Meta-Learning Siamese Network for Few-Shot Text Classification

Chengcheng Han[1], Yuhe Wang[1], Yingnan Fu[1], Xiang Li[1*],

Minghui Qiu[2], Ming Gao[1], Aoying Zhou[1]

[1]School of Data Science and Engineering, East China Normal University

[2]Alibaba Group

Table of Contents

- Background
  - Problem Definition
  - Related Work
- Method
- Experiments
- Conclusion

Problem Definition

Training data $\mathcal{Y}_{train}$:
technique, travel, media, beauty, taste, style, culture

Testing data $\mathcal{Y}_{test}$:
politics, entertainment, sports

$$\mathcal{Y}_{train} \cap \mathcal{Y}_{test} = \emptyset$$

For each of these new classes, we only have a few labeled examples.

Related Work

Transfer Learning

Pretrained Language Model

- Bert, XLNet

Prompt-based Methods

- GPT3, ParaBART

Meta Learning

Metric-based

- Matching Network, PROTO, Relation network ...

Model-based

- MANNs, Meta networks, SNAIL ...

Optimization-based

- MAML, Reptile, LSTM optimizer ...

Method

Prototypical Networks (PROTO)

Algorithm 1 Training episode loss computation for prototypical networks. N is the number of examples in the training set, K is the number of classes in the training set, $N_C \leq K$ is the number of classes per episode, N_S is the number of support examples per class, N_Q is the number of query examples per class. RANDOMSAMPLE(S, N) denotes a set of N elements chosen uniformly at random from set S, without replacement.

Input: Training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where each $y_i \in \{1, \ldots, K\}$. $\mathcal{D}_k$ denotes the subset of $\mathcal{D}$ containing all elements $(\mathbf{x}_i, y_i)$ such that $y_i = k$.

Output: The loss J for a randomly generated training episode.

$V \leftarrow$ RANDOMSAMPLE$(\{1, \ldots, K\}, N_C)$ ▷ Select class indices for episode

for k in $\{1, \ldots, N_C\}$ **do**

$S_k \leftarrow$ RANDOMSAMPLE$(\mathcal{D}_{V_k}, N_S)$ ▷ Select support examples

$Q_k \leftarrow$ RANDOMSAMPLE$(\mathcal{D}_{V_k} \setminus S_k, N_Q)$ ▷ Select query examples

$$\mathbf{c}_k \leftarrow \frac{1}{N_C} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_{\phi}(\mathbf{x}_i)$$ ▷ Compute prototype from support examples

end for

$J \leftarrow 0$ ▷ Initialize loss

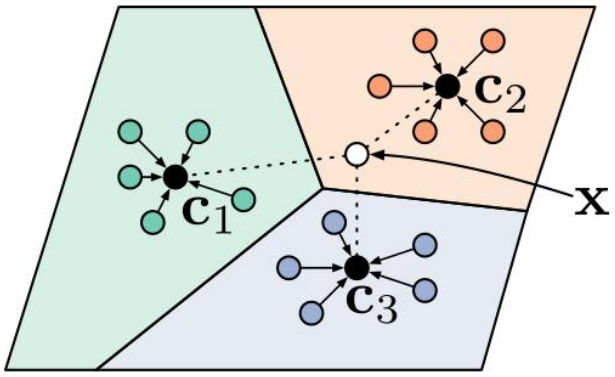
for k in $\{1, \ldots, N_C\}$ **do**

for $(\mathbf{x}, y)$ in Q_k **do**

$$J \leftarrow J + \frac{1}{N_C N_Q} \left[d(f_{\phi}(\mathbf{x}), \mathbf{c}_k)) + \log \sum_{k'} \exp(-d(f_{\phi}(\mathbf{x}), \mathbf{c}_k)) \right]$$ ▷ Update loss

end for

end for



$$p_{\phi}(y = k \,|\, \mathbf{x}) = \frac{\exp(-d(f_{\phi}(\mathbf{x}), \mathbf{c}_k))}{\sum_{k'} \exp(-d(f_{\phi}(\mathbf{x}), \mathbf{c}_{k'}))}$$

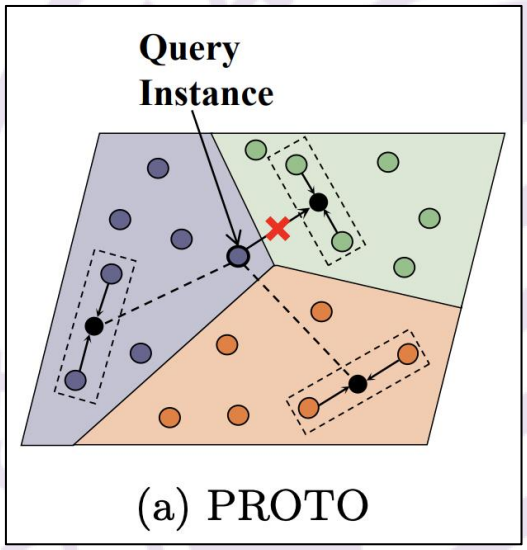
Method

There are three main problems in PROTO:

(1) ignore the randomness of the sampled support sets when computing prototype vectors;

(2) disregard the importance of labeled samples;

(3) construct meta-tasks in a purely random manner.

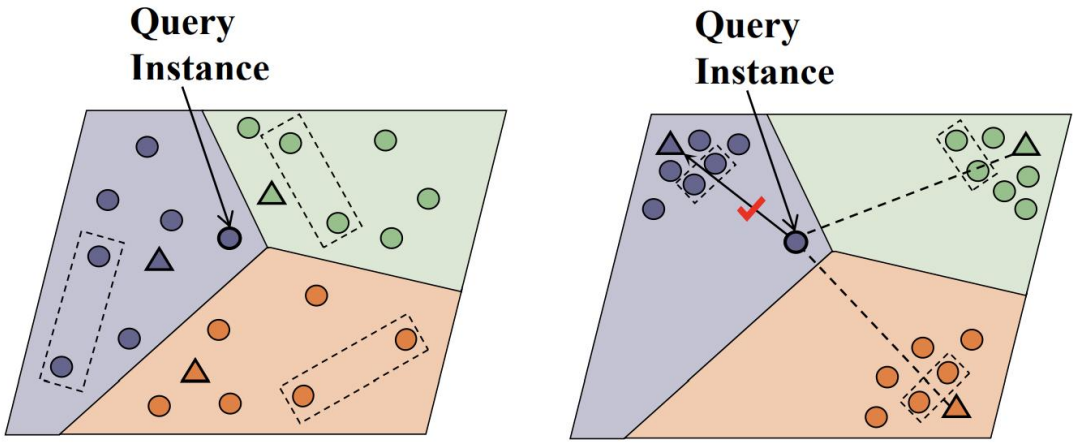


(a) PROTO

Method

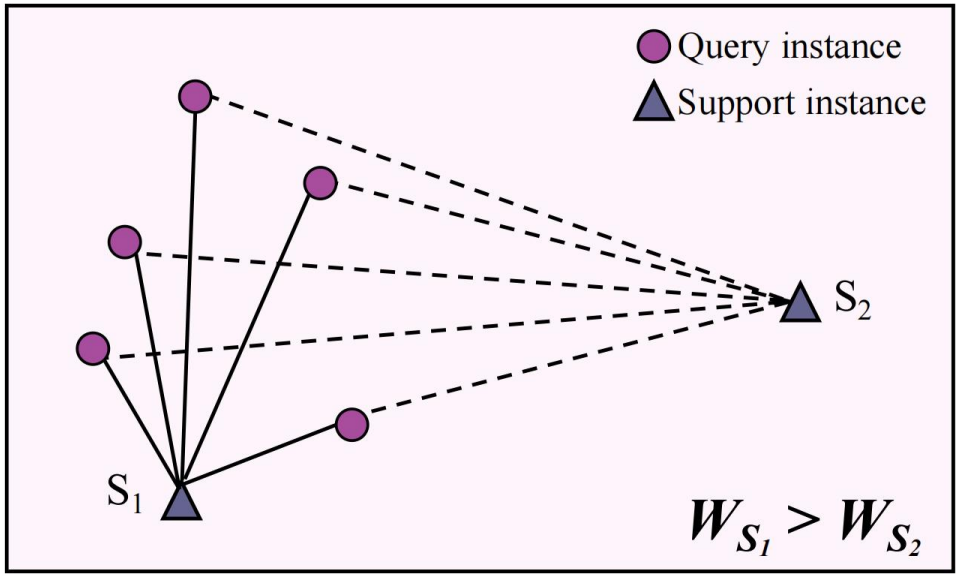
Unbiased prototype vectors

Support set	
(A) *Politics*	(1) Trump's Crackdown On Immigrant Parents Puts More Kids In An Already Strained System. (2) Ireland Votes To Repeal Abortion Amendment In Landslide Referendum.
(B) *Entertainment*	(1) Hugh Grant Marries For The First Time At Age 57. (2) Mike Myers Reveals He'd 'Like To' Do A Fourth Austin Powers Film.
(C) *Sports*	(1) U.S. Olympic Committee Ignored Sexual Abuse Complaints Against Taekwondo Stars: Lawsuit. (2) MLB Pitcher Punches Himself In Face Really Hard After Blowing Game.
Query instance	
Which class?	'Crazy Rich Asians' Trailer Is Already A Magnificent Masterpiece.
External knowledge (class name and related descriptive texts)	
(A)	**Politics** is the set of activities that are associated with making decisions in groups.
(B)	**Entertainment** is a form of activity that holds the attention and interest of an audience.
(C)	**Sports** pertain to any form of competitive physical activity or game.



(b)Initialization in Meta-SN (c) Refinement in Meta-SN

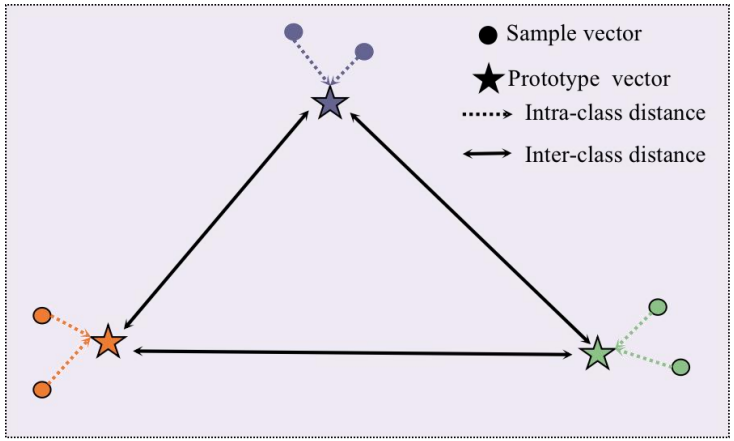
Weights of Samples in the Support Set



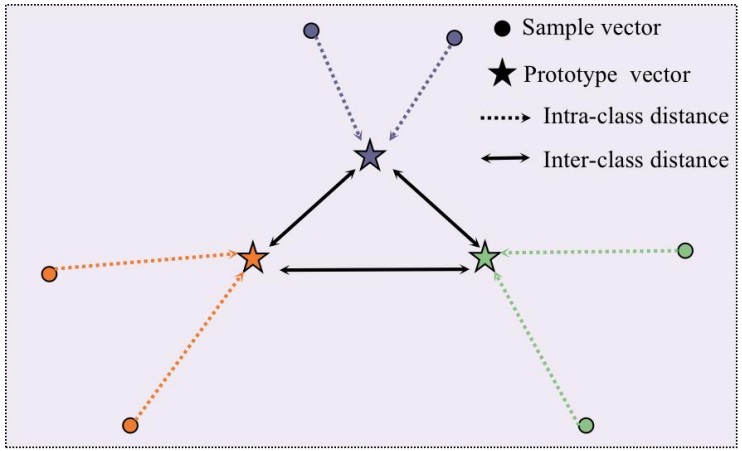
$$w_{\langle s_i, \phi_j \rangle} = \texttt{softmax}\left[-\frac{1}{L}\sum_{l=1}^{L} dis(f_\theta(s_i), f_\theta(q_l))\right]$$

Method

Task Sampler



Easy Task



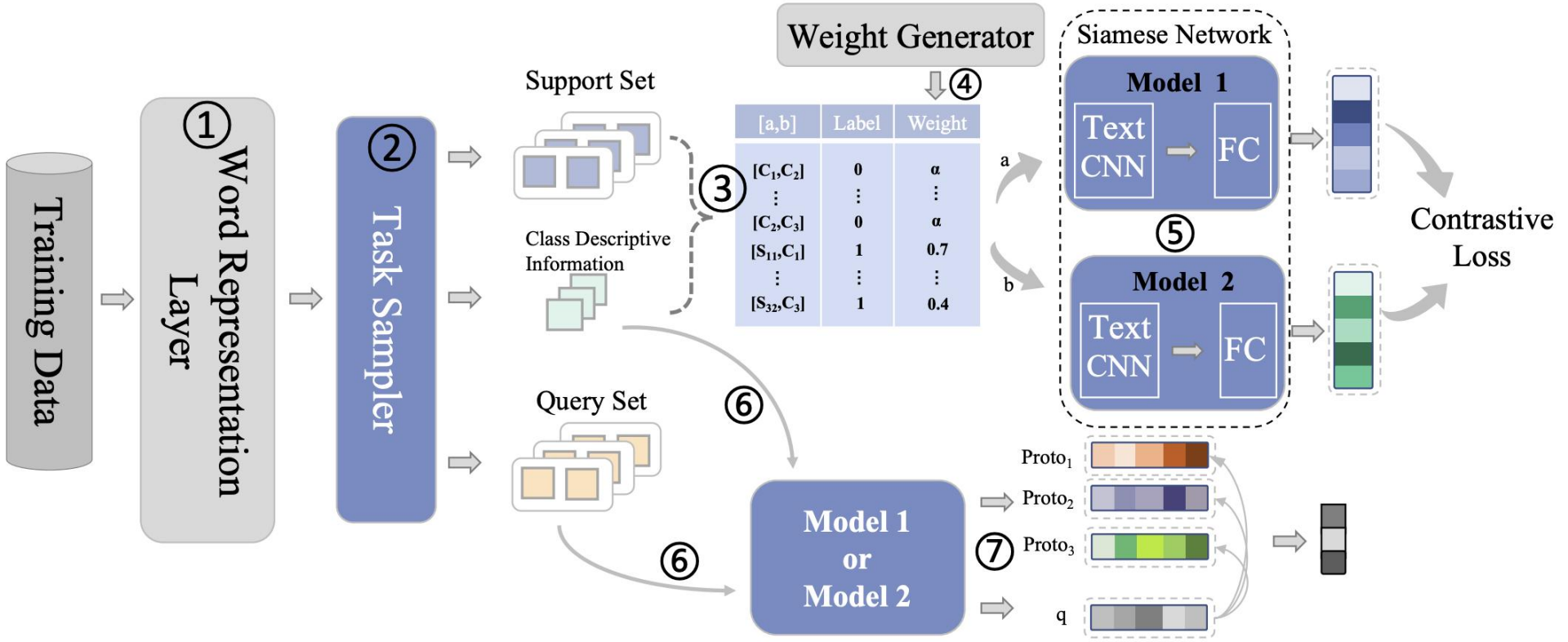
Hard Task

We assign higher sampling probability to tasks that are hard to classify.

The closer the distance between prototype vectors, the more difficult the corresponding classes can be separated; the more distant a sample in the support set is to the prototype vector, the more difficult the task will be.

Method

Model Architecture of Meta-SN



Experiments

Main Results

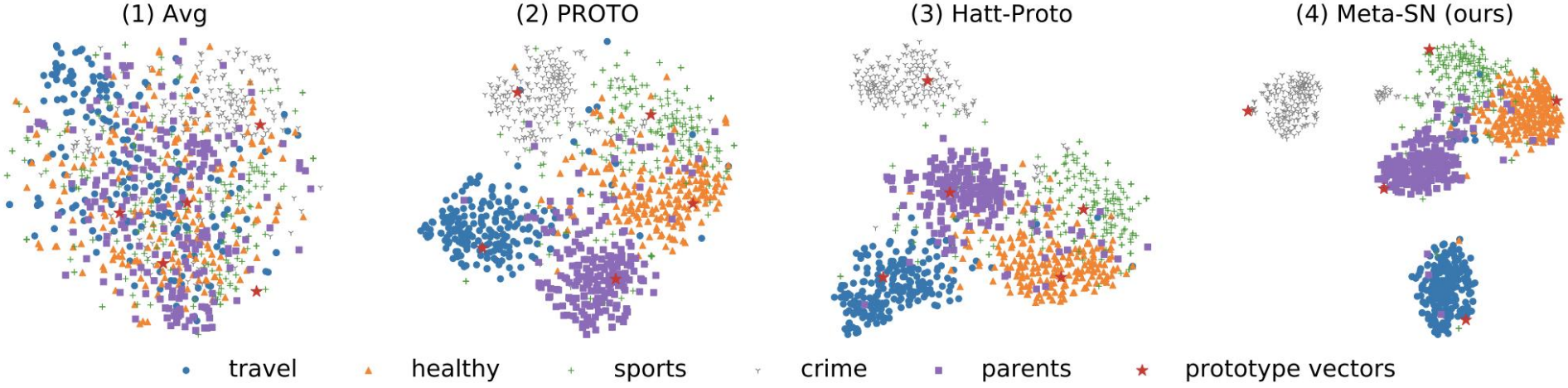
	Methods	HuffPost		Amazon		Reuters		20News		RCV1		FewRel		Average	
		1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
fastText-	MAML [5]	35.9	49.3	39.6	47.1	54.6	62.9	33.8	43.7	39.0	51.1	51.7	66.9	42.4	53.5
	PROTO [37]	35.7	41.3	37.6	52.1	59.6	66.9	37.8	45.3	32.1	35.6	49.7	65.1	42.1	51.1
	Induct [8]	38.7	49.1	34.9	41.3	59.4	67.9	28.7	33.3	33.4	38.3	50.4	56.1	40.9	47.6
	Hatt-Proto [6]	41.1	56.3	59.1	76.0	73.2	86.2	44.2	55.0	43.2	64.3	77.6	90.1	56.4	71.3
	DS-FSL [1]	43.0	63.5	62.6	81.1	81.8	96.0	52.1	68.3	54.1	75.3	67.1	83.5	60.1	78.0
	MLADA [10]	45.0	64.9	68.4	86.0	82.3	96.7	59.6	77.8	55.3	80.7	81.1	90.8	65.3	82.8
	Meta-SN	**54.7**	**68.5**	**70.2**	**87.7**	**84.0**	**97.1**	**60.7**	**78.9**	**60.0**	**86.1**	**84.8**	**93.1**	**69.1**	**85.2**
BERT-	ContrastNet [2]	52.7	64.4	75.4	85.2	86.2	95.3	71.0	81.3	65.7	87.4	85.3	92.7	72.7	84.3
	Meta-SN	**63.1**	**71.3**	**77.5**	**89.1**	**87.9**	**96.7**	**72.1**	**83.2**	**67.3**	**88.9**	**86.8**	**94.6**	**73.6**	**87.3**

Ablation Study

Models	HuffPost		Amazon		Reuters		20News		RCV1		FewRel		Average	
	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
Meta-SN-rpv	46.0	61.5	62.9	80.1	75.0	89.9	52.2	70.0	52.3	79.1	76.1	85.9	60.8	77.9
Meta-SN-ew	51.4	64.9	68.4	83.3	81.1	93.4	57.9	74.4	57.7	82.5	81.7	89.3	66.4	81.3
Meta-SN-rts	52.1	66.1	69.3	84.5	81.6	95.1	60.0	76.8	58.7	84.2	79.7	88.8	66.9	82.6
Meta-SN-ln	53.8	68.0	69.6	87.1	83.3	96.0	59.8	78.0	59.5	85.4	84.3	92.6	68.4	84.5
Meta-SN	**54.7**	**68.5**	**70.2**	**87.7**	**84.0**	**97.1**	**60.7**	**78.9**	**60.0**	**86.1**	**84.8**	**93.1**	**69.1**	**85.2**

Experiments

Visualization



These results show the superiority of Meta-SN in generating high-quality sentence embeddings.

Conclusion

DaSE Data Science &Engineering

- We construct the prototype vectors with the external descriptive information of class labels and further refines these vectors with a Siamese network, which alleviates the adverse impact of sampling randomness.
- We further learn the importance of a labeled sample by considering its average distance to the query set.
- We present an effective sampling strategy to construct meta-tasks, which assigns higher sampling probability to the hard-to-classify samples. This boosts the model's generalization ability.

Thank you!