CLUR: Uncertainty Estimation for Few-Shot Text Classification with Contrastive Learning

Jianfeng He, Xuchao Zhang, Shuo Lei, Abdulaziz Alhamadani, Fanglan Chen, Bei Xiao, Chang-Tien Lu

Virginia Tech

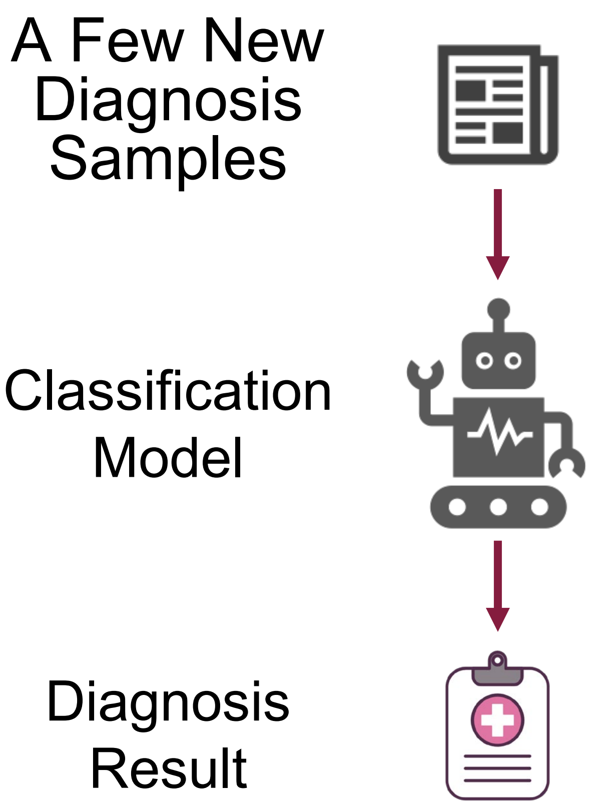
Microsoft

American University

Few-shot text classification is important.

- <u>Few-shot text classification</u> learns a classifier by a few training or even only one training example per class.

- E.g., a new disease with only a few recorded diagnosis at beginning

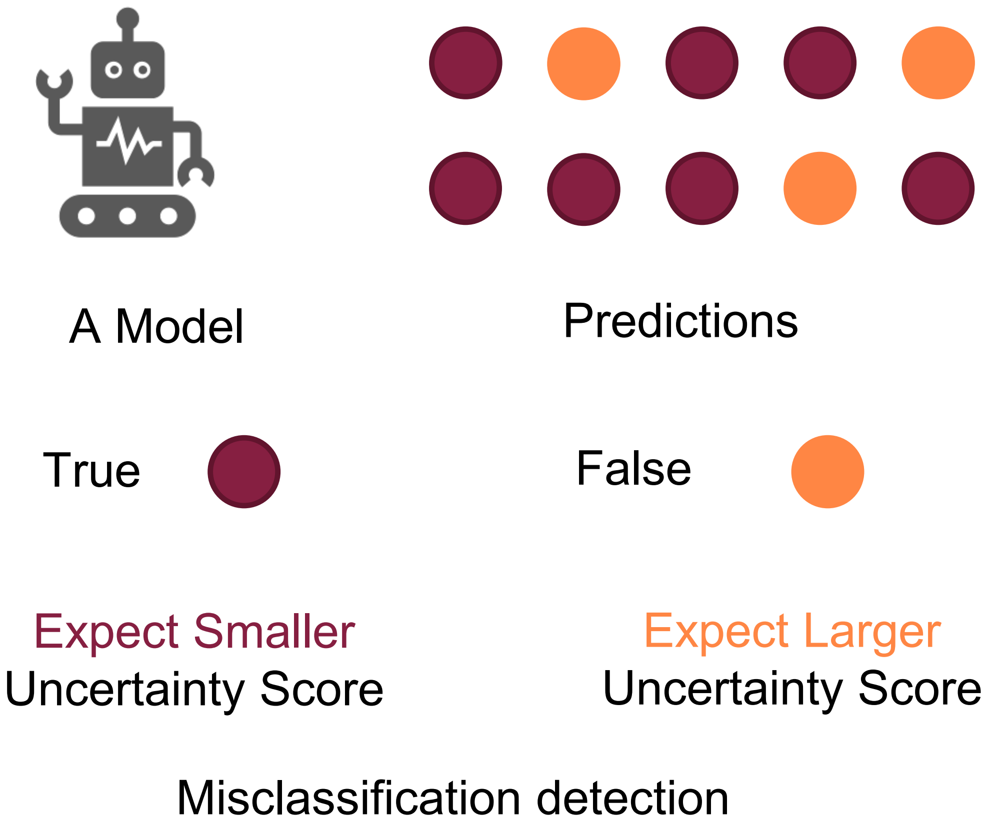


Trust the diagnosis results?
Ask human expert for recheck?

Therefore, we need uncertainty estimation to detect false prediction in few-shot scenerios.

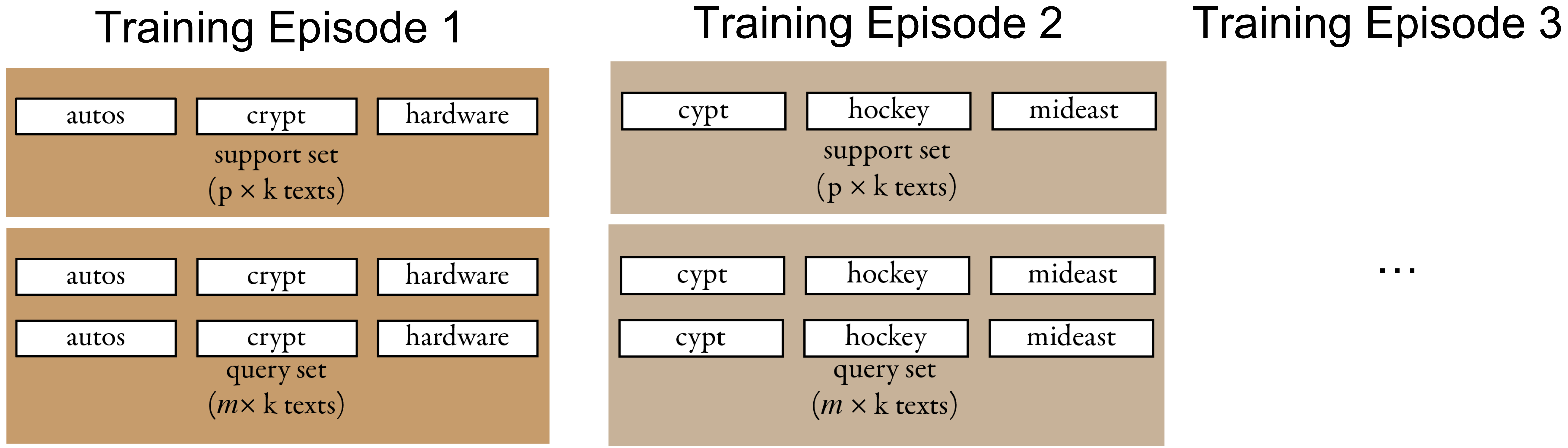
- Uncertainty estimation quantifies to which degree we should discard a model prediction.

- Applications of uncertainty estimation
  - ☐ Out-of-domain detection
  - ☐ Active learning
  - ☐ Misclassification detection (Our focus)



Our Task: Uncertainty Estimation in Few-Shot Text Classification (UEFTC)

Task Setting: Based on meta-learning (meta-training & meta-testing)



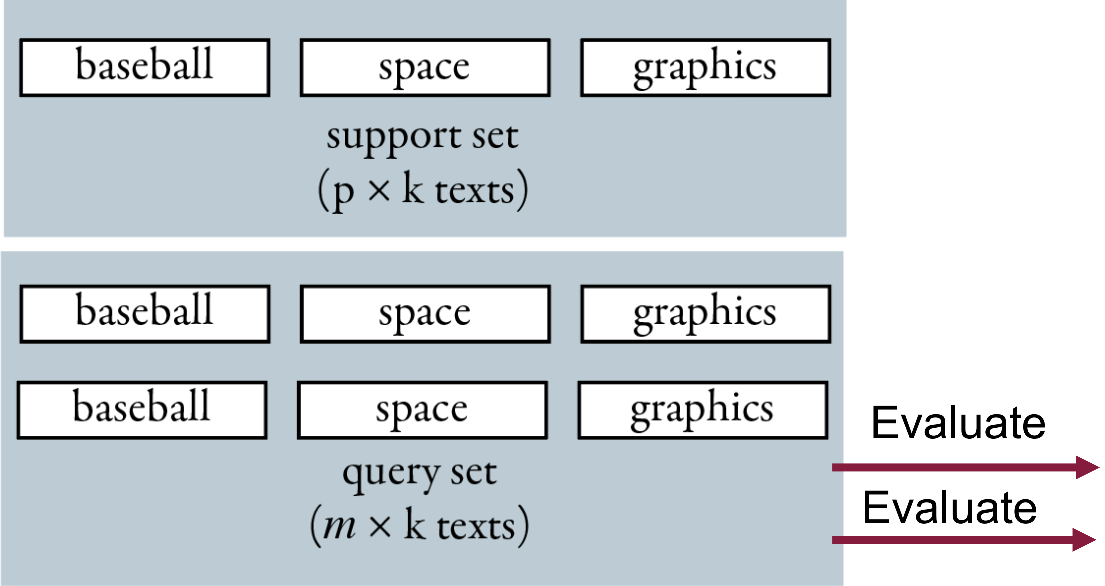
Meta-training: p-shot (sample size) k-way (class size), 1-shot 3-way

Samples in both support and query sets are given labels for minimizing loss.

We aim to improve Uncertainty Estimation in Few-Shot Text Classification (UEFTC).

Task Setting: Based on meta-learning (meta-training & meta-testing)

Testing Episode 1 Testing Episode 2 …



Classification

Uncertainty Estimation

Another sampled episode

Meta-testing Process (use disjoint classes to meta-training)

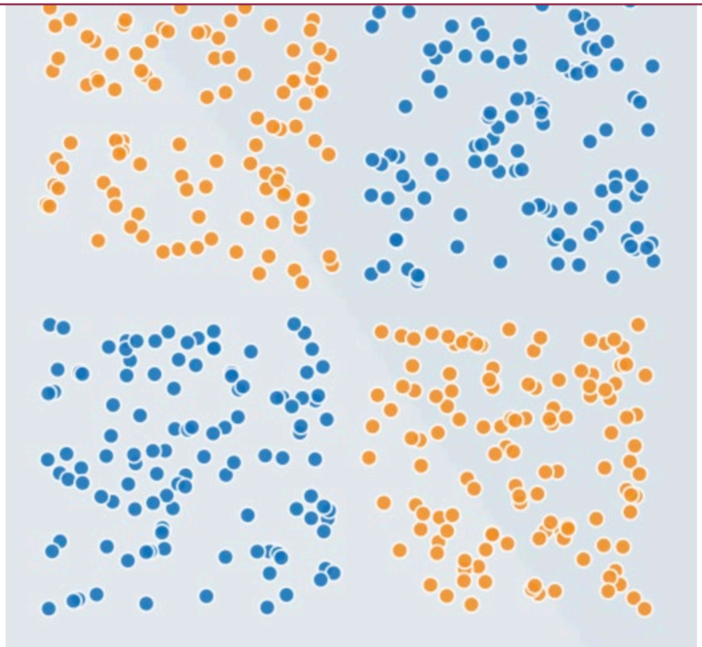
Only in support samples are given labels.

Evaluation: <u>classification</u> & <u>uncertainty estimation</u>

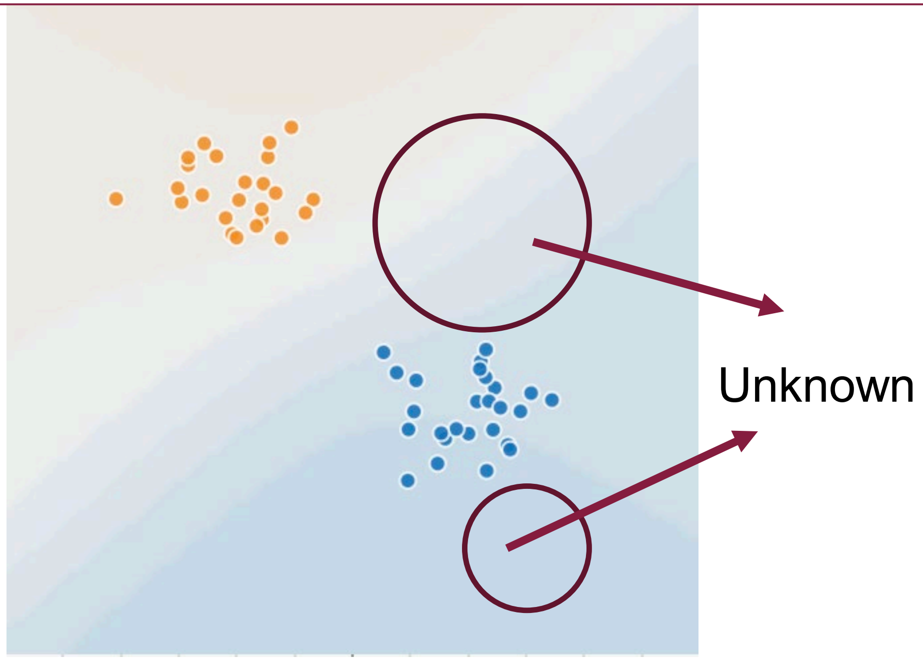
Challenge in UEFTC: Few Support Samples

Sufficient training samples → accurate sample or parameter distribution.

Few support samples → inaccurate sample or parameter distribution. (i.e., 1 support sample per class)



Previous: Uncertainty estimation on traditional text classification



Ours: Uncertainty estimation on few-shot text classification (UEFTC)

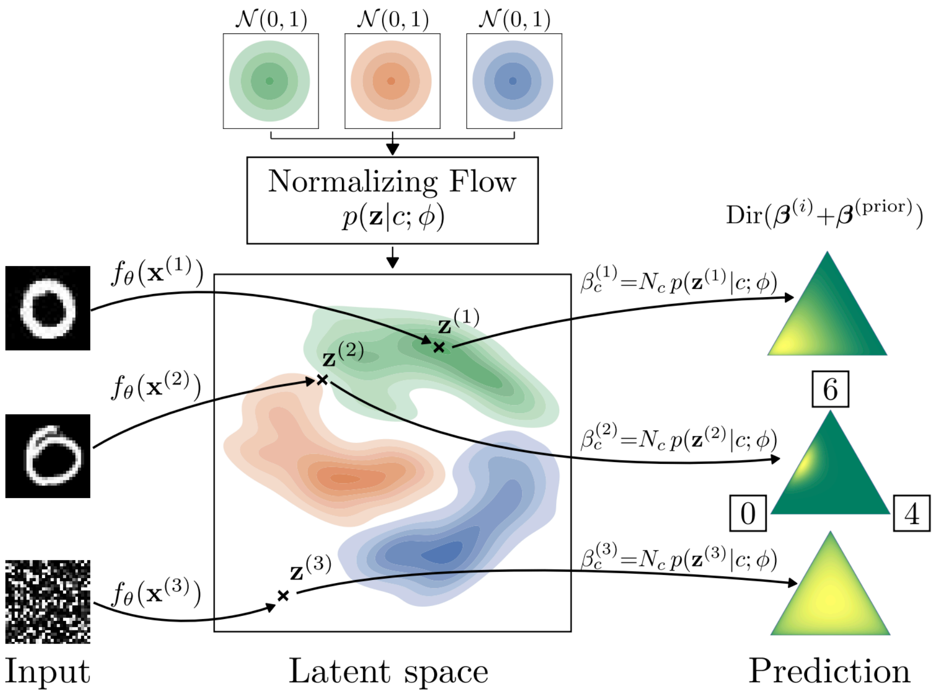
Few-support-sample Impacts on Current Uncertainty Estimation Models in UEFTC

1. Sample-distribution-based methods

➢ probability/distance to distribution of each class of training samples

☐ e.g., Posterior Neural Network

Sample distribution in UEFTC is inaccurate.



Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts. NIPS 2020.

A simple unified framework for detecting out-of-distribution samples and adversarial attacks. NIPS 2018.

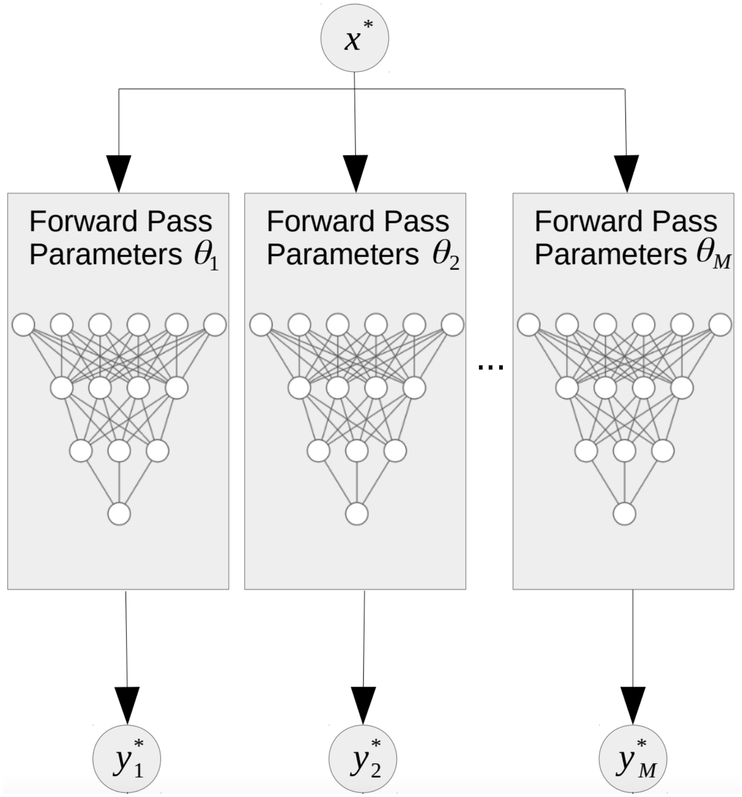
Few-support-sample Impacts on Current Uncertainty Estimation Models in UEFTC

2. Parameter-distribution-based methods

- e.g., Bayesian Neural Network (BNN)

Feasible parameter set has a larger size

<u>Inaccurate</u> parameters distribution



A Survey of Uncertainty in Deep Neural Networks. Arxiv 2021.

Few-support-sample Impacts on Current Uncertainty Estimation Models in UEFTC

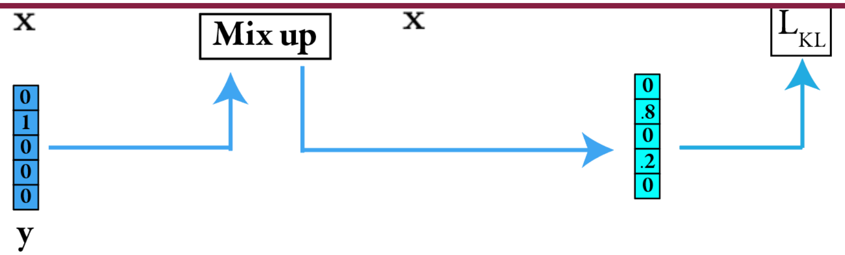
3. Pseudo-label-based methods

- Augment samples
- Manually set their psuedo ground-truth uncertainty score given a specific model structure.
- E.g., Mix-up

Advantage: Independent on sample size

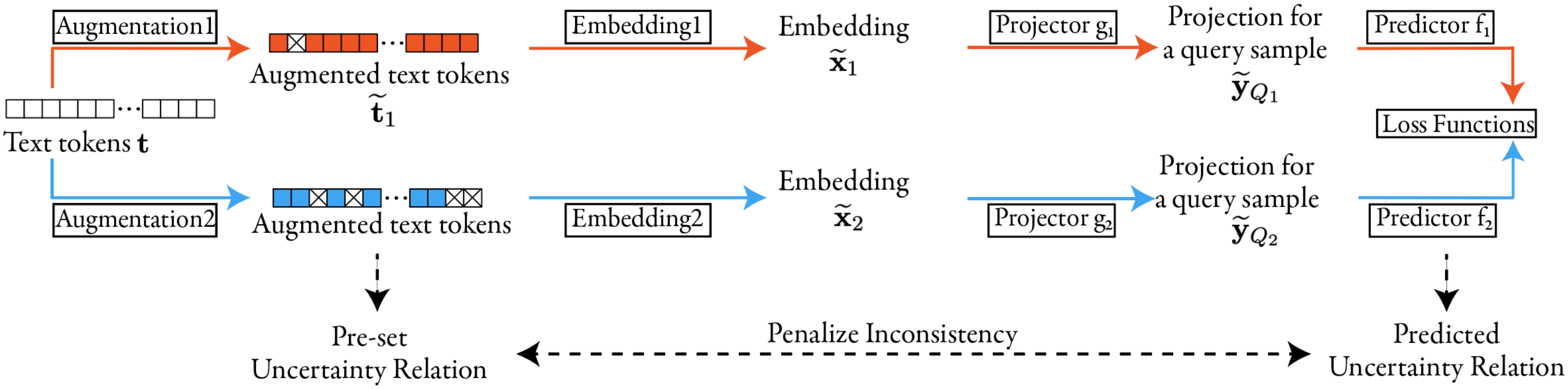
Drawback: Manually set pseudo uncertainty scores (inaccurate).

Thus, we propose a method to self-adaptively learn pseudo ground-truth uncertainty scores.



Towards More Accurate Uncertainty Estimation In Text Classification. EMNLP 2020.

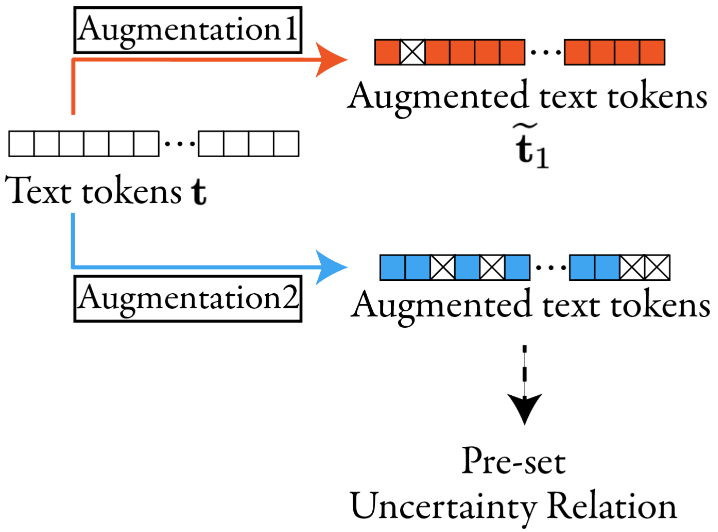
Our Model: Contrastive Learning from Uncertainty Relations (CLUR)



Global View

Main motivation: self-adaptively learn pseudo ground-truth uncertainty scores given a model.

CLUR: Augmentation & Unequal Relation

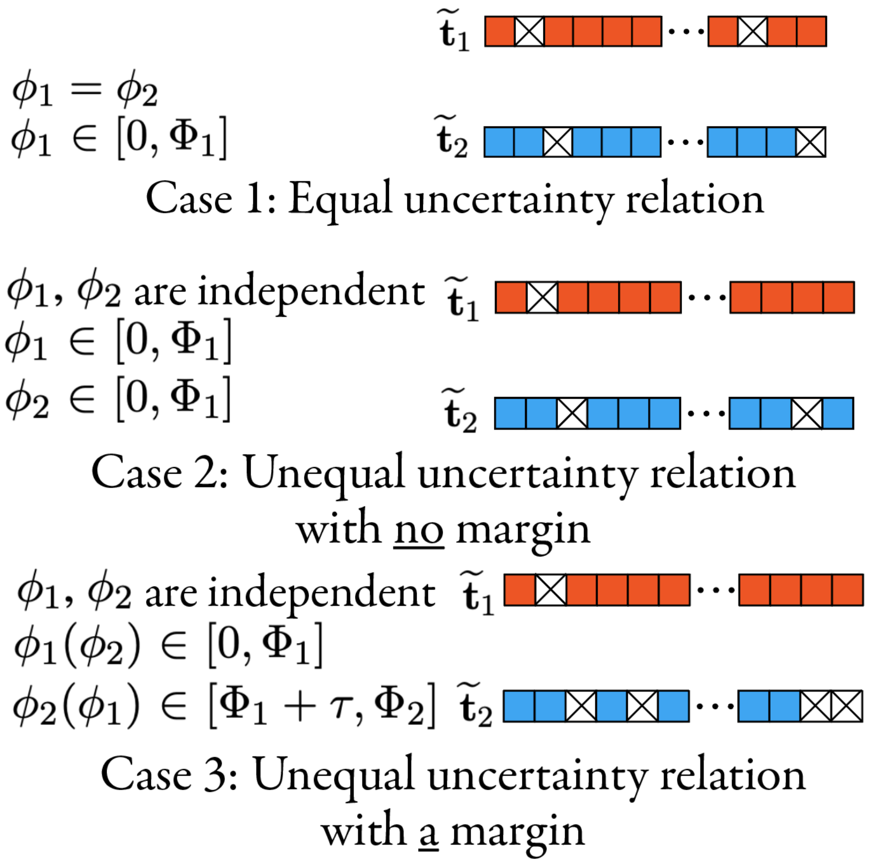


Augmentation: Token-mask

$$\tilde{\mathbf{t}}_1 = \mathbf{t} \cdot \mathbf{m}_{\phi_1}$$

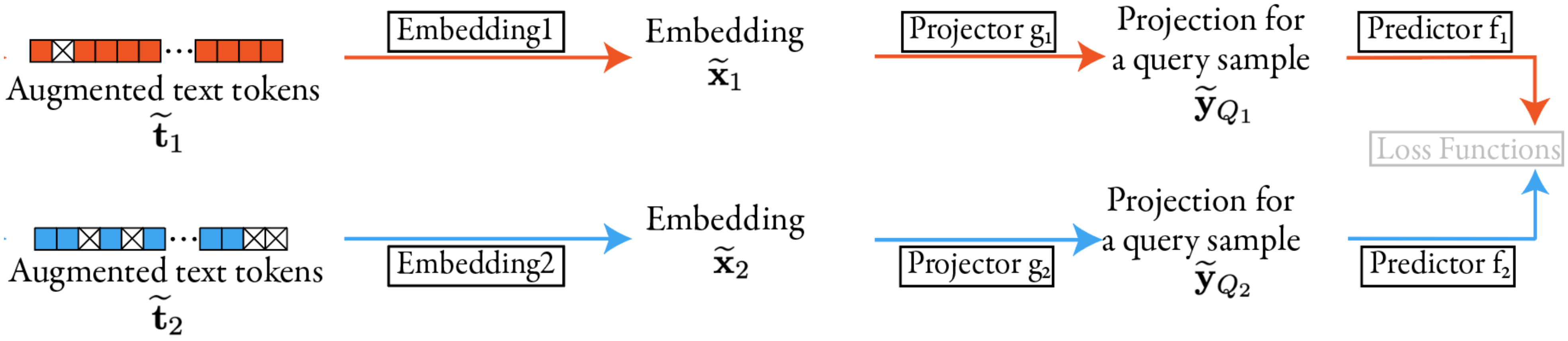
$$\tilde{\mathbf{t}}_2 = \mathbf{t} \cdot \mathbf{m}_{\phi_2}$$

$\mathbf{m}_{\phi_1}$ and $\mathbf{m}_{\phi_2}$: binary vectors to randomly mask by ratios of ϕ_1 and ϕ_2 .



Uncertainty Relations

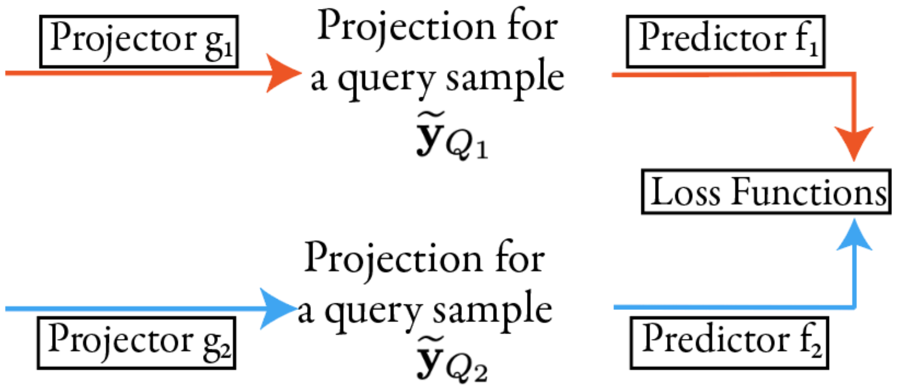
CLUR: General Modules

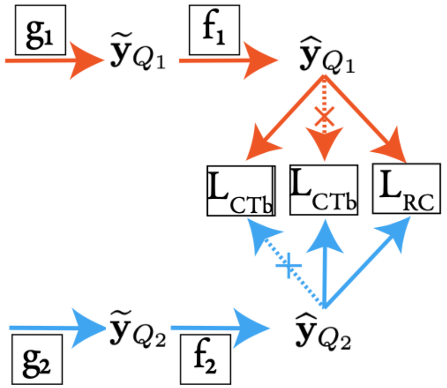


General Module

- Projection & prediction
  - Both k-dimensions (k classes)
  - Follow contrastive learning (SOTA usage of augmented samples)

CLUR: Loss Functions





(b)

- Motivated from SimSiam
  - No negative pairs & large batch size (Few-support-sample limitation)

Contrastive loss <u>equal</u> uncertainty relation (D: Cosine distance; o: detach):

$$L_{CT_b} = D[\hat{\mathbf{y}}_{Q_1}, o(\hat{\mathbf{y}}_{Q_2})] + D[\hat{\mathbf{y}}_{Q_2}, o(\hat{\mathbf{y}}_{Q_1})]$$

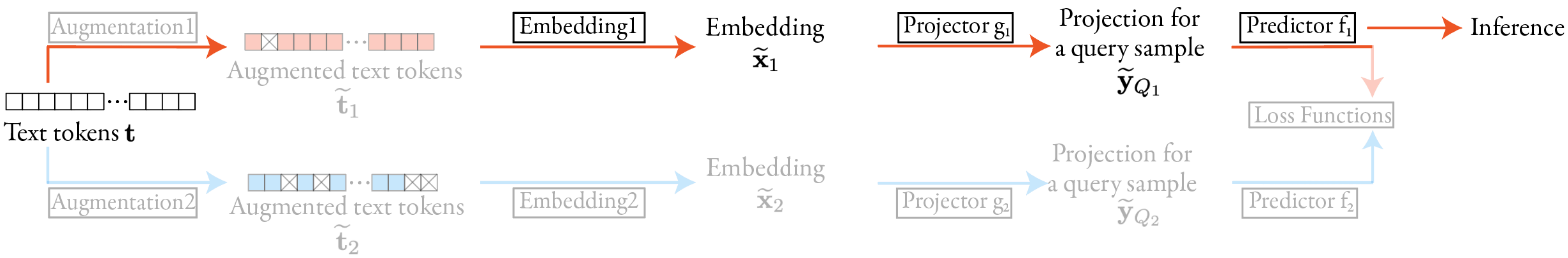
Contrastive loss in <u>unequal</u> uncertainty relation: (H: entropy for uncertainty)

$$L_{CT_b} = max\{\underbrace{[H(\hat{\mathbf{y}}_{Q_1}) - H(o(\hat{\mathbf{y}}_{Q_2}))]}_{\text{Predicted uncertainty relation}} \times \underbrace{(\phi_2 - \phi_1)}_{\text{Pseudo ground-truth uncertainty relation}}, 0\} + max\{[H(\hat{\mathbf{y}}_{Q_2}) - H(o(\hat{\mathbf{y}}_{Q_1}))] \times (\phi_1 - \phi_2), 0\}$$

Total loss: $L_{SUM_b} = L_{RC} + \gamma L_{CT_b}$ $\longrightarrow$ L_{RC} avoids overconfidence (not closing to 1)

Exploring simple siamese representation learning. CVPR 2021.

CLUR: Inference



Only use the first submodel (first row) & skip augmentation

Classification: arg max

Uncertainty Score: reciprocal of maximum probability $\frac{1}{\max(\widehat{\mathbf{y}}_{Q_1})}$

Exploring simple siamese representation learning. CVPR 2021.

Experimental Settings

- Five public datasets
  - News domain: 20News, HuffPost, RCV1
  - User review domain: Amazon Reviews
  - Medical domain: Med-Domain
- Metrics
  - AUROC
  - AUPR
  - F1 scores in eliminated ratios
    - Simulate human recheck
    - Replace the most uncertain parts by the ground truth
- Our CLUR and baselines are all default based on a classical few-shot model, FTC-DS.

Few-shot Text Classification with Distributional Signatures. ICLR 2020.

Our CLUR model performs better than baselines in UEFTC on 5-way 1-shot setting.

Methods	Uncertainty Ratio (F1 Score, Eliminated Ratio)↑					AUROC ↑	AUPR↑
	0%	10%	20%	30%	40%		
20News in the 5-way 1-shot setting							
FTC-DS	47.56±1.56	55.76±1.38	62.92±1.25	69.86±1.11	75.77±1.04	68.17±2.15	68.20±1.29
DE	52.32±1.70	59.45±1.59	65.71±1.47	72.12±1.32	77.57±1.27	67.69±2.44	69.38±1.57
DE+Metric	52.33±1.61	59.63±1.44	65.73±1.36	72.04±1.26	77.61±1.15	68.02±2.38	69.44±1.45
MSD1	53.11±1.60	60.47±1.47	66.61±1.36	72.87±1.26	78.38±1.09	68.40±2.35	70.01±1.36
MSD2	52.54±1.32	60.09±1.19	66.54±1.10	72.59±1.04	77.96±0.93	68.49±1.91	69.78±1.01
SimSiam(CLUR-a-1)	53.30±1.57	60.63±1.43	66.86±1.32	73.19±1.23	78.59±1.16	68.74±2.29	70.89±1.36
CLUR-b-3	**54.53**±1.50	**62.06**±1.37	**68.29**±1.25	**74.59**±1.11	**80.02**±0.98	**70.50**±2.13	**73.71**±1.22
RCV1 in the 5-way 1-shot setting							
FTC-DS	51.32±1.64	59.71±1.49	66.16±1.33	72.83±1.23	78.65±1.12	70.48±2.32	73.99±1.22
DE	55.42±1.62	62.96±1.50	68.91±1.37	74.99±1.22	80.09±1.14	70.72±2.34	75.12±1.12
DE+Metric	54.89±1.68	62.50±1.52	68.41±1.34	74.59±1.25	79.78±1.20	70.61±2.46	74.51±1.24
MSD1	54.91±1.79	62.32±1.64	68.27±1.48	74.60±1.36	79.82±1.26	70.11±2.50	73.67±1.35
MSD2	55.54±1.65	62.96±1.50	68.91±1.39	75.18±1.30	80.39±1.17	71.12±2.37	75.34±1.23
SimSiam(CLUR-a-1)	54.12±1.97	61.66±1.79	67.98±1.67	74.47±1.49	79.71±1.38	71.10±2.73	74.24±1.56
CLUR-b-3	**55.89**±1.60	**63.48**±1.44	**69.47**±1.35	**75.62**±1.23	**80.91**±1.12	**72.31**±2.26	**77.00**±1.10

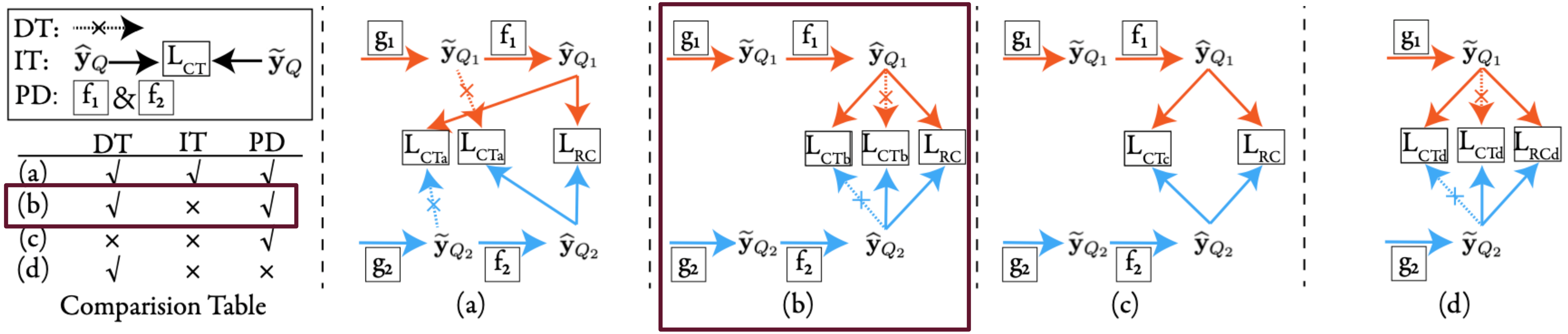
UEFTC results on 5-way 1-shot on 20News & RCV1

Our CLUR model performs better than baselines in UEFTC 5-way 5-shot setting.

Methods	Uncertainty Ratio (F1 Score, Eliminated Ratio)↑					AUROC ↑	AUPR↑
	0%	10%	20%	30%	40%		
20News in the 5-way 1-shot setting							
FTC-DS	47.56±1.56	55.76±1.38	62.92±1.25	69.86±1.11	75.77±1.04	68.17±2.15	68.20±1.29
DE	52.32±1.70	59.45±1.59	65.71±1.47	72.12±1.32	77.57±1.27	67.69±2.44	69.38±1.57
DE+Metric	52.33±1.61	59.63±1.44	65.73±1.36	72.04±1.26	77.61±1.15	68.02±2.38	69.44±1.45
MSD1	53.11±1.60	60.47±1.47	66.61±1.36	72.87±1.26	78.38±1.09	68.40±2.35	70.01±1.36
MSD2	52.54±1.32	60.09±1.19	66.54±1.10	72.59±1.04	77.96±0.93	68.49±1.91	69.78±1.01
SimSiam(CLUR-a-1)	53.30±1.57	60.63±1.43	66.86±1.32	73.19±1.23	78.59±1.16	68.74±2.29	70.89±1.36
CLUR-b-3	**54.53**±1.50	**62.06**±1.37	**68.29**±1.25	**74.59**±1.11	**80.02**±0.98	**70.50**±2.13	**73.71**±1.22
RCV1 in the 5-way 1-shot setting							
FTC-DS	51.32±1.64	59.71±1.49	66.16±1.33	72.83±1.23	78.65±1.12	70.48±2.32	73.99±1.22
DE	55.42±1.62	62.96±1.50	68.91±1.37	74.99±1.22	80.09±1.14	70.72±2.34	75.12±1.12
DE+Metric	54.89±1.68	62.50±1.52	68.41±1.34	74.59±1.25	79.78±1.20	70.61±2.46	74.51±1.24
MSD1	54.91±1.79	62.32±1.64	68.27±1.48	74.60±1.36	79.82±1.26	70.11±2.50	73.67±1.35
MSD2	55.54±1.65	62.96±1.50	68.91±1.39	75.18±1.30	80.39±1.17	71.12±2.37	75.34±1.23
SimSiam(CLUR-a-1)	54.12±1.97	61.66±1.79	67.98±1.67	74.47±1.49	79.71±1.38	71.10±2.73	74.24±1.56
CLUR-b-3	**55.89**±1.60	**63.48**±1.44	**69.47**±1.35	**75.62**±1.23	**80.91**±1.12	**72.31**±2.26	**77.00**±1.10

UEFTC results on 5-way 5-shot on 20News & RCV1

Designed Loss for Ablation Studies of Contrastive Learning Modules



Summary and comparisons between our designed four loss functions (DT: Detach operation, IT: Intersection comparison, PD: Predictor).

- ➢ Designed another three losses
  - ◻ Main one: choice (b)

Ablation studies of CLUR

Methods	Detach	Intersection	Predictor	Uncertainty Ratio (F1 Score, Eliminated Ratio) ↑					AUROC ↑	AUPR ↑
				0%	10%	20%	30%	40%		
Amazon in the 5-way 5-shot setting										
CLUR-b-3	✓	×	✓	**81.95**±1.09	**87.37**±0.90	**91.49**±0.76	**94.47**±0.57	**96.21**±0.51	**82.35**±1.79	**95.16**±0.36
CLUR-c-3	×	×	✓	81.44±1.09	86.91±0.94	90.59±0.77	93.63±0.70	95.76±0.61	81.26±1.92	94.52±0.43
CLUR-d-3	✓	×	×	80.17±2.09	85.90±1.76	89.93±1.48	93.33±1.23	95.58±1.02	81.13±3.05	94.33±0.92
CLUR-a-2	✓	✓	✓	80.83±1.29	86.32±1.12	90.14±0.96	93.33±0.82	95.50±0.71	80.69±2.15	94.23±0.55
CLUR-b-2	✓	×	✓	80.59±1.23	86.11±1.06	90.00±0.91	93.25±0.80	95.42±0.70	80.79±2.07	94.17±0.52
CLUR-c-2	×	×	✓	80.90±1.19	86.31±1.01	90.05±0.84	93.08±0.75	95.20±0.66	80.11±2.05	93.91±0.48

UEFTC results on 5-way 5-shot on Amazon dataset

CLUR with loss choice (b) using unequal uncertainty relation with a margin (case 3) performs the best.

Besides, the p-values of our t-test indicate that module contribution is Predictor > Detach > Intersection

Generalization of CLUR

We test CLUR on another classical few-shot model, Prototypical Network, and it is still effective.

Methods	Uncertainty Ratio (F1 Score, Eliminated Ratio)↑					AUROC ↑	AUPR↑
	0%	10%	20%	30%	40%		
FTC-DS	27.12±3.58	35.75±3.43	43.48±3.29	51.64±3.09	58.96±2.95	55.75±5.96	37.11±6.91
DE	29.83±3.52	38.09±3.36	45.55±3.18	53.51±3.00	60.72±2.88	58.81±5.70	41.14±6.81
DE+Metric	31.09±3.04	39.22±2.89	46.54±2.76	54.35±2.56	61.35±2.41	58.76±4.74	42.17±5.05
MSD1	30.96±2.84	39.06±2.68	46.36±2.58	54.08±2.48	61.04±2.38	57.75±4.76	40.13±4.33
MSD2	30.36±3.53	38.44±3.34	45.71±3.17	53.53±2.99	60.60±2.77	57.72±5.26	40.54±5.85
SimSiam(CLUR-a-1)	30.39±3.42	38.52±3.28	45.81±3.14	53.55±2.97	60.66±2.76	57.58±5.32	40.62±5.90
CLUR-b-3	**31.77**±3.32	**40.16**±3.09	**47.54**±2.92	**55.37**±2.73	**62.47**±2.56	**59.20**±5.18	**43.89**±5.75

UEFTC results on 5-way 1-shot on 20News based on Prototypical Network.

Prototypical Networks for Few-shot Learning. NIPS 2017.

Conclusion

- We define and provide a benchmark for Uncertainty Estimation on Few-shot Text Classification (UEFTC).
- For few-support-sample challenge in UEFTC, we propose Contrastive Learning with Unequal Relation (CLUR) to self-adaptively learn the pseudo ground-truth uncertainty scores given a specific model structure.
- Propose unequal uncertainty relation (>, <), which is ignored by the contrastive learning using only equal relation (=, ≠).
- The data split and code is coming soon, where the link has been attached in the paper.

Thanks! Q & A





Jianfeng He
Virginia Tech

Xuchao Zhang
Microsoft

Shuo Lei
Virginia Tech

Abdulaziz Alhamadani
Virginia Tech

Fanglan Chen
Virginia Tech


Bei Xiao
American University

Chang-Tien Lu
Virginia Tech

National Science Foundation (NSF) grants CNS-2141095 and 2050727.

Appendix: More Related Work

- Uncertainty estimation on text classification
  - Training process

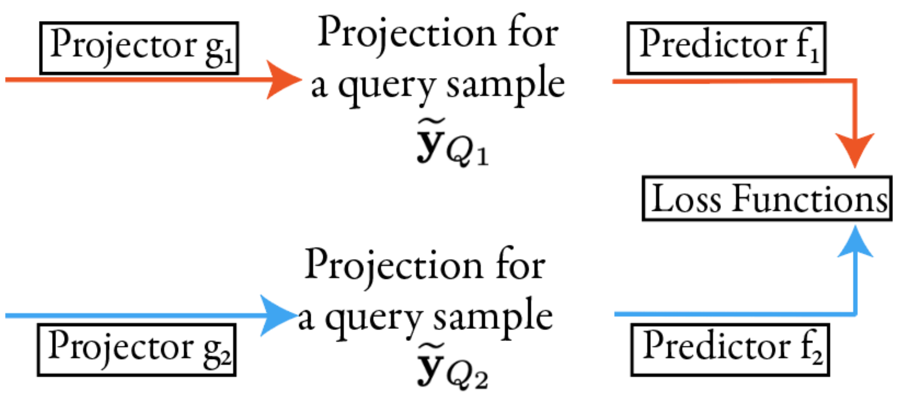
    e.g., active learning
  - Testing process

    e.g., out-of-distribution detection, <u>misclassification</u>
- Few-shot text classification
  - <u>Meta-learning based</u>
  - Transfer-learning based
- Contrastive learning
  - Equal relation

    e.g., same(=)/different($\neq$) instance, same(=)/different($\neq$) class
  - <u>Unequal relation</u> (Our proposed)

    e.g., larger ($>$) /smaller($<$) uncertainty to be classified

Appendix: Revised Cross-Entropy loss



Contrastive loss in <u>unequal</u> uncertainty relation: (H: entropy for uncertainty)

$$L_{CT_b} = max\{[\underbrace{H(\hat{\mathbf{y}}_{Q_1}) - H(o(\hat{\mathbf{y}}_{Q_2}))}_{\text{Predicted uncertainty relation}}] \times \underbrace{(\phi_2 - \phi_1)}_{\text{Pseudo ground-truth uncertainty relation}}, 0\} + max\{[H(\hat{\mathbf{y}}_{Q_2}) - H(o(\hat{\mathbf{y}}_{Q_1}))] \times (\phi_1 - \phi_2), 0\}$$

- ☐ Revised cross-entropy loss: probability of correct class is within $[\beta\,, 1)$, instead of closing 1

$$L_{RC} = \max\{L_{CE}(\hat{\mathbf{y}}_{Q_1}, \mathbf{y}_Q) + \log(\beta), 0\} + \max\{L_{CE}(\hat{\mathbf{y}}_{Q_2}, \mathbf{y}_Q) + \log(\beta), 0\}$$

Total loss: $L_{SUM_b} = L_{RC} + \gamma L_{CT_b}$

Exploring simple siamese representation learning. CVPR 2021.

Appendix: Experiments on Medical Domain

We also test CLUR on a medical domain dataset, and it is still effective.

Methods	Uncertainty Ratio (F1 Score, Eliminated Ratio)↑					AUROC ↑	AUPR↑
	0%	10%	20%	30%	40%		
FTC-DS	50.63±1.79	58.98±1.55	65.63±1.40	71.69±1.28	77.08±1.23	67.42±2.37	70.24±1.66
DE	56.01±1.83	63.13±1.67	69.36±1.53	75.17±1.44	80.36±1.32	70.94±2.54	75.53±1.43
DE+Metric	54.98±2.12	62.06±1.96	68.32±1.85	74.31±1.71	79.80±1.55	71.01±2.89	75.62±1.79
MSD1	55.93±1.99	62.88±1.82	69.04±1.70	74.85±1.60	80.02±1.44	70.10±2.71	74.39±1.65
MSD2	55.99±1.50	62.96±1.39	69.04±1.32	74.78±1.21	79.94±1.08	70.15±2.10	75.82±1.08
SimSiam(CLUR-a-1)	54.48±1.69	61.49±1.62	67.78±1.51	73.89±1.39	79.43±1.32	70.64±2.36	74.31±1.49
CLUR-b-3	**56.81**±1.69	**63.87**±1.51	**70.16**±1.42	**76.10**±1.32	**81.44**±1.21	**72.31**±2.36	**77.29**±1.31

UEFTC results on 5-way 1-shot on the Med-Domain dataset.

Prototypical Networks for Few-shot Learning. NIPS 2017.