ContrastNet: A Contrastive Learning Framework for Few-shot Text Classification

Junfan Chen[1,2], Richong Zhang[1,2]*, Yongyi Mao[3], Jie Xue[4]

[1]Beijing Advanced Institution for Big Data and Brain Computing, Beihang University, Beijing, China
[2]SKLSDE, School of Computer Science and Engineering, Beihang University, Beijing, China
[3]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada
[4]Department of Computer Science, University of Leeds, UK
chenjf@act.buaa.edu.cn, zhangrc@act.buaa.edu.cn, ymao@uottawa.ca, j.xu@leeds.ac.uk

(AAAI-2022)

Reported by Zhaoze Gao

1. Introduction

2. Approach

3. Experiments

什么是Few-shot（小样本）？

在图像分类的问题下，正确率可以很轻松的达到94%之上。然而，deep learning是一种data hungry的技术，需要大量的标注样本才能发挥作用。但现实世界中，有很多问题是没有那么多的标注数据的，获取标注数据的成本也非常大。因此，我们讨论的是这样一个问题的场景，也就是小样本问题。它面临的问题是：

- 训练过程中有从未见过的新类，只能借助每类少数几个标注样本；
- 不改变已经训练好的模型；

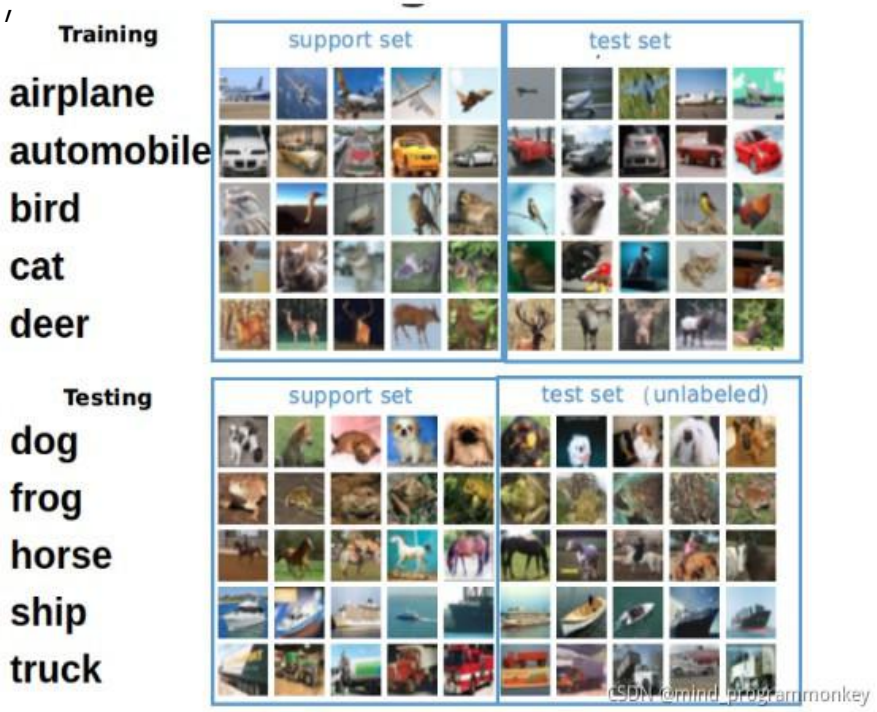


图 1

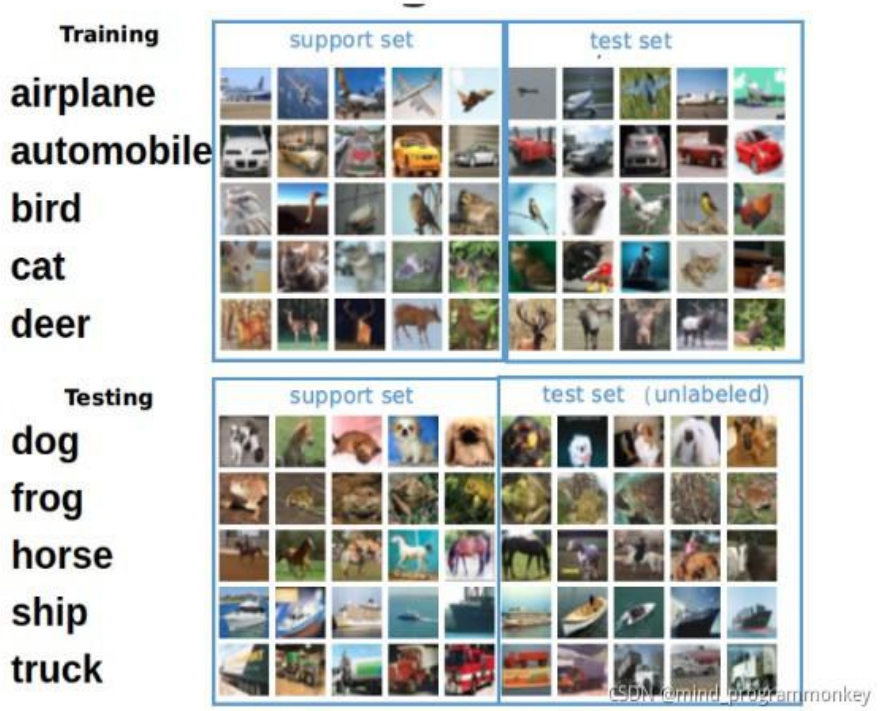
传统的方法是基于左边这些训练集，获得模型，然后对右边测试集进行自动标注。

而小样本问题如图 1 所示，我们大量拥有的是上方这5类的数据，而新问题（下方这5类）是只有很少的标注数据。

模型在学习了一定类别的大量数据后，对于新的类别，只需要少量的样本就能快速学习，即Few-shot learning要解决的问题。

Few-shot的测试阶段

test stage：在测试阶段，我们当前拿到的测试集的label当然是已知的。此时也进行类似episode training的机制，每一个episode从测试集中随机采样N个类别，每个类别采样K个标注样本构成支持集，此时支持集的label是已知的也是可用的，之后每个类别的剩余一小部分样本构成查询集，这些样本是需要模型分类的，label是未知的，通过模型计算得到查询集样本的accuracy，此时就模拟了我们在每个类别小样本label已知的情况下，去预测其他相同label空间样本。



Few-shot的基本概念

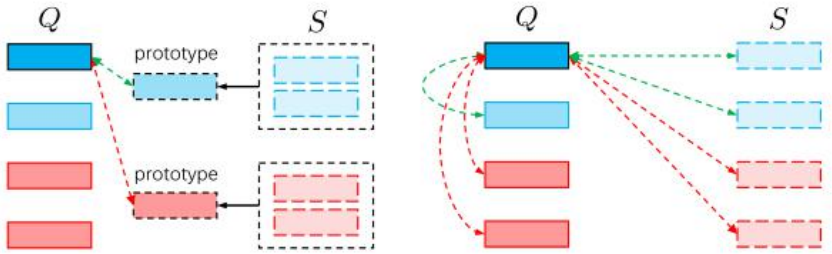
- episode：每一个episode便是训练一次，类似于传统训练方法里的一个batch，然后这个episode内部由两部分构成：support set和query set。

- support set：在构建这个episode时候，会从全量数据中每个episode都随机选择一些类别，比如C个类别，然后从数据集中同样随机从这选定的C个类别中选取同样数量的K个样本，这便构成了support set，总共包含C * K个样本。一般而言模型会在这个上面进行一次训练。这样构造出来的任务便是C-way K-shot。

- query set: 和support set类似，会在剩下的数据集样本中的C个类（注意这里的类别是从对应的support set类别选择）分别采样m个样本作为query set。

- task：一个episode包含一个support set和一个query set，一个episode对应的便是让模型在support set上进行学习后能够在query set上有比较好的预测表现。因此一个episode便对应为一个task。

Introduction



(a) Prototypical Networks (b) ContrastNet

Figure 1: The learning strategies of Prototypical Network and proposed ContrastNet. Q and S respectively denote the query set and support set. The rectangles with different colors denote text representations from different classes. The green and red dashed arrow lines respectively indicate pulling closer and pushing away the representations. Picture (a) shows that Prototypical Networks learn to align a given query-text representations to prototypes computed by support-text representations. Picture (b) shows that ContrastNet learns to pull closer the given query-text representation with text representations belonging to the same class and push away text representations with different classes.

- When these two sentences are sampled in the same query set, they are hard to distinguish from each other and bring about contradiction in prediction because they will obtain similar measurements aligning to each prototype,

"who covered the song one more cup of coffee" with intent music-query

"play the song one more cup of coffee" with intent music-play

- Another challenge in few-shot text classification is that the models are prone to overfit the source classes based on the biased distribution formed by a few training examples.

Approach

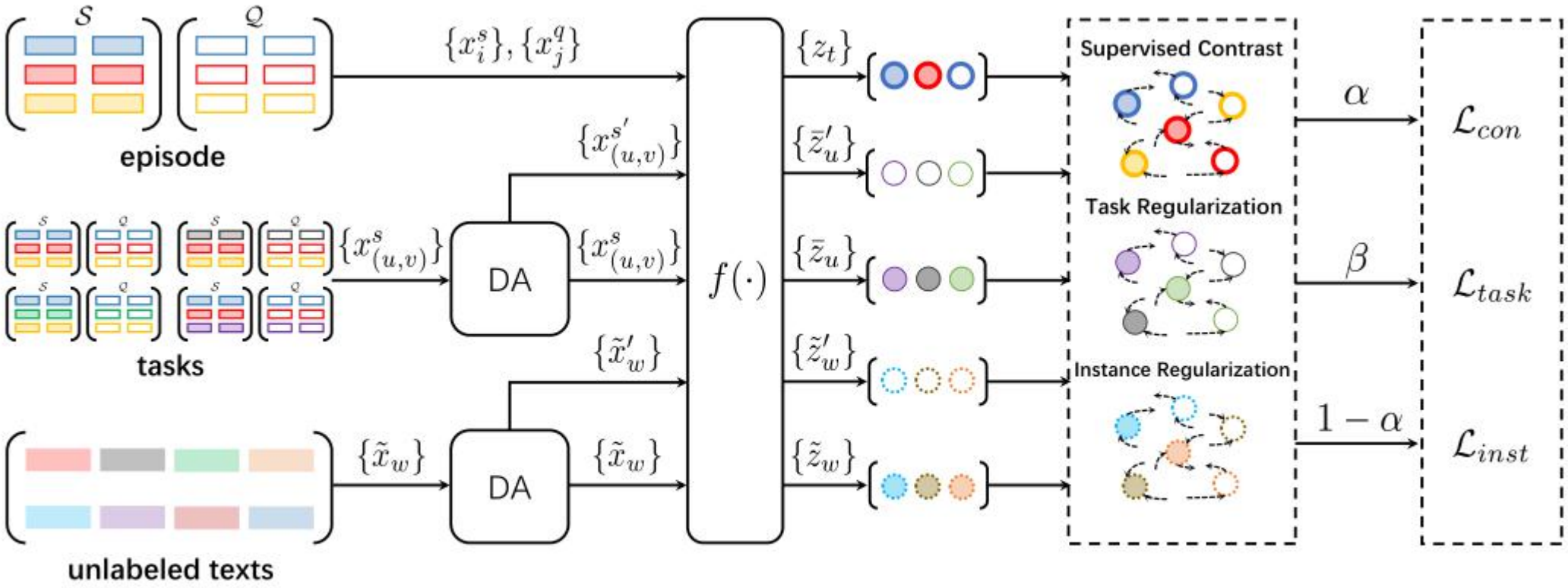


Figure 2: The overall model structure of ContrastNet. The DA blocks represent data augmentation.

Approach

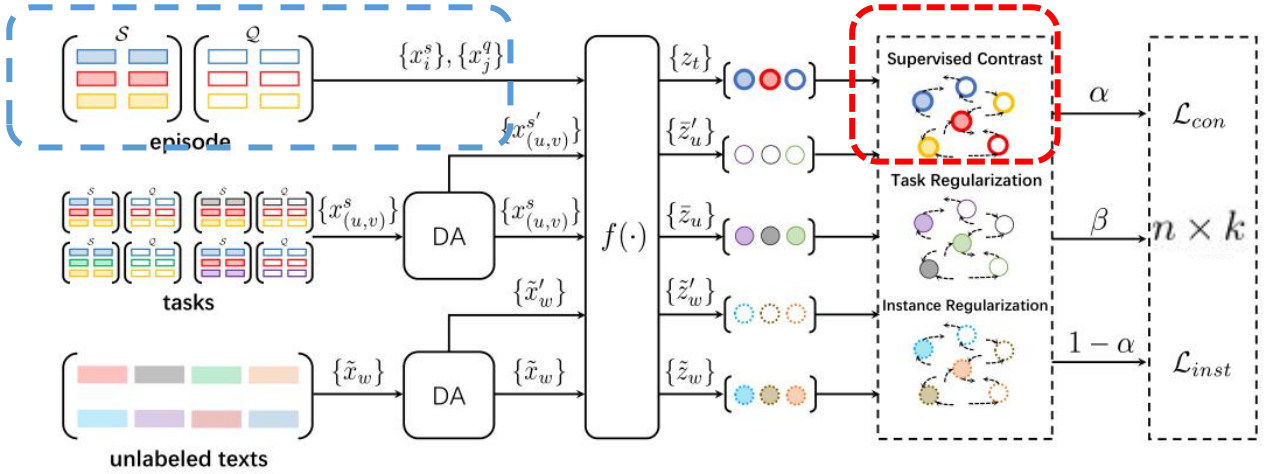


Figure 2: The overall model structure of ContrastNet. The DA blocks represent data augmentation.

$$(x_i^s, y_i^s) \quad n \times k$$
$$x_j^q \quad n \times m$$

$$\mathcal{B} = \{x_1, x_2, \cdots, x_{n(k+m)}\},$$

$$x_t = \begin{cases} x_t^s, & t \leqslant nk \\ x_{t-nk}^q, & t > nk \end{cases} \tag{1}$$

$$\mathcal{L}_{con} = -\sum_{\mathbf{x}_t \in \mathcal{B}} \frac{1}{c} \log \frac{\sum_{y_r = y_t} \exp(z_t \cdot z_r / \tau)}{\sum_{y_r = y_t} \exp(z_t \cdot z_r / \tau) + \sum_{y_{r'} \neq y_t} \exp(z_t \cdot z_{r'} / \tau)} \tag{2}$$

$$c = k + m - 1$$

Approach

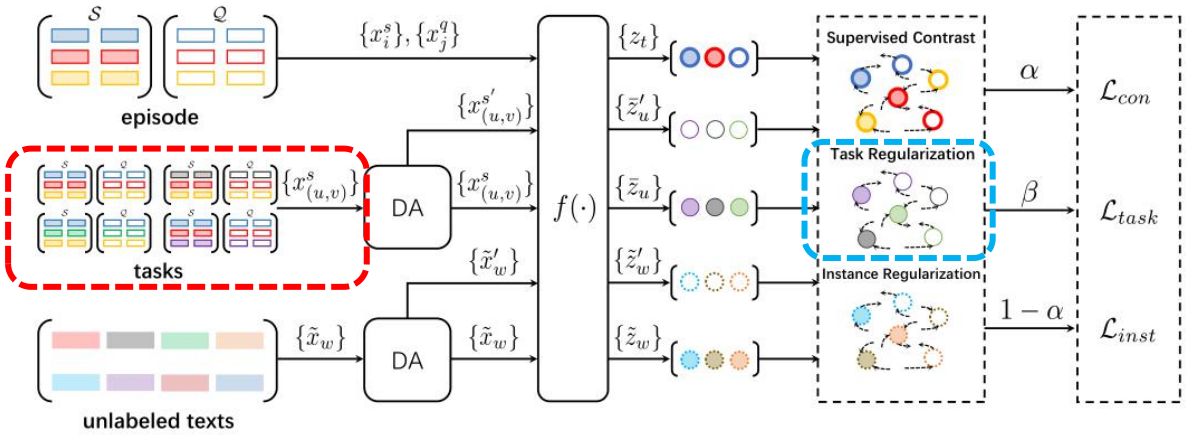


Figure 2: The overall model structure of ContrastNet. The DA blocks represent data augmentation.

$$\{(\mathcal{Q}_1, \mathcal{S}_1), (\mathcal{Q}_2, \mathcal{S}_2), \cdots, (\mathcal{Q}_{N_{task}}, \mathcal{S}_{N_{task}})\}$$

$$\mathcal{L}_{task} = -\sum_{u=1}^{2N_{task}} \log \frac{\exp(\bar{z}_u \cdot \bar{z}'_u/\tau)}{\exp(\bar{z}_u \cdot \bar{z}'_u/\tau) + \sum_{\bar{z}_{u'} \neq \bar{z}'_u} \exp(\bar{z}_u \cdot \bar{z}_{u'}/\tau)} \quad (3)$$

Approach

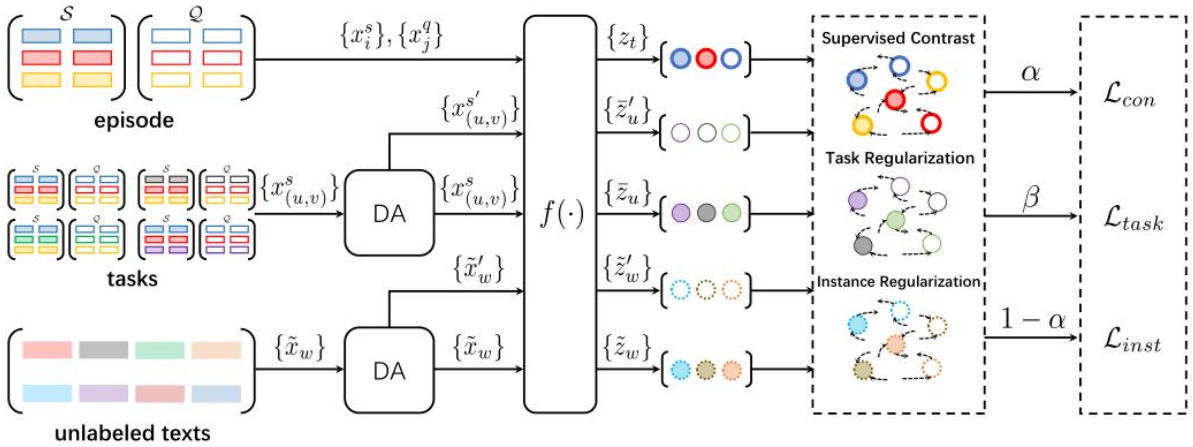


Figure 2: The overall model structure of ContrastNet. The DA blocks represent data augmentation.

$$\{\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_{N_{inst}}\}$$

$$\mathcal{L}_{inst} = -\sum_{w=1}^{2N_{inst}} \log \frac{\exp(\tilde{z}_w \cdot \tilde{z}'_w/\tau)}{\exp(\tilde{z}_w \cdot \tilde{z}'_w/\tau) + \sum_{\tilde{z}_{w'} \neq \tilde{z}'_w} \exp(\tilde{z}_w \cdot \tilde{z}_{w'}/\tau)} \quad (4)$$

$$\mathcal{L} = \alpha\mathcal{L}_{con} + (1-\alpha)\mathcal{L}_{inst} + \beta\mathcal{L}_{task} \quad (5)$$

$$i^* = \arg\max_i f(x^q) \cdot f(x^s_i) \quad (6)$$

Experiments

dataset	train/valid/test classes	sentences	avg_sent_class	avg_tok_sent
Banking77	25/25/27	13,083	170	12
HWU64	23/16/25	11,036	172	7
Clinic150	50/50/50	22,500	150	9
Liu	18/18/18	25,478	472	8
HuffPost	20/5/16	36,900	900	11
Amazon	10/5/9	24,000	1000	140
Reuters	15/5/11	620	20	168
20News	8/5/7	18,820	941	340

Table 1: The statistics of few-shot text classification datasets. The *avg_sent_class* denotes average sentences per class and *avg_tok_sent* denotes average tokens per sentence.

Experiments

Method	Banking77		HWU64		Liu		Clinic150		Average	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Prototypical Networks	86.28	93.94	77.09	89.02	82.76	91.37	96.05	98.61	85.55±2.20	93.24±1.22
PROTAUGMENT	86.94	94.50	82.35	91.68	84.42	92.62	94.85	98.41	87.14±1.36	94.30±0.60
PROTAUGMENT (bigram)	88.14	94.70	84.05	92.14	85.29	93.23	95.77	98.50	88.31±1.43	94.64±0.59
PROTAUGMENT (unigram)	89.56	94.71	84.34	92.55	86.11	93.70	96.49	**98.74**	89.13±1.13	94.92±0.57
ContrastNet ($\mathcal{L}_{task}$&$\mathcal{L}_{inst}$ /o)	88.53	95.22	84.62	91.93	80.53	93.47	94.29	98.09	86.99±1.57	94.68±0.74
ContrastNet ($\mathcal{L}_{inst}$ /o)	89.75	95.36	85.14	91.69	**86.79**	93.28	96.32	98.25	89.50±1.30	94.65±0.64
ContrastNet	**91.18**	**96.40**	**86.56**	**92.57**	85.89	**93.72**	**96.59**	98.46	**90.06±1.02**	**95.29±0.53**

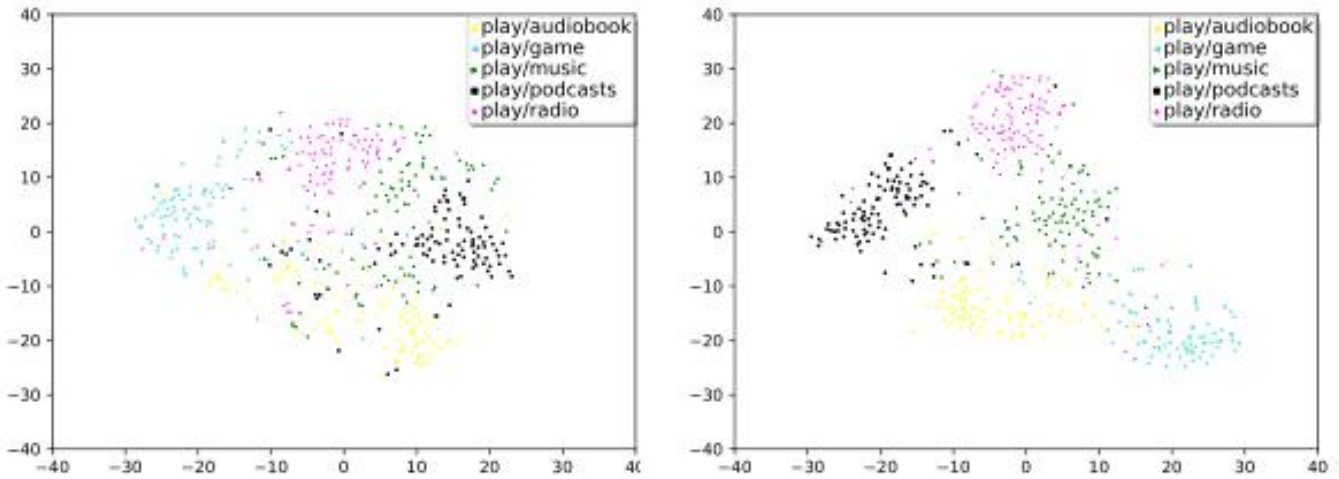
Table 2: The 5-way 1-shot and 5-way 5-shot text classification results on the Banking77, HWU64, Liu and Clinic150 intent classification datasets. The ContrastNet ($\mathcal{L}_{task}$&$\mathcal{L}_{inst}$ /o) model denote the ContrastNet only using supervised contrastive text representation without any unsupervised regularization and the ContrastNet ($\mathcal{L}_{inst}$ /o) model denotes the ContrastNet with only task-level unsupervised regularization. We compute the mean and the standard deviation over 5 runs with different class splitting. The **Average** denotes the averaged mean and standard deviation over all datasets for each setting of each model.

Experiments

Method	HuffPost		Amazon		Reuters		20News		Average	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MAML	35.9	49.3	39.6	47.1	54.6	62.9	33.8	43.7	40.9	50.8
Prototypical Networks	35.7	41.3	37.6	52.1	59.6	66.9	37.8	45.3	42.7	51.4
Induction Networks	38.7	49.1	34.9	41.3	59.4	67.9	28.7	33.3	40.4	47.9
HATT	41.1	56.3	49.1	66.0	43.2	56.2	44.2	55.0	44.4	58.4
DS-FSL	43.0	63.5	62.6	81.1	81.8	96.0	52.1	68.3	59.9	77.2
MLADA	45.0	64.9	68.4	**86.0**	82.3	**96.7**	59.6	77.8	63.9	81.4
ContrastNet ($\mathcal{L}_{task}$&$\mathcal{L}_{inst}$ /o)	52.74	63.59	74.70	84.47	83.74	93.28	70.61	80.04	70.45±3.28	80.35±3.32
ContrastNet ($\mathcal{L}_{inst}$ /o)	52.85	64.88	75.33	84.21	85.10	93.65	70.35	80.19	70.91±3.00	80.73±2.79
ContrastNet	**53.06**	**65.32**	**76.13**	85.17	**86.42**	95.33	**71.74**	**81.57**	**71.84±2.81**	**81.85±2.03**

Table 3: The 5-way 1-shot and 5-way 5-shot text classification results on the HuffPost, Amazon, Reuters and 20News datasets.

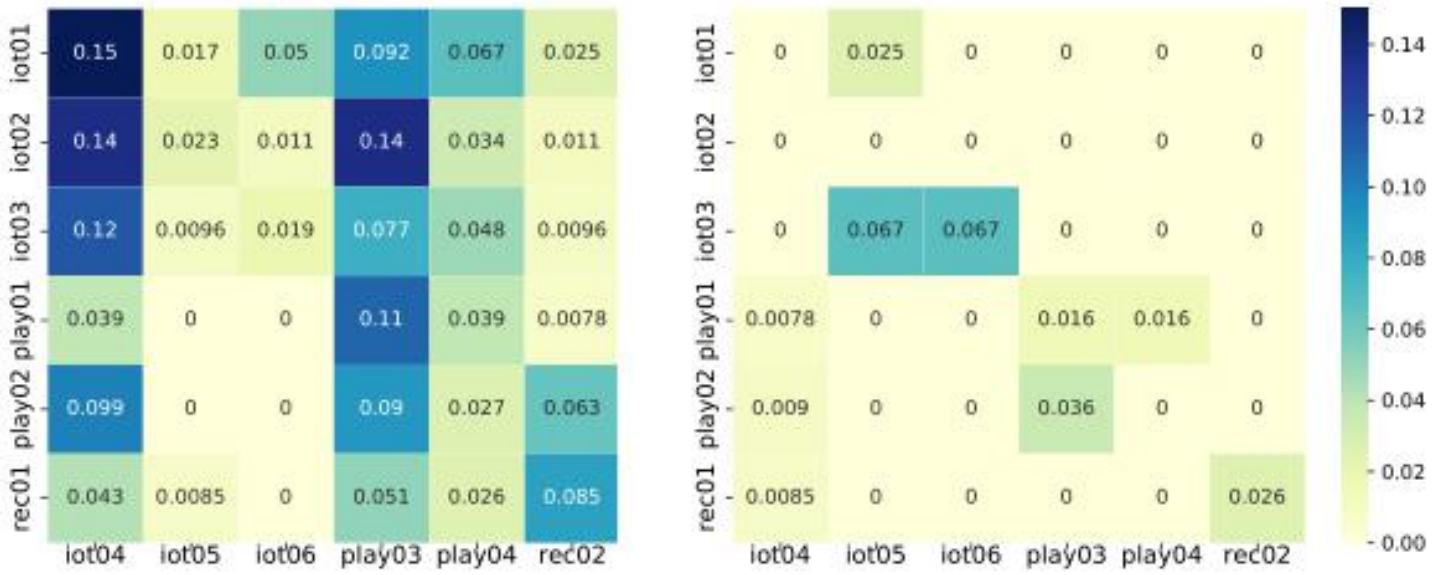
Experiments



(a) Prototypical Networks (b) ContrastNet

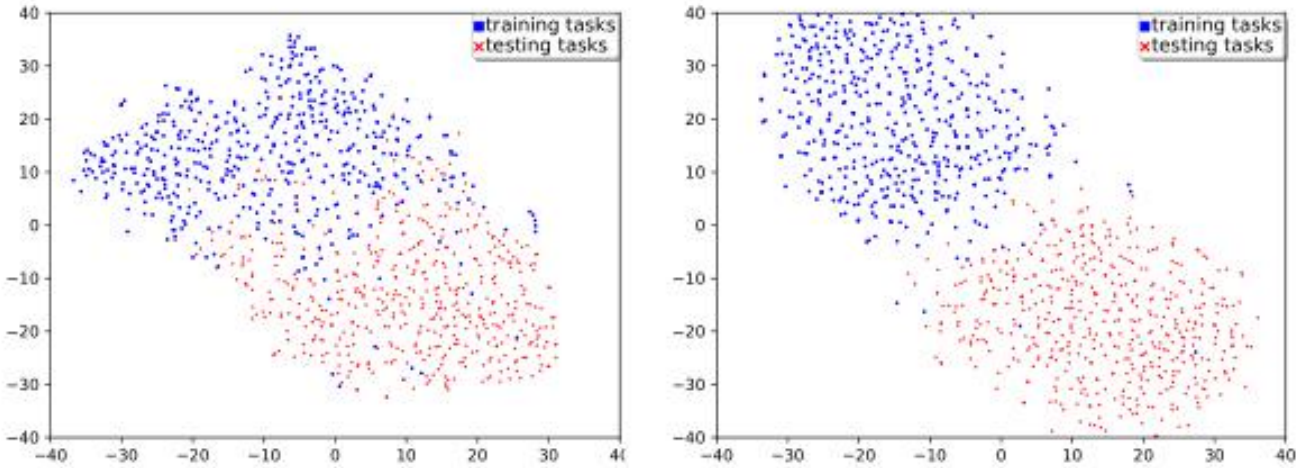
Figure 3: Visualization of query text representations sampled from similar target classes on HWU64.

Experiments



(a) Prototypical Networks (b) ContrastNet

Experiments



(a) Prototypical Networks (b) ContrastNet

Figure 5: Task-representation vasualization on Banking77.

Experiments

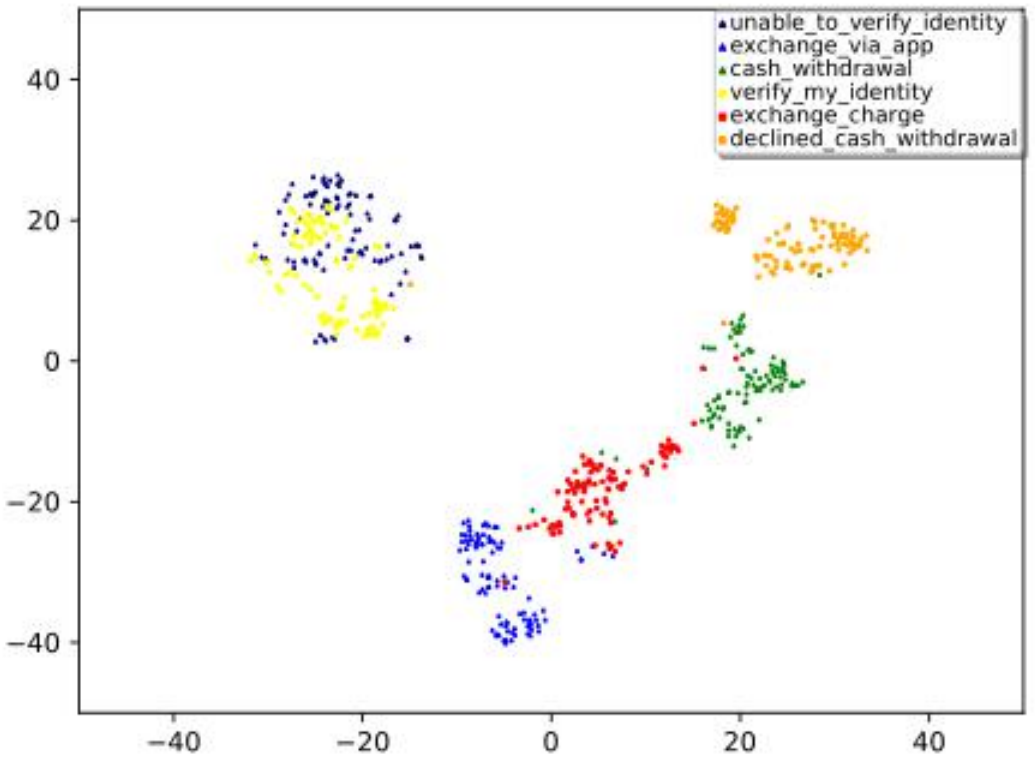


Figure 6: Text-representation of Prototypical Networks.

Experiments

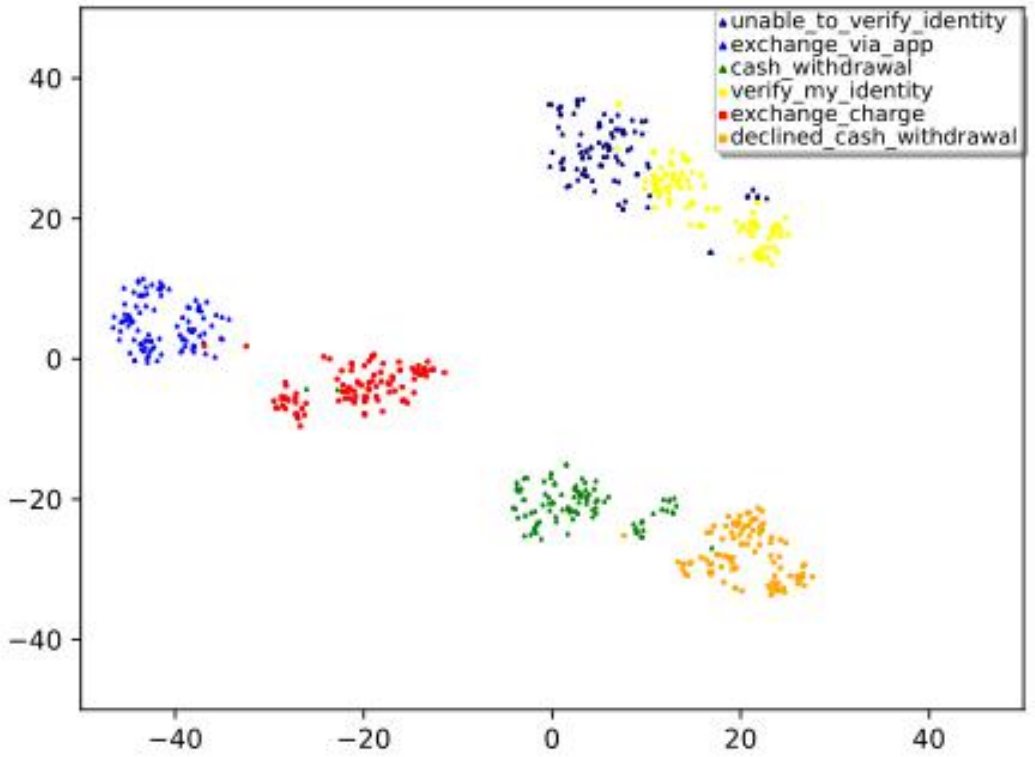


Figure 7: Text-representation of ContrastNet.

Thank you !