Few-shot Text Classification with Distributional Signatures

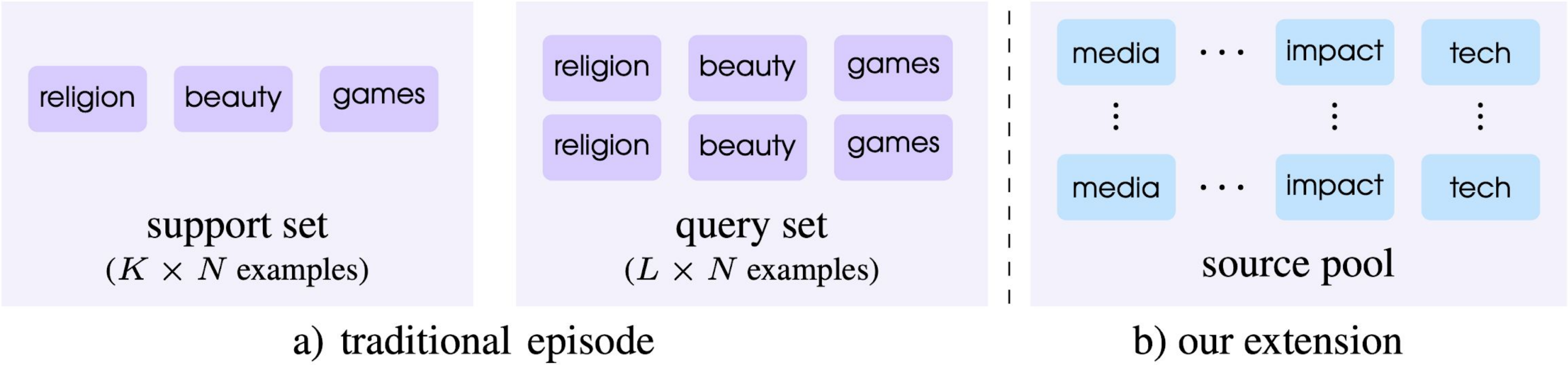
Yujia Bao, Menghua Wu, Shiyu Chang, Regina Barzilay

ICLR 2020

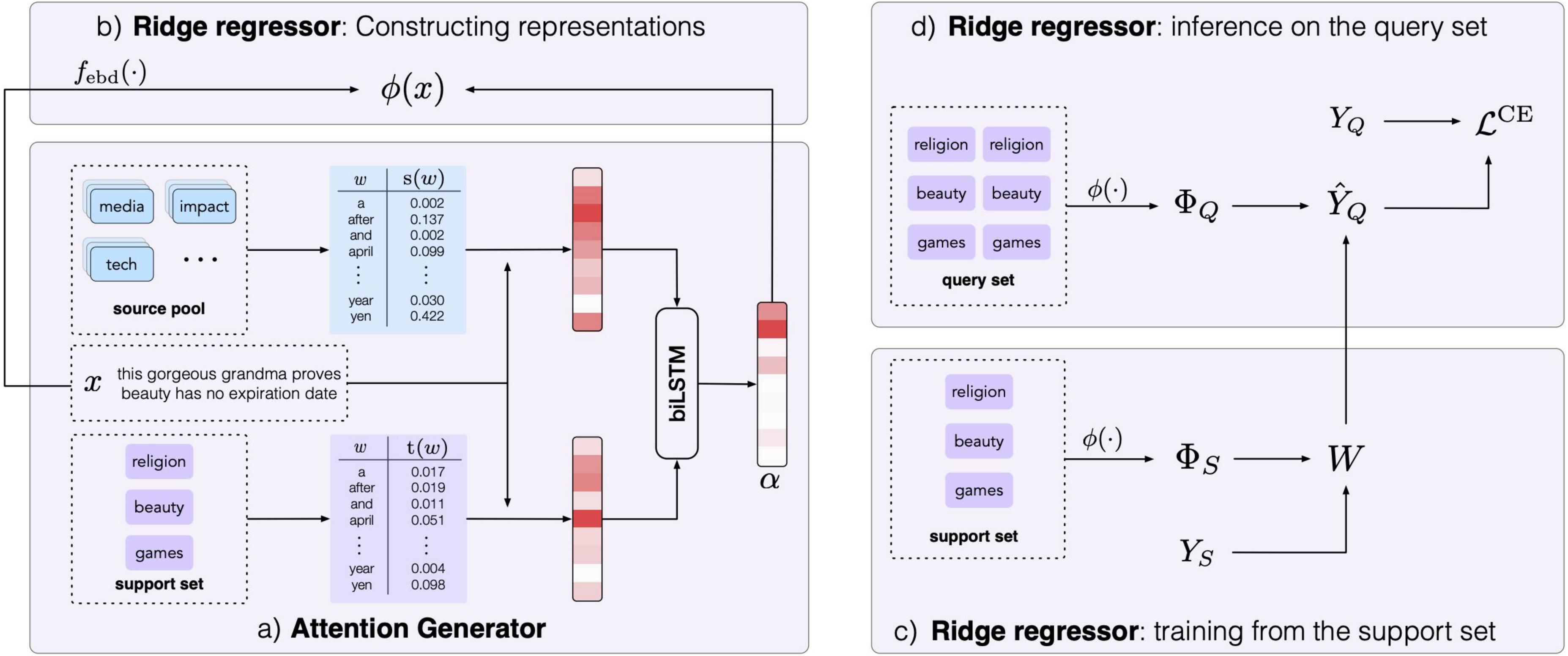
Motivation

1. In meta-learning, learning soly from word is not enough
    1. Matching information
    2. Interaction
    3. Underlying distribution
2. Model word’s distributional signatures across classes
3. Using this distributional signature as attention weight

Settings


religion
beauty
games
support set
($K \times N$ examples)
religion
beauty
games
religion
beauty
games
query set
($L \times N$ examples)
media
impact
tech
media
impact
tech
source pool
a) traditional episode
b) our extension

Framework



b) **Ridge regressor**: Constructing representations

a) **Attention Generator**

c) **Ridge regressor**: training from the support set

d) **Ridge regressor**: inference on the query set

w	$\mathrm{s}(w)$
a	0.002
after	0.137
and	0.002
april	0.099
⋮	⋮
year	0.030
yen	0.422

w	$\mathrm{t}(w)$
a	0.017
after	0.019
and	0.011
april	0.051
⋮	⋮
year	0.004
yen	0.098

x this gorgeous grandma proves beauty has no expiration date

Attention Generator

Distribution from the pool

$$\mathrm{s}(x_i) := \frac{\varepsilon}{\varepsilon + \mathrm{P}\left(x_i\right)}$$

Distribution learned from the support set

$$\mathrm{t}(x_i) := \mathcal{H}(\mathrm{P}\left(y \mid x_i\right))^{-1}$$

Attention weight learned from the support set

$$h = \mathrm{biLSTM}([\mathrm{s}(x); \mathrm{t}(x)])$$

$$\alpha_i := \frac{\exp\left(v^T h_i\right)}{\sum_j \exp\left(v^T h_j\right)}$$

Ridge regressor

Construct sentence representation

$$\phi(x) := \sum_i \alpha_i \cdot f_{\text{ebd}}(x_i)$$

Learn from support set

$$\mathcal{L}^{RR}(W) := \|\Phi_S W - Y_S\|_F^2 + \lambda \|W\|_F^2$$

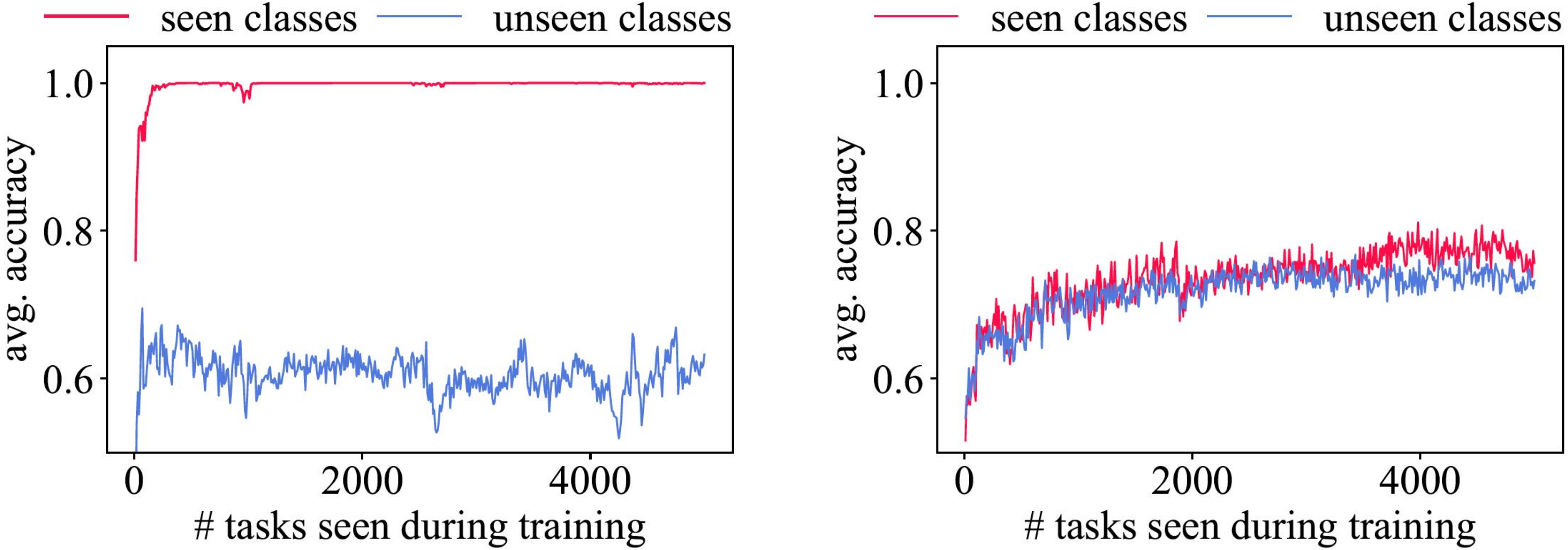
$$W = \Phi_S^T(\Phi_S \Phi_S^T + \lambda I)^{-1} Y_S$$

Predict on query set

$$\hat{Y}_Q = a\Phi_Q W + b$$

Method		20 News		Amazon		HuffPost		RCV1		Reuters		FewRel		Average	
Rep.	Alg.	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
AVG	NN	33.9	45.8	46.7	60.3	31.4	41.5	43.7	60.8	56.5	80.5	47.5	60.6	43.3	58.2
IDF	NN	38.8	51.9	51.4	67.1	31.5	42.3	41.9	58.2	57.8	82.9	46.8	60.6	44.7	60.5
CNN	FT	33.0	47.1	45.7	63.9	32.4	44.1	40.3	62.3	70.9	91.0	54.0	71.1	46.0	63.2
AVG	PROTO	36.2	45.4	37.2	51.9	35.6	41.6	28.4	31.2	59.5	68.1	44.0	46.5	40.1	47.4
IDF	PROTO	37.8	46.5	41.9	59.2	34.8	50.2	32.1	35.6	61.0	72.1	43.0	61.9	41.8	54.2
CNN	PROTO	29.6	35.0	34.0	44.4	33.4	44.2	28.4	29.3	65.2	74.3	49.7	65.1	40.1	48.7
AVG	MAML	33.7	43.9	39.3	47.2	36.1	49.6	39.9	50.6	54.6	62.5	43.8	57.8	41.2	51.9
IDF	MAML	37.2	48.6	43.6	62.4	38.9	53.7	42.5	54.1	61.5	72.0	48.2	65.8	45.3	59.4
CNN	MAML	28.9	36.7	35.3	43.7	34.1	45.8	39.0	51.1	66.6	85.0	51.7	66.9	42.6	54.9
AVG	RR	37.6	57.2	50.2	72.7	36.3	54.8	48.1	72.6	63.4	90.0	53.2	72.2	48.1	69.9
IDF	RR	44.8	64.3	60.2	79.7	37.6	59.5	48.6	72.8	69.1	93.0	55.6	75.3	52.6	74.1
CNN	RR	32.2	44.3	37.3	53.8	37.3	49.9	41.8	59.4	71.4	87.9	56.8	71.8	46.1	61.2
OUR		**52.1**	**68.3**	**62.6**	**81.1**	**43.0**	**63.5**	**54.1**	**75.3**	**81.8**	**96.0**	**67.1**	**83.5**	**60.1**	**78.0**
OUR w/o $t(\cdot)$		50.1	67.5	61.7	80.5	42.0	60.8	51.5	75.1	76.7	93.7	66.9	83.2	58.1	76.8
OUR w/o $s(\cdot)$		41.9	60.7	51.1	75.3	40.1	60.2	48.5	72.8	78.1	94.8	65.8	82.6	54.2	74.4
OUR w/o biLSTM		50.3	66.9	61.9	80.9	42.2	63.0	51.8	74.1	77.2	95.4	66.4	82.9	58.3	77.2
OUR w EBD		39.7	57.5	56.5	76.3	40.6	58.6	48.6	71.5	81.7	95.8	61.5	80.9	54.8	73.4

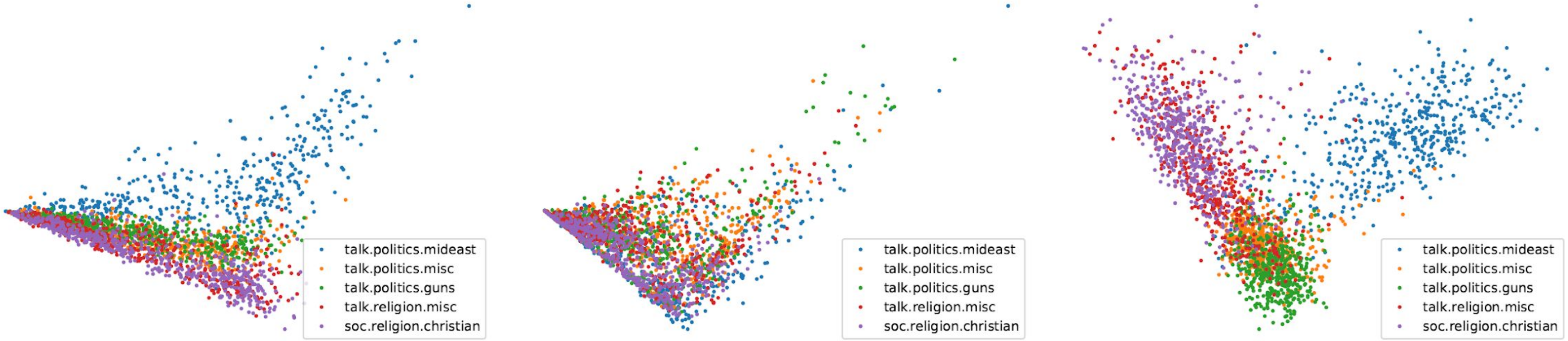
Generalization and Overfitting



(a) CNN+PROTO

(b) OUR

PCA visualization



(a) $s(\cdot)$

(b) $t(\cdot)$

(c) OUR