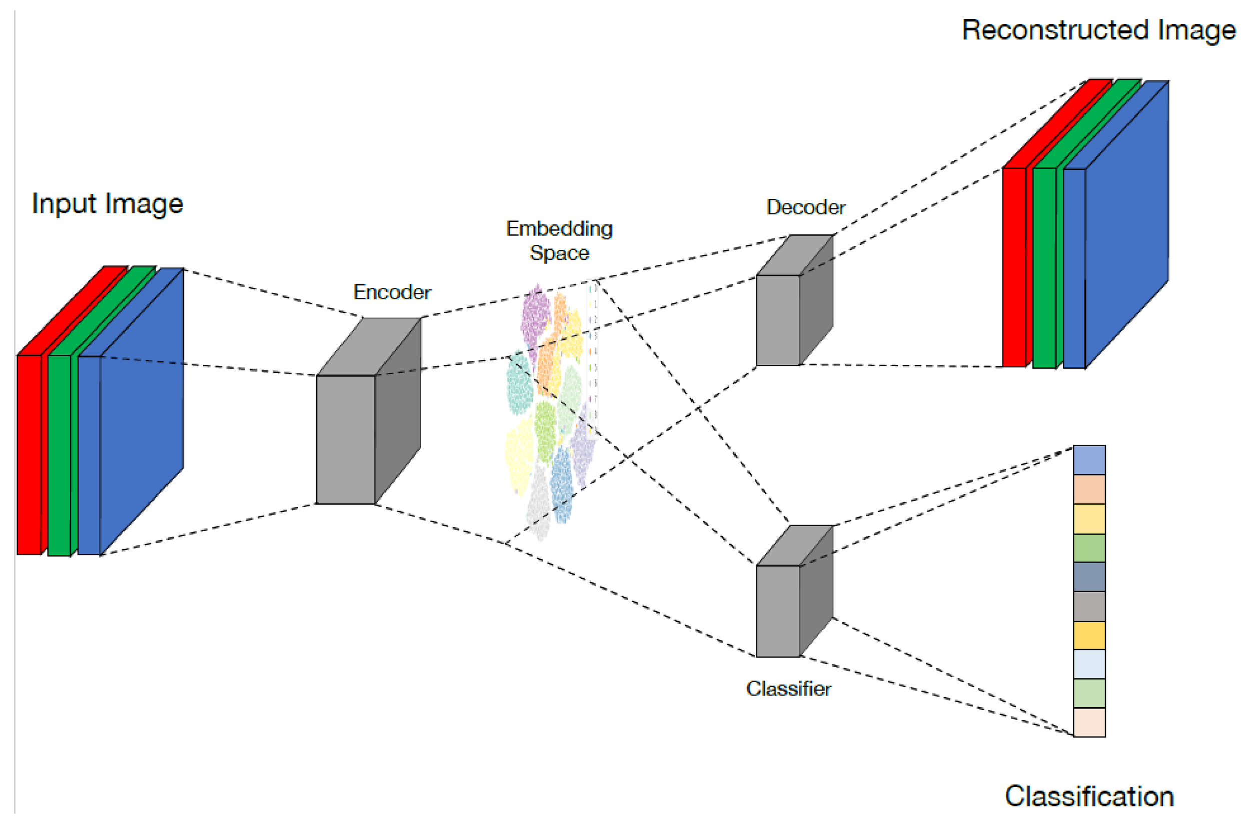
where $\mathbb{E}(\cdot)$ is the expectation function and ℓ is the loss function. We define $\ell = 1$ if $y = *$.

Contribution

- We propose an end-to-end architecture with associated loss function that jointly optimizes out-of-distribution detection and classification by learning a label-clustered embedding
- We understand outliers by backtracking them visually in our embedding space and devise a training process that dynamically removes outliers in training set
- We empirically compare our model with other OOD detection algorithms in various datasets that are mixed with OOD examples

Our Approach

- In order to perform classification and OOD at the same time, we adopt a commonly used architecture as following:



- Suggested by Zhang et al., unsupervised loss can improve the accuracy of classifier. However, directly using this structure cannot sufficiently solve our problem.
- We integrate this architecture with Deep Embedding Clustering, an embedding space regularizer that can cluster the similar points and push away dissimilar ones
- We calculate current embedding metric by

$$q_{ij} = \frac{(1 + ||z_i - (1-\alpha)\mu_j - \alpha\mu'_j||^2)^{-1}}{\sum_j (1 + ||z_i - (1-\alpha)\mu_j - \alpha\mu'_j||^2)^{-1}}$$

which use α as interpolation between current cluster center and label-based average, which is defined as

$$\mu'_j = \frac{1}{|y_i = j|} \sum_k \mathbb{1}(y_i = j) z_k$$

- To achieve our goal of clustering points, we calculate a target distribution to can make the similar points and dissimilar ones var away

$$p_{ij} = \frac{q^2_{ij} / \sum_i q_{ij}}{\sum_{j'} q^2_{ij'} / \sum_i q_{ij}}$$

- Then we construct our embedding loss as

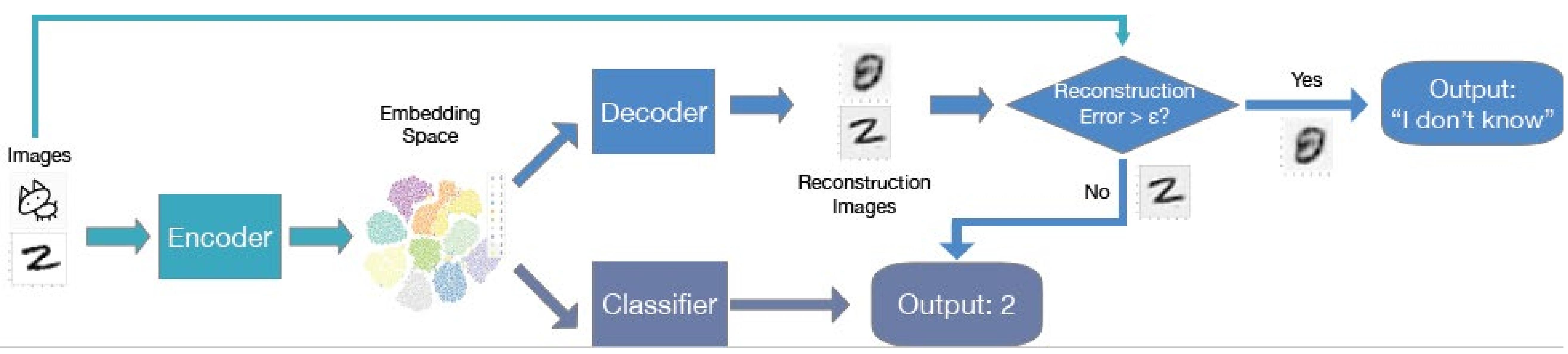
$$J_e = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

- With the reconstruction error for autoencoder, which is defined as

$$J_r = ||x - f_{DE}(f_{EN}(x))||^2_2$$

We devise our new loss function

$$J = J_r + \lambda_1 J_e + \lambda_2 J_c$$



A Flow Chart of Our Proposed Approach

Evaluation

- We compare our model with other models on different types of datasets
- We have OOD datasets different from MNIST, such as iSUN, Omniglot, notMNIST, CIFAR-bw, random noise

In-Distribution/ Out-of-Distribution	AUROC	AUPR-In	AUPR-Out
	Hendrycks et al.[6]/Our Method		
CIFAR-10/SUN	99.99/100	99.95/100	99.04/100
CIFAR-10/Gaussian	100/100	100/100	99.24/100
MNIST/Omniglot	99.45/99.50	99.49/99.45	99.40/99.38
MNIST/notMNIST	100/100	100/100	99.97/100
MNIST/CIFAR-10bw	99.97/100	99.97/100	99.97/100
MNIST/Gaussian	100/100	100/100	100/100
MNIST/Uniform	100/100	100/100	100/100

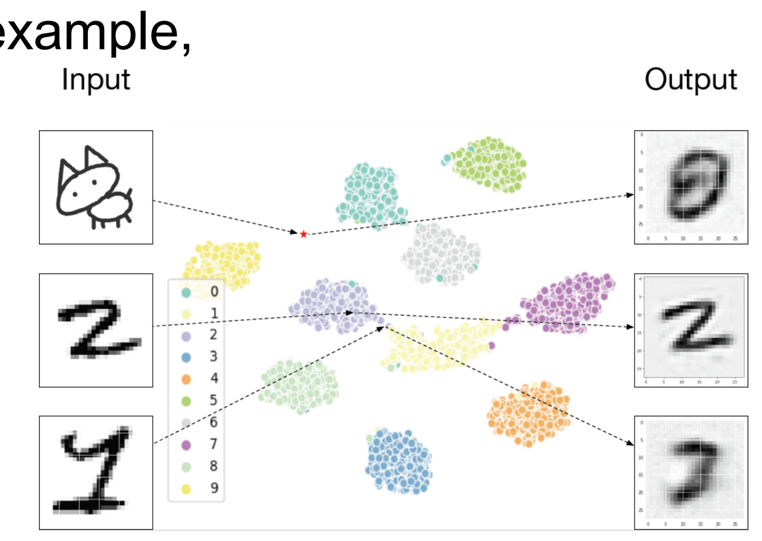
TABLE I
THE EVALUATION OF IN- AND OUT-OF-DISTRIBUTION DETECTION FOR THE DATASETS IN BASELINE. THE BOLD TEXT INDICATES BETTER NOVELTY DETECTION PERFORMANCE. EACH VALUE CELL IS IN "BASELINE/OUR METHOD" FORMAT.

- We claim the original benchmark proposed by Hendrycks et al. is not challenging enough. Thus, we propose MNIST-{x}:
- This dataset is generated from original MNIST dataset by removing points with label x. For example, MNIST-{0,1} means we use data points with label 2,3,4,5,6,7,8,9 as in-samples and points with label 0 and 1 as outliers

In-Distribution/ Out-of-Distribution	AUROC	AUPR-In	AUPR-Out
	Hendrycks et al.[6]/Our Method		
MNIST'/MNIST-{0,1}	93.57/94.65	98.10/98.16	75.98/89.10
MNIST'/MNIST-{2,3}	90.46/98.85	96.44/99.72	75.48/95.18
MNIST'/MNIST-{4,5}	92.82/96.78	98.25/99.24	72.98/86.88
MNIST'/MNIST-{6,7}	95.37/95.49	98.76/98.81	82.90/84.13
MNIST'/MNIST-{8,9}	94.63/95.52	98.69/98.86	75.19/82.68

TABLE II

- In order to see why our model works, we go back to our previous motivating example,



- In terms of functionality, we compare with other state of art models in OOD detection,

	OOD Detection	Joint Optimization for Classification	Clustering	Interpretable Embedding space
ODIN [10]	✓			
DEC [8]			✓	✓
Shaol et al.[21]		✓	✓	O
Devries et al.[11]	✓	✓		
Hendrycks et al.[6]	✓	✓		
Our model	✓	✓	✓	✓

References

D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," arXiv preprint arXiv:1610.02136, 2016
J. Xie, R. Girshick, and A. Farhadi, Unsupervised deep embedding for clustering analysis," in International conference on machine learning, 2016, pp. 478–487.