scenes

task

TART: Improved Few-shot Text Classification Using Task-Adaptive Reference Transformation

method

Advisor : Jia-Ling, Koh

Speaker : Ting-I, Weng

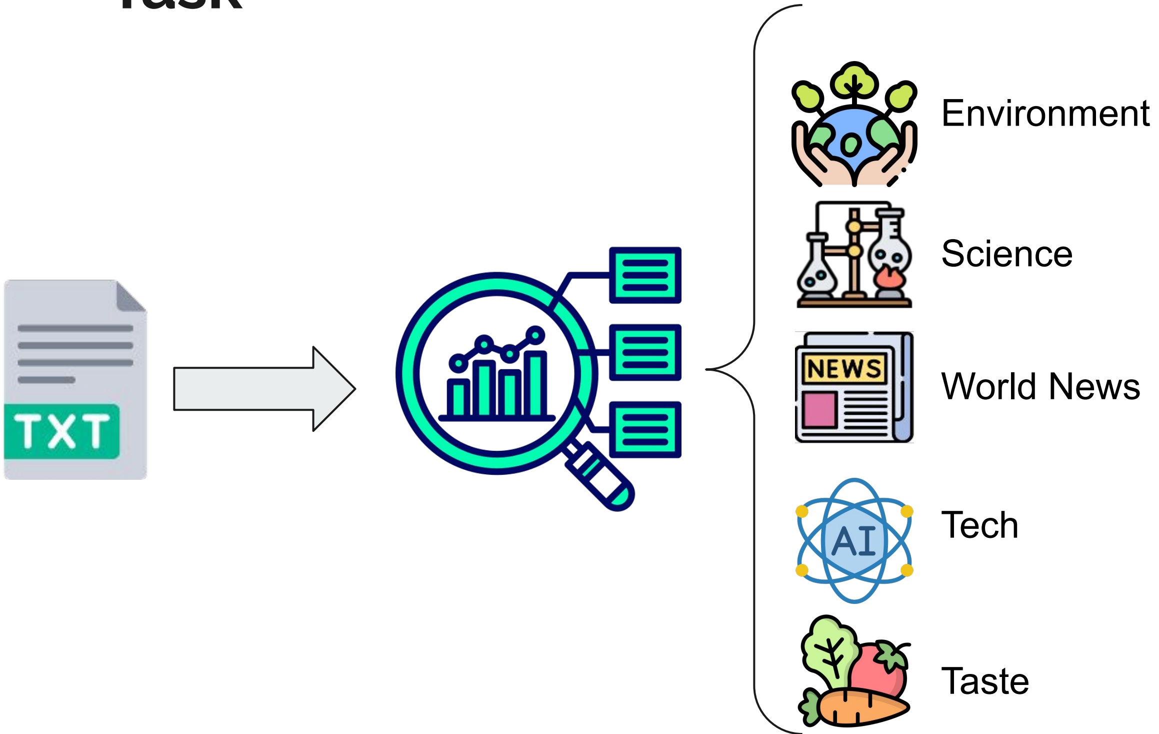
Source : ACL'23

Date : 2024/03/05

Outline

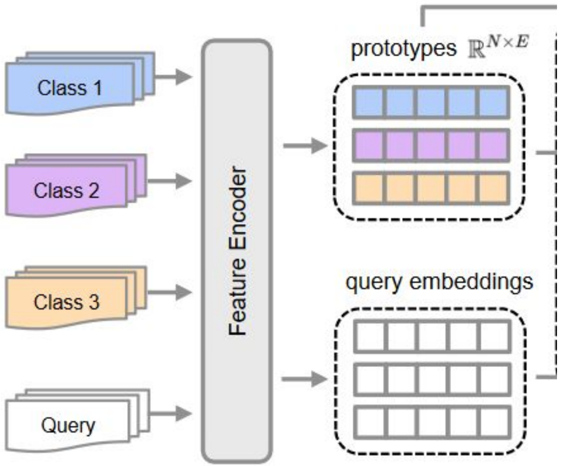
- Introduction
- Method
- Experiment
- Conclusion

Task

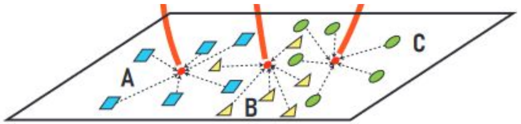


text classification

Problem

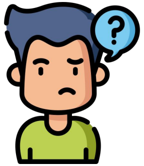


Indistinguishable between categories



original feature space

Issue



query

method

Class	Testing Sample	Task 1: Support class: 1,2,3,4 MLADA	Task 1: Support class: 1,2,3,4 Ours	Task 2: Support class: 1,2,3,5 MLADA	Task 2: Support class: 1,2,3,5 Ours
1	Animal photos of the week: baby tiger goes for a swim.	1	1	1	1
2	Twitter helps confirm X-shaped bulge at Center of Milky Way.	4	2	2	2
3	Toronto van attack suspect's Facebook post praised misogynist mass killer.	4	3	2	3
4	Apple just solved one of the iphone's most harmful features.	2	4	-	-
5	Apple fritter season is here, and so are the recipes you'll need.	-	-	5	5

Class 1: Environment *Class 2: Science* *Class 3: World News* *Class 4: Tech* *Class 5: Taste*

technology company

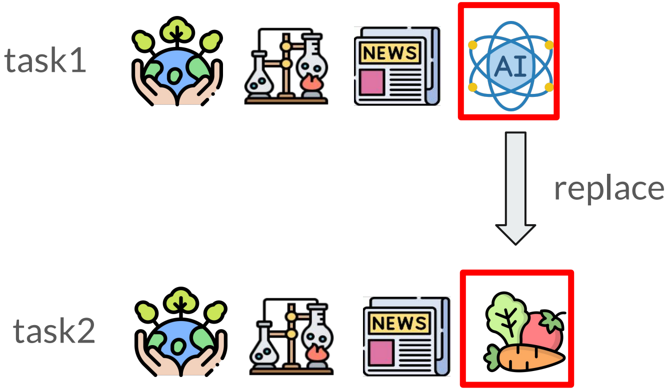
class similar

Guess

method

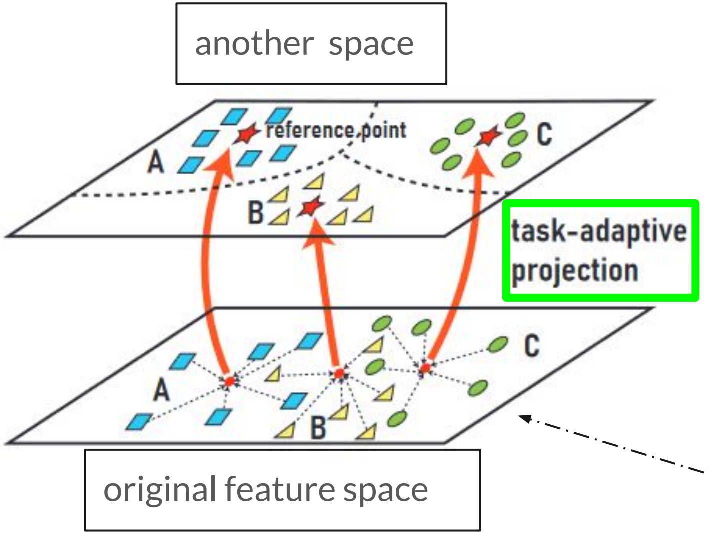
Class	Testing Sample	Task 1: Support class: 1,2,3,4		Task 2: Support class: 1,2,3,5	
		MLADA	Ours	MLADA	Ours
1	Animal photos of the week: baby tiger goes for a swim.	1	1	1	1
2	Twitter helps confirm X-shaped bulge at Center of Milky Way.	4	2	2	2
3	Toronto van attack suspect's Facebook post praised misogynist mass killer.	4	3	2	3
4	Apple just solved one of the iphone's most harmful features.	2	4	-	-
5	Apple fritter season is here, and so are the recipes you'll need.	-	-	5	5

Class 1: Environment *Class 2: Science* *Class 3: World News* *Class 4: Tech* *Class 5: Taste*



consider the inter-class variance of support sets

Solution



helpful to enhance the divergence between class prototypes

another space

Task-Adaptive Reference Transfer Module

projection

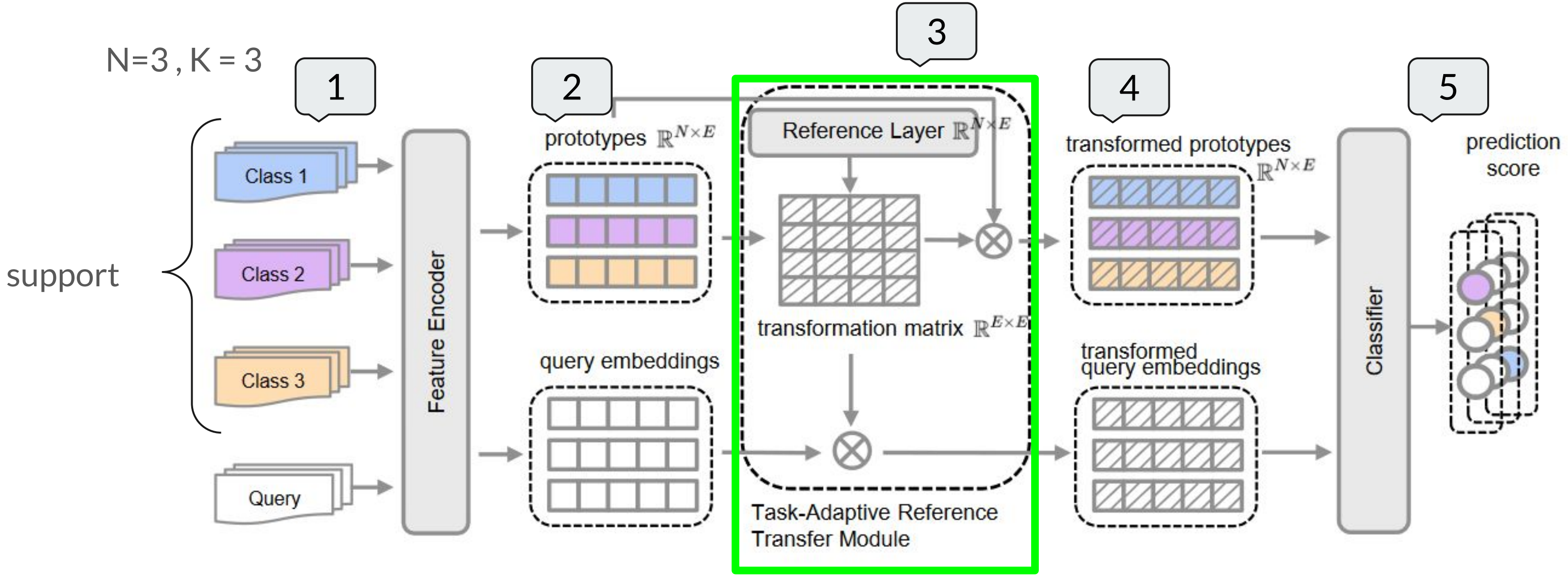
cannot distinguish between categories

Outline

- Introduction
- Method
- Experiment
- Conclusion

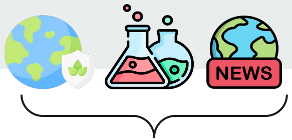
Architecture

$$C_{train} \cap C_{test} = \emptyset.$$



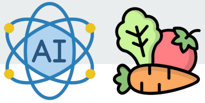
train

support



query

test

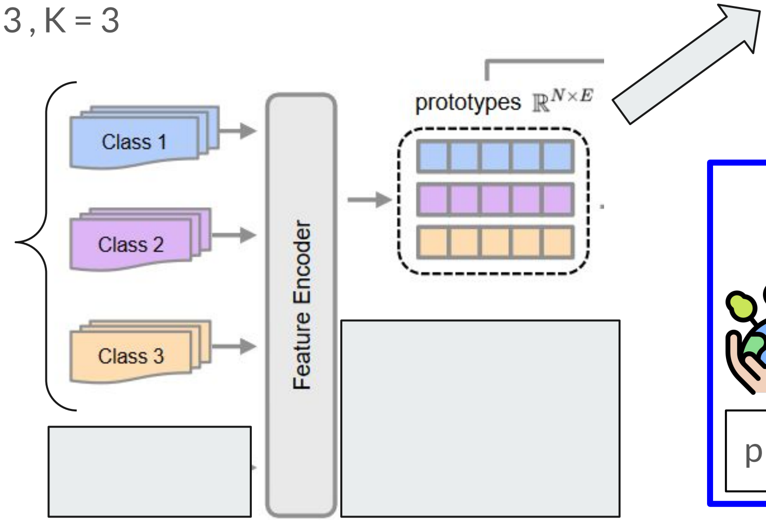


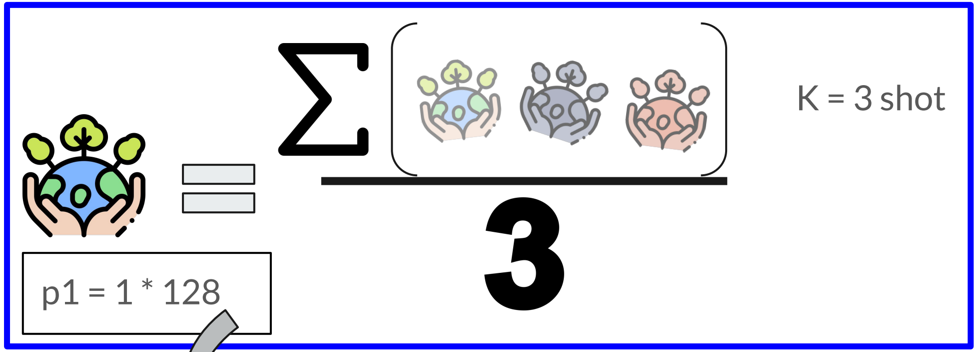
Prototype matrix

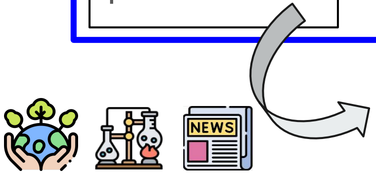
$$\boldsymbol{p}_c = \frac{1}{|S_c|} \sum_{(\boldsymbol{x}_i, y_i) \in S_c} f_\theta(\boldsymbol{x}_i),$$

N=3 , K = 3

support







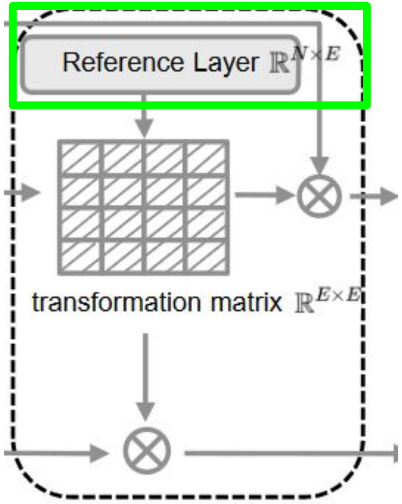
P = 3 * 128

generate reference vector

Reference layer

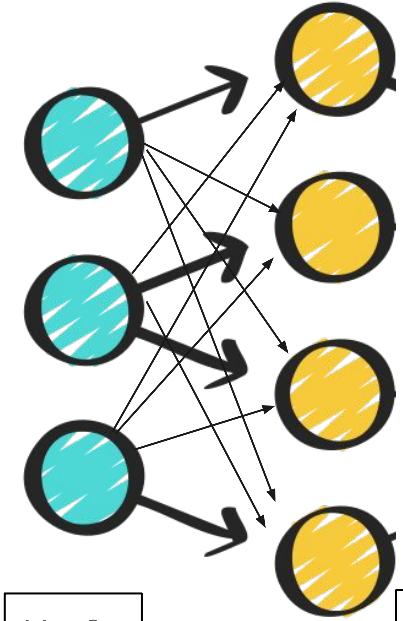
reference vectors $\{r_1, \ldots, r_N\}$

N = 3



random init

linear layer



N = 3

E = 128

r1 = 1 * 128

r2 = 1 * 128

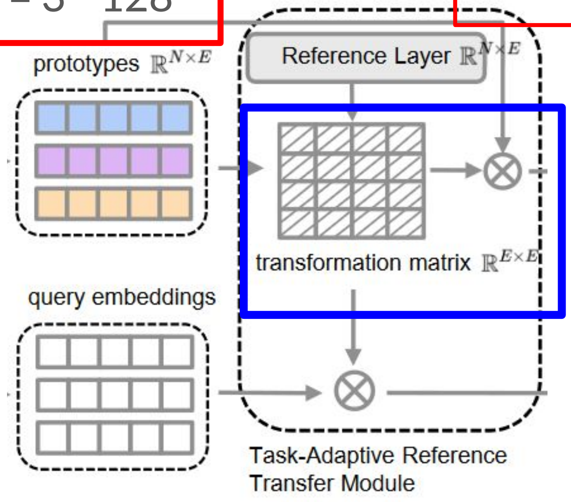
r3 = 1 * 128

R = 3 * 128

Transformation matrix

P = 3 * 128

R = 3 * 128



$$P\boxed{W} = R$$

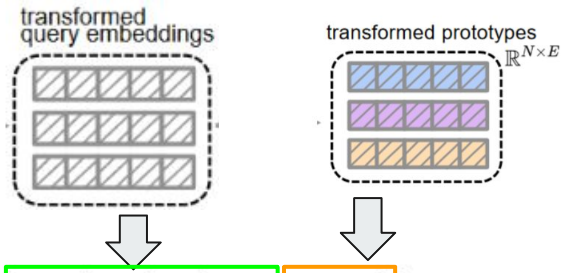
$$P^{+}PW = P^{+}R$$

normal equation $\Rightarrow$ $P^{+} = \{P^{T}P\}^{-1}P^{T}$

$$W = P^{+}R$$

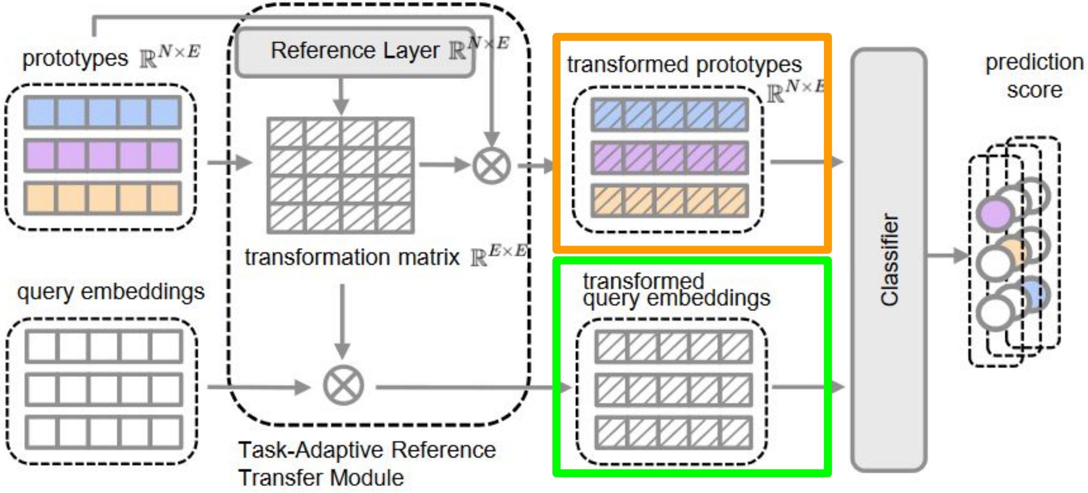
W = 128*128

Predict query



softmax

$$p(y=c|\boldsymbol{x}_q)=\frac{\exp\left(\boxed{-d(f_\theta(\boldsymbol{x}_q)W,}\ \boxed{\boldsymbol{p}_c W)}\right)}{\sum_{\boldsymbol{p}_c\in\mathcal{P}}\exp\left(-d(f_\theta(\boldsymbol{x}_q)W,\boldsymbol{p}_c W)\right)}$$



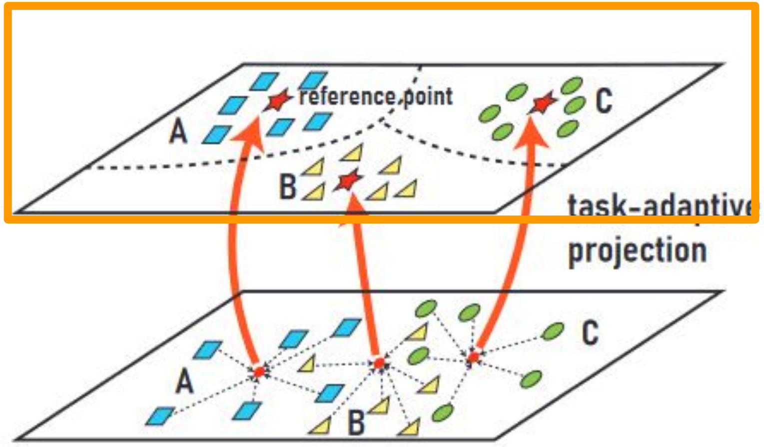
Classification Loss

$$\mathcal{L}_{cls} = \frac{1}{|Q|} \sum_{\boldsymbol{x}_q \in Q} [d(f_\theta(\boldsymbol{x}_q)W, \boldsymbol{p}_c W) + \log \sum_{\boldsymbol{p}_c \in \mathcal{P}} \exp\left(-d(f_\theta(\boldsymbol{x}_q)W, \boldsymbol{p}_c W)\right)]$$

$$= \frac{-\log \left\{ p(y = c|\boldsymbol{x}_q) = \overbrace{\frac{\exp\left(-d(f_\theta(\boldsymbol{x}_q)W, \boldsymbol{p}_c W)\right)}{\sum_{\boldsymbol{p}_c \in \mathcal{P}} \exp\left(-d(f_\theta(\boldsymbol{x}_q)W, \boldsymbol{p}_c W)\right)}}^{\text{softmax}} \right\}}{|Q|}$$

= Negative Log Likelihood (NLL loss)

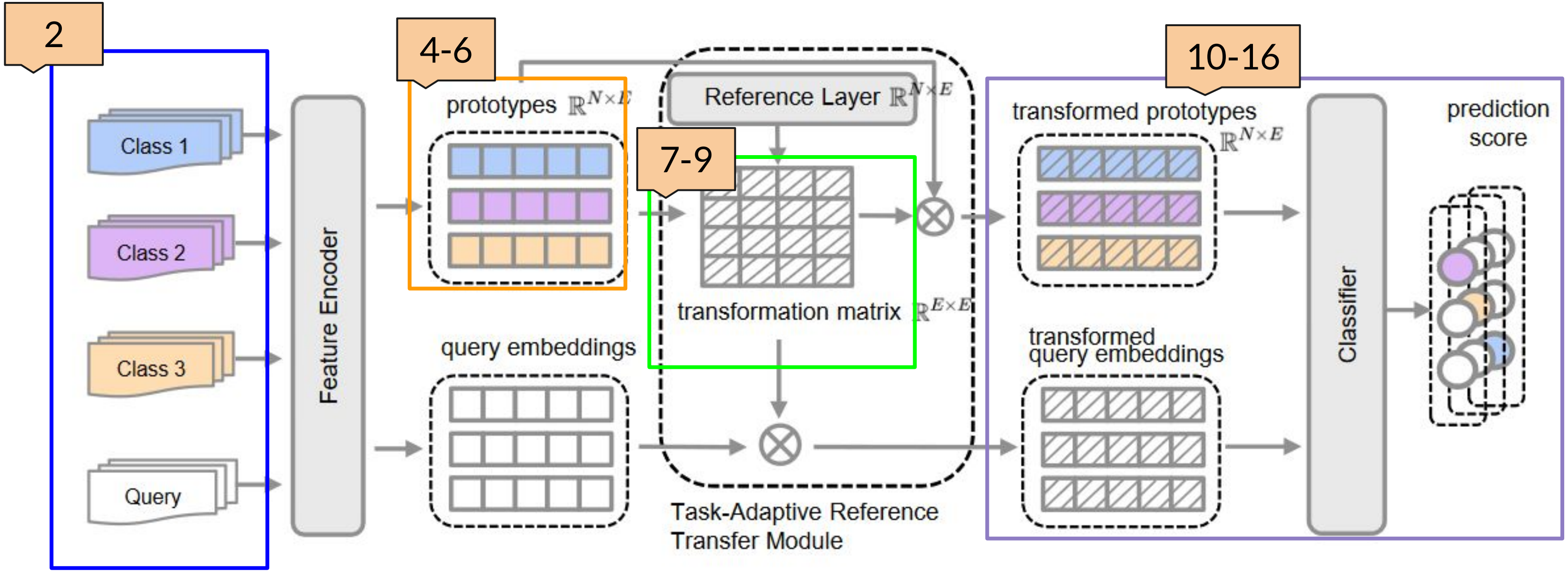
Discriminative Reference Regularization



Maximize distance between prototypes

$$\mathcal{L}_{drr} = \sum_{i \neq j, p \in \mathcal{P}} -d(\boxed{\mathbf{p}_i W, \mathbf{p}_j W})$$

Algorithm

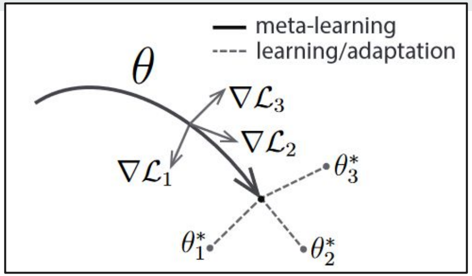

2
4-6
7-9
10-16
Class 1
Class 2
Class 3
Query
Feature Encoder
prototypes $\mathbb{R}^{N\times E}$
Reference Layer $\mathbb{R}^{N\times E}$
transformation matrix $\mathbb{R}^{E\times E}$
query embeddings
Task-Adaptive Reference Transfer Module
transformed prototypes $\mathbb{R}^{N\times E}$
transformed query embeddings
Classifier
prediction score

Outline

- Introduction
- Method
- Experiment
- Conclusion

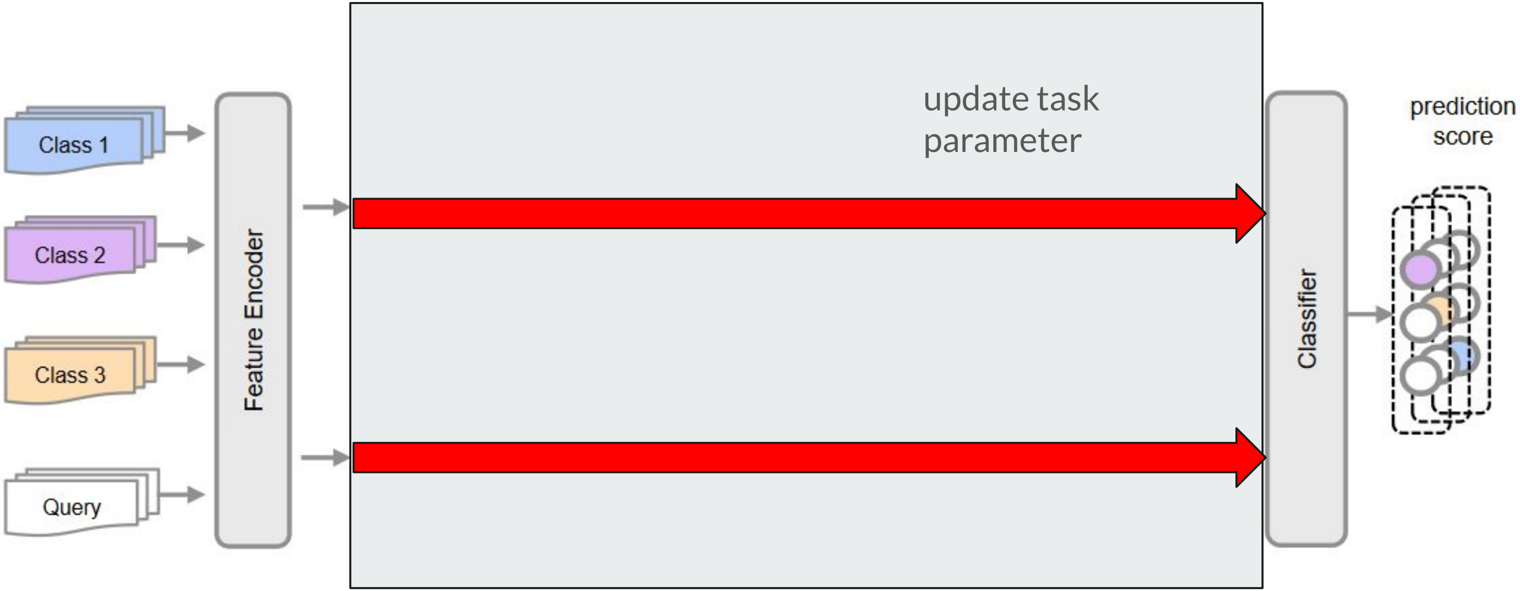
Datasets

Datasets	Content	E.g.	Avg. # tokens/sample	class	samples	# train/val/test classes
HuffPost headlines	news headlines	犯罪、娛樂、世界新聞、政治 ...	11	41	36900	20/5/16
Amazon product data	product reviews	書、電子、電影、電玩遊戲 ...	140	24	24000	10/5/9
Reuters-21578	Reuters Articles	貿易、糧食、原油、植物油、黃金 ...	168	31	620	15/5/11
20 Newsgroups	newsgroups	電子、醫學、宗教、政治、電腦硬體 ...	340	20	18820	8/5/7



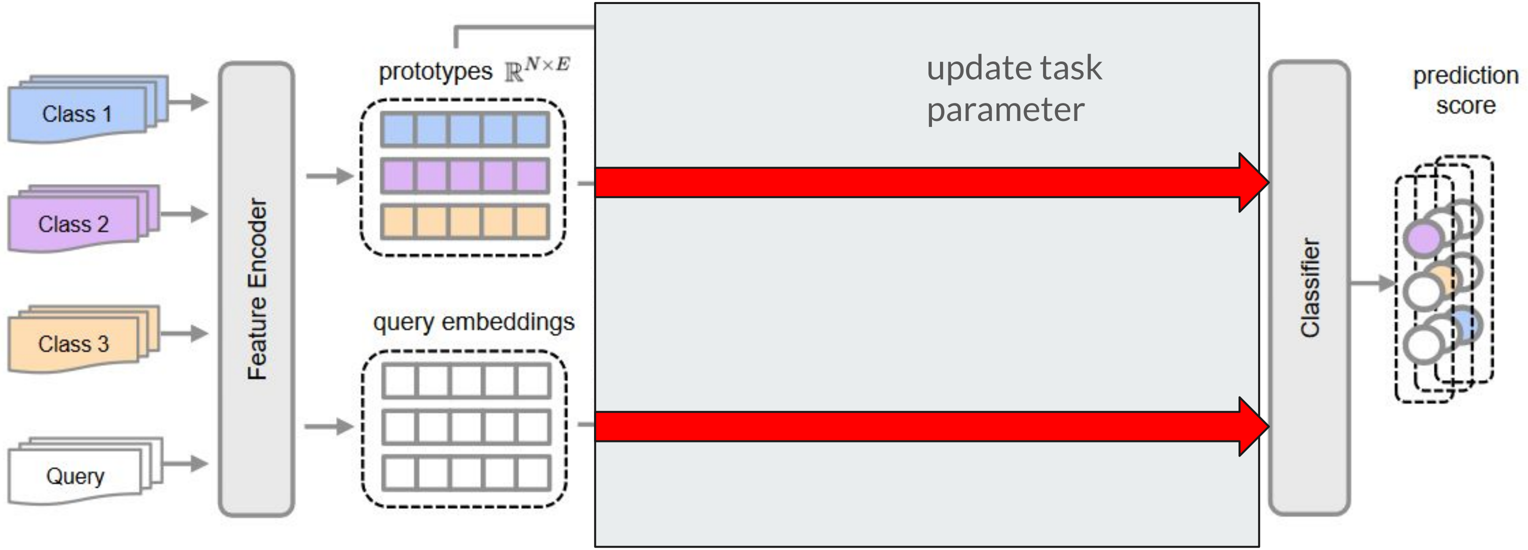
meta framework

Baseline - MAML



prototype

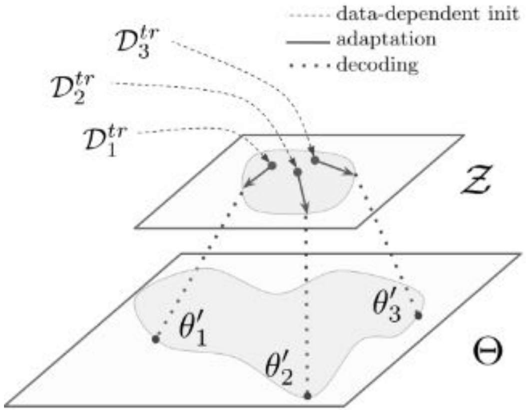
Baseline - PROTO

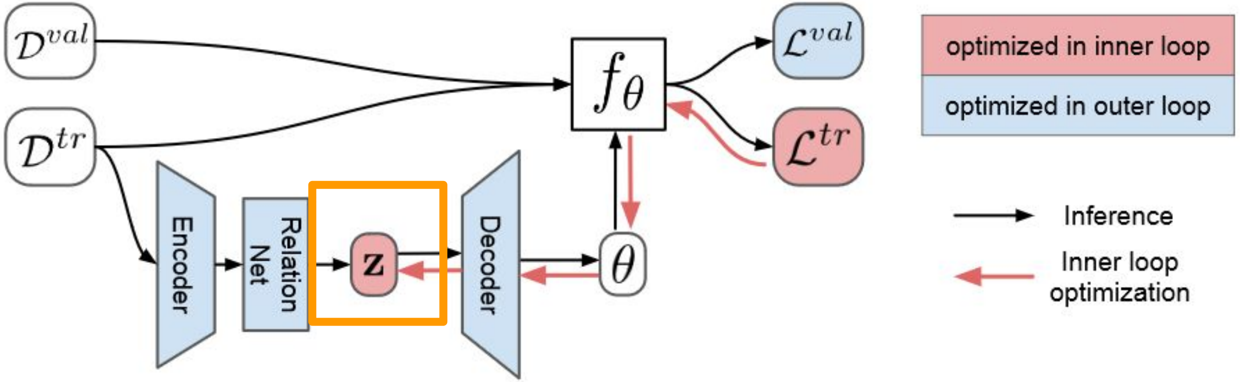


learns a low-dimensional latent embedding

Baseline - Latent Embedding Optimization (LEO)

1 shot、5 shot -> model overfitting

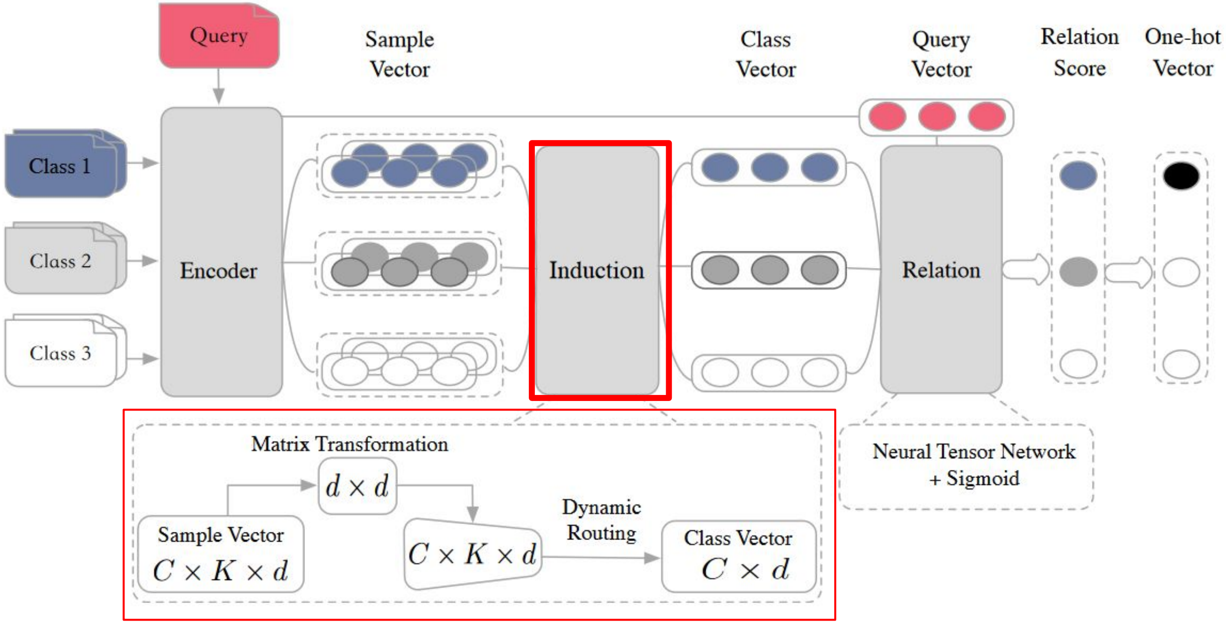




prototype with dynamic routing

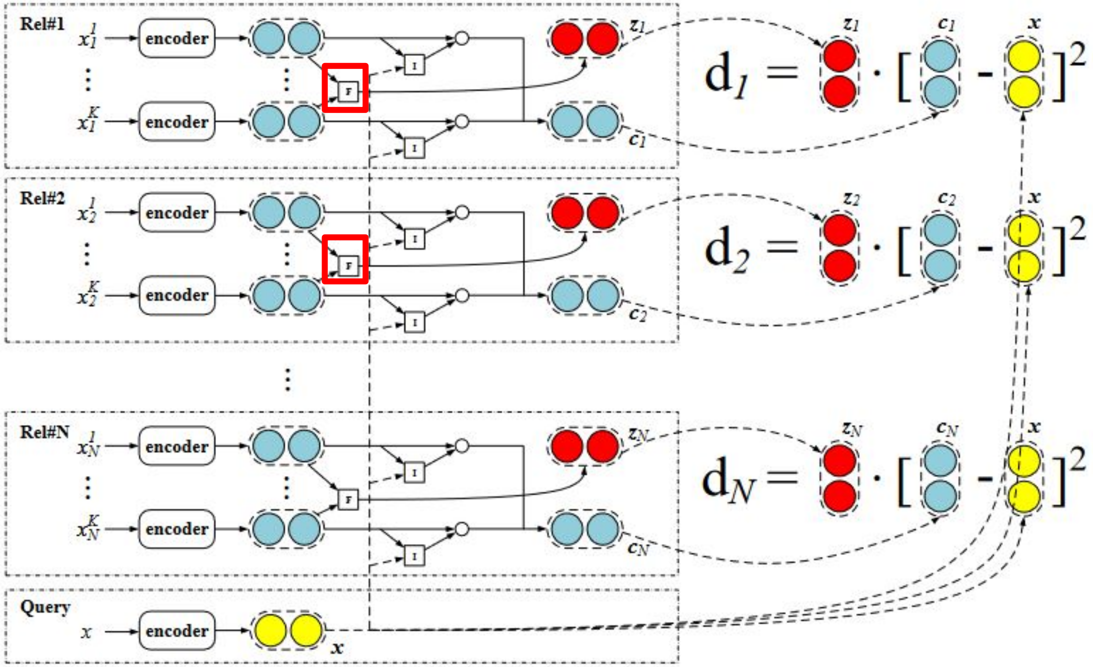
Baseline - Induction Networks

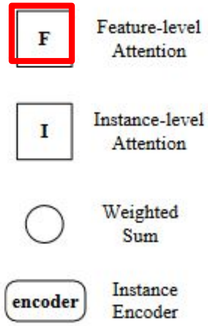
learn a general representation of each class in the support set and then compare it to new queries



hybrid attention-based prototype

Baseline - Hybrid Attention(HATT)





透過每個input feature計算一個重要性分數

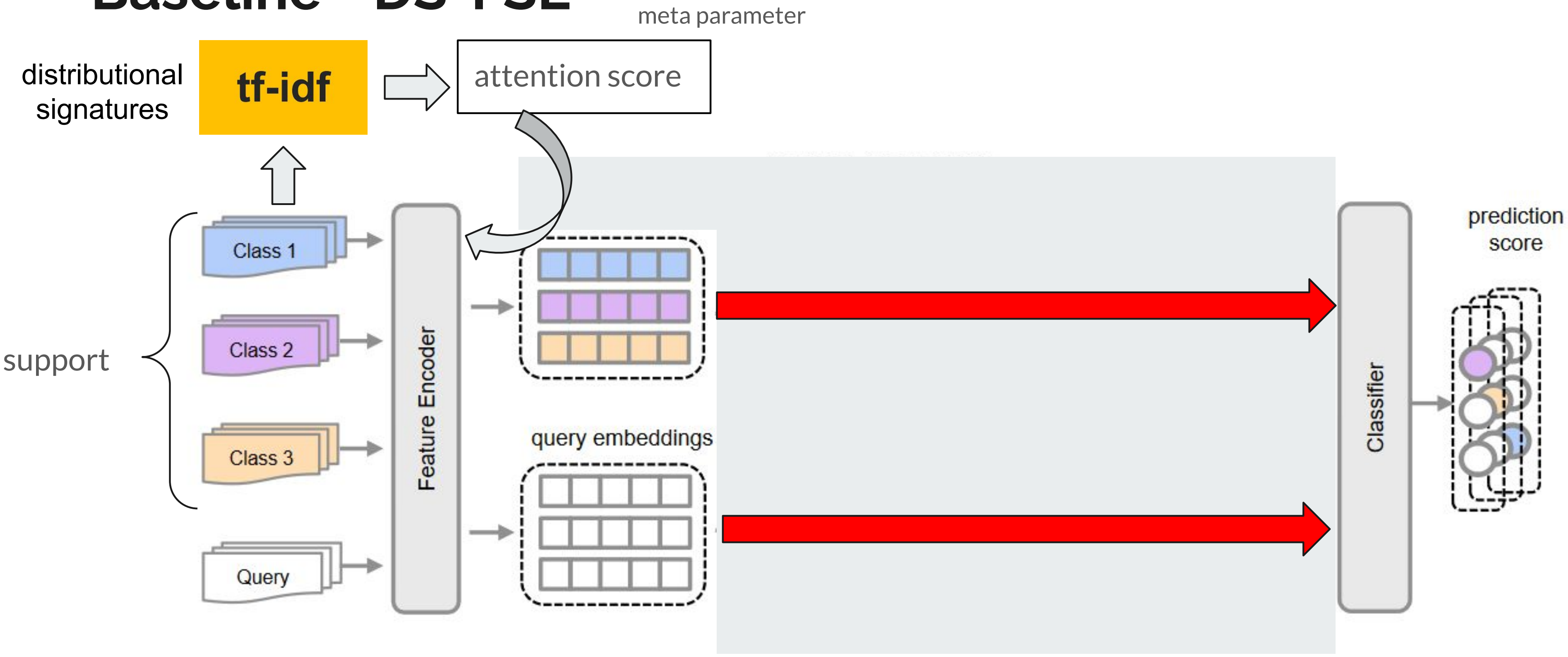
透過注意力，選擇與query 最相似的instance

$$\mathbf{c}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_i^j, \Rightarrow \mathbf{c}_i = \sum_{j=1}^{n_i}\alpha_j\mathbf{x}_i^j.$$

prototype

Attention-based prototype

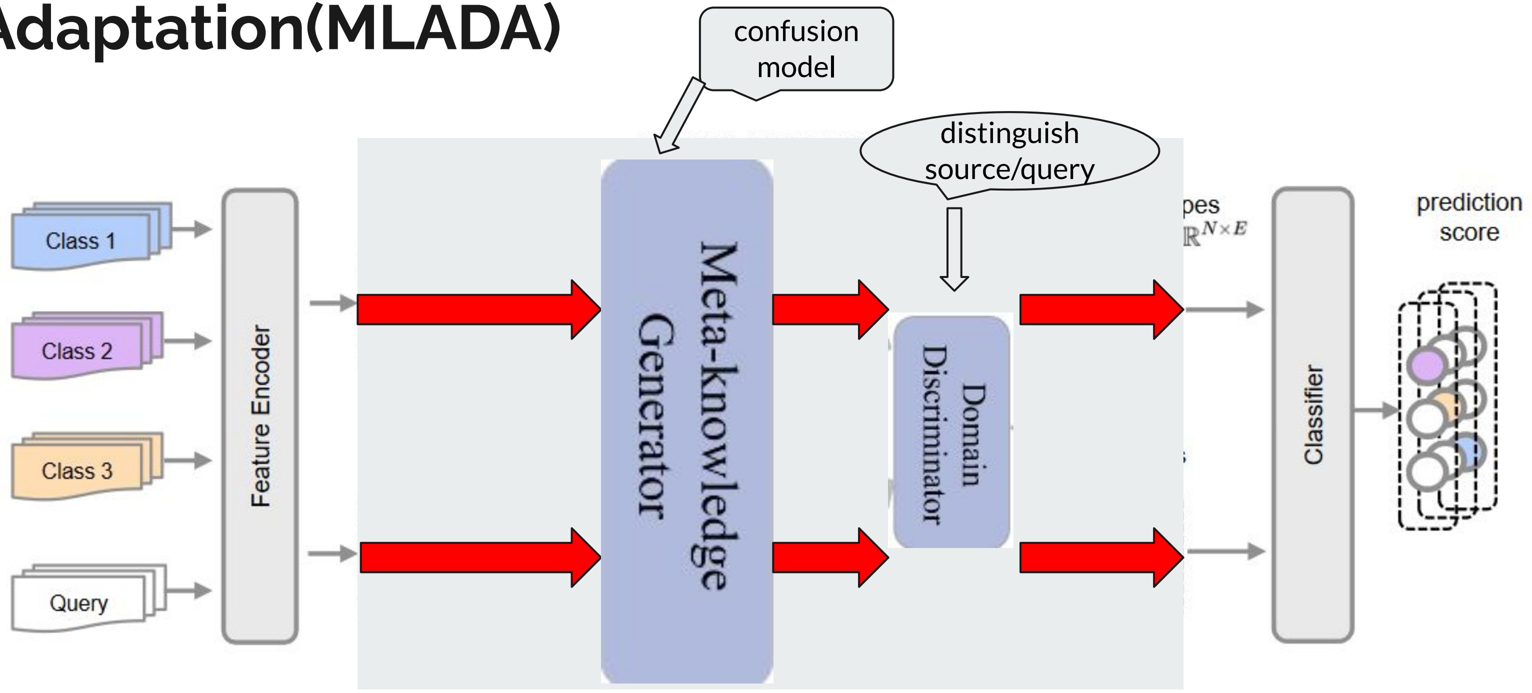
Baseline - DS-FSL



domain adversarial network + meta-learning
= transferable features

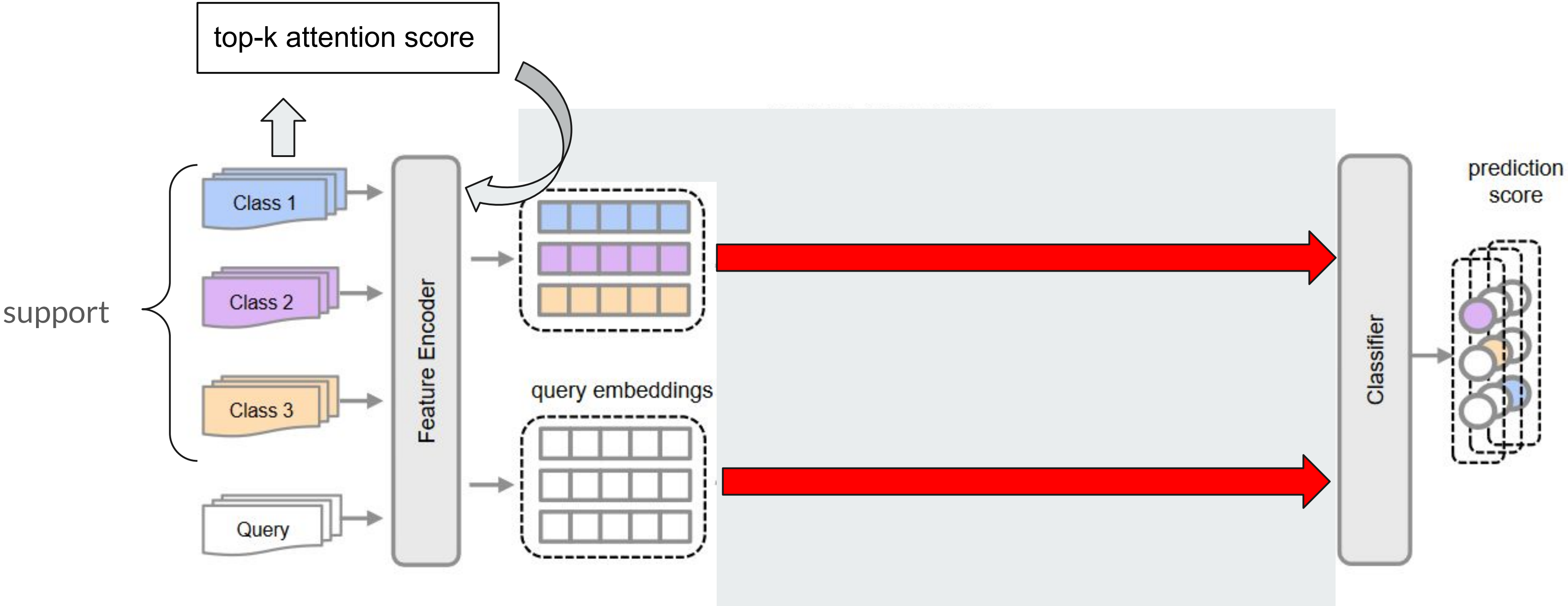
Adversarial Domain Adaptation

Baseline - Meta-Learning Adversarial Domain Adaptation(MLADA)

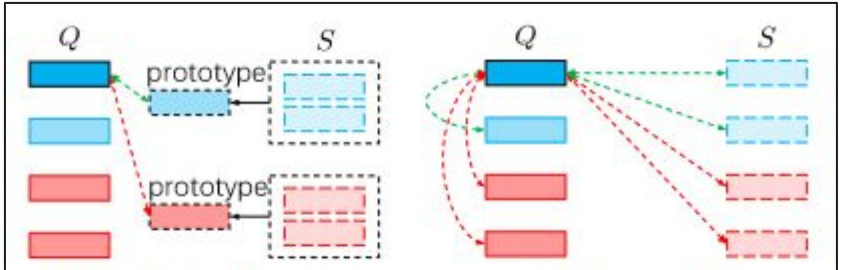


Attention + Meta

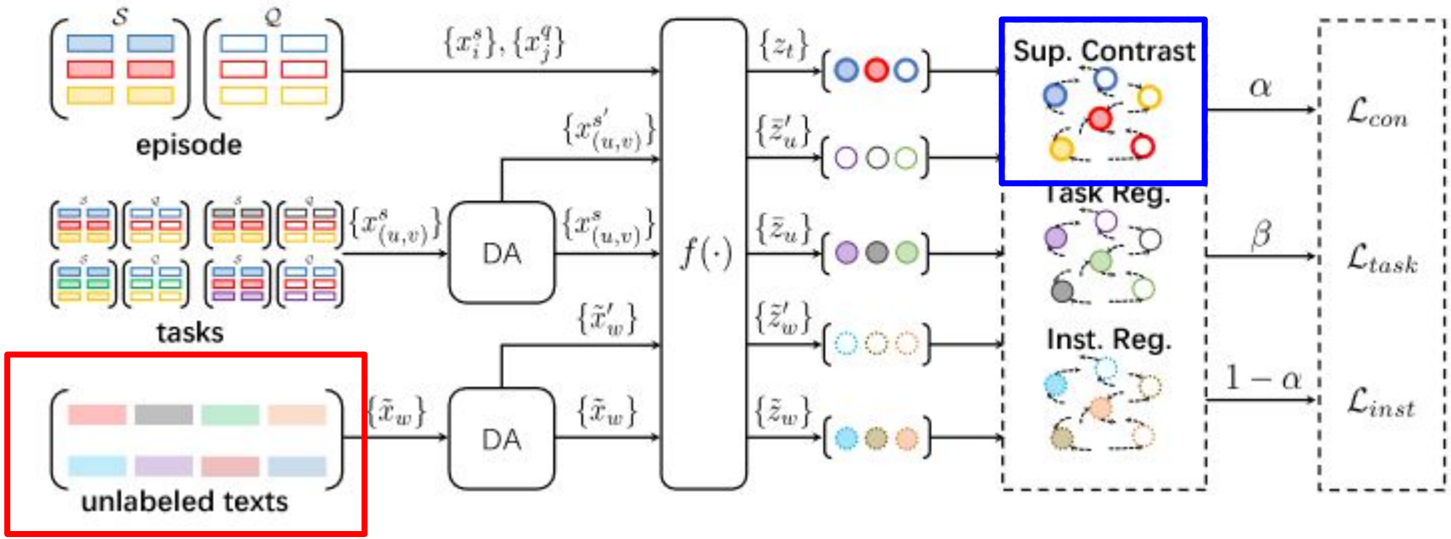
Baseline - LEarning-to-Attend(LEA)



Baseline - ContrastNet



(a) Prototypical Networks (b) ContrastNet



Experiment

Method	HuffPost		Amazon		Reuters		20 News		Average	
	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
MAML (2017)	35.9	49.3	39.6	47.1	54.6	62.9	33.8	43.7	40.9	50.8
PROTO (2017)	35.7	41.3	37.6	52.1	59.6	66.9	37.8	45.3	42.7	51.4
LEO* (2018)	28.8	42.3	39.5	52.5	35.4	54.1	36.4	52.2	35.0	50.3
Induct (2019)	38.7	49.1	34.9	41.3	59.4	67.9	28.7	33.3	40.4	47.9
HATT (2019)	41.1	56.3	49.1	66.0	43.2	56.2	44.2	55.0	44.4	58.4
DS-FSL (2020)	43.0	63.5	62.6	81.1	81.8	96.0	52.1	68.3	59.9	77.2
MLADA (2021)	45.0	64.9	68.4	**86.0**	82.3	96.7	59.6	77.8	63.9	81.4
LEA (2022)	46.2	65.8	66.5	83.5	69.0	89.0	54.1	60.2	58.9	74.6
TART	46.9	**66.8**	**70.1**	82.4	**92.2**	**96.7**	**67.0**	**83.2**	**69.0**	**82.3**

solve time-consuming

Experiment

Method	HuffPost		Amazon		Reuters		20 News		Average	
	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
MAML (2017)	35.9	49.3	39.6	47.1	54.6	62.9	33.8	43.7	40.9	50.8
PROTO (2017)	35.7	41.3	37.6	52.1	59.6	66.9	37.8	45.3	42.7	51.4
LEO* (2018)	28.8	42.3	39.5	52.5	35.4	54.1	36.4	52.2	35.0	50.3
Induct (2019)	38.7	49.1	34.9	41.3	59.4	67.9	28.7	33.3	40.4	47.9
HATT (2019)	41.1	56.3	49.1	66.0	43.2	56.2	44.2	55.0	44.4	58.4
DS-FSL (2020)	43.0	63.5	62.6	81.1	81.8	96.0	52.1	68.3	59.9	77.2
MLADA (2021)	45.0	64.9	68.4	**86.0**	82.3	96.7	59.6	77.8	63.9	81.4
LEA (2022)	46.2	65.8	66.5	83.5	69.0	89.0	54.1	60.2	58.9	74.6
TART	46.9	**66.8**	**70.1**	82.4	**92.2**	**96.7**	**67.0**	**83.2**	**69.0**	**82.3**

solve time-consuming

Experiment

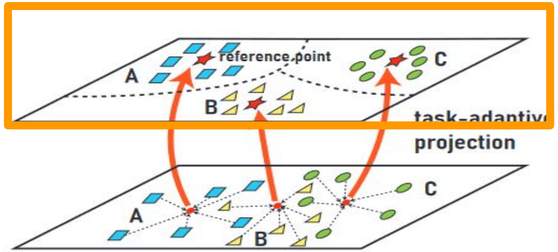
Method	HuffPost		Amazon		Reuters		20 News		Average	
	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
MAML (2017)	35.9	49.3	39.6	47.1	54.6	62.9	33.8	43.7	40.9	50.8
PROTO (2017)	35.7	41.3	37.6	52.1	59.6	66.9	37.8	45.3	42.7	51.4
LEO* (2018)	28.8	42.3	39.5	52.5	35.4	54.1	36.4	52.2	35.0	50.3
Induct (2019)	38.7	49.1	34.9	41.3	59.4	67.9	28.7	33.3	40.4	47.9
HATT (2019)	41.1	56.3	49.1	66.0	43.2	56.2	44.2	55.0	44.4	58.4
DS-FSL (2020)	43.0	63.5	62.6	81.1	81.8	96.0	52.1	68.3	59.9	77.2
MLADA (2021)	45.0	64.9	68.4	**86.0**	82.3	96.7	59.6	77.8	63.9	81.4
LEA (2022)	46.2	65.8	66.5	83.5	69.0	89.0	54.1	60.2	58.9	74.6
TART	46.9	**66.8**	**70.1**	82.4	**92.2**	**96.7**	**67.0**	**83.2**	**69.0**	**82.3**

solve time-consuming

Experiment

Method	HuffPost		Amazon		Reuters		20 News		Average	
	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
MAML (2017)	35.9	49.3	39.6	47.1	54.6	62.9	33.8	43.7	40.9	50.8
PROTO (2017)	35.7	41.3	37.6	52.1	59.6	66.9	37.8	45.3	42.7	51.4
LEO* (2018)	28.8	42.3	39.5	52.5	35.4	54.1	36.4	52.2	35.0	50.3
Induct (2019)	38.7	49.1	34.9	41.3	59.4	67.9	28.7	33.3	40.4	47.9
HATT (2019)	41.1	56.3	49.1	66.0	43.2	56.2	44.2	55.0	44.4	58.4
DS-FSL (2020)	43.0	63.5	62.6	81.1	81.8	96.0	52.1	68.3	59.9	77.2
MLADA (2021)	45.0	64.9	68.4	**86.0**	82.3	96.7	59.6	77.8	63.9	81.4
LEA (2022)	46.2	65.8	66.5	83.5	69.0	89.0	54.1	60.2	58.9	74.6
TART	46.9	**66.8**	**70.1**	82.4	**92.2**	**96.7**	**67.0**	**83.2**	**69.0**	**82.3**

solve time-consuming



$$\mathcal{L}_{drr} = \sum_{i \neq j, p \in \mathcal{P}} -d(\boldsymbol{p}_i W, \boldsymbol{p}_j W)$$

Maximize distance between prototypes

Ablation Study - Discriminative Reference Regularization(DRR)

Method	HuffPost		Amazon		Reuters		20 News		Average	
	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
TART w/o DRR	**48.4**	66.0	68.9	83.5	90.4	96.2	66.4	82.2	68.5	81.9
TART	46.9	**66.8**	**70.1**	82.4	**92.2**	**96.7**	**67.0**	**83.2**	**69.0**	**82.3**

- PLM denotes prompting language model
- EK denotes extra knowledge (unlabeled data)

Ablation Study - Using BERT

top-k attention
GNN
prompt-based
contrast-base

Method	PLM	EK	HuffPost		Amazon		Reuters		20 News		Average	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
TART	×	×	46.5	68.9	73.7	84.3	86.9	95.6	**73.2**	**84.9**	70.1	**83.4**
TART with fastText + BiLSTM			46.9	**66.8**	**70.1**	82.4	**92.2**	**96.7**	**67.0**	**83.2**	**69.0**	**82.3**

Google BERT

bert has richer semantic representation than fastText

Outline

- Introduction
- Method
- Experiment
- Conclusion

Conclusion

- propose a novel TART for fewshot text classification
- enhance the generalization by transforming the class prototypes to per-class fixed reference points in task-adaptive metric spaces
- discriminative reference regularization to maximize divergence between transformed prototypes