模式识别与神经计算研究组

PAttern Recognition and NEural Computing

TrustAL: Trustworthy Active Learning using Knowledge Distillation

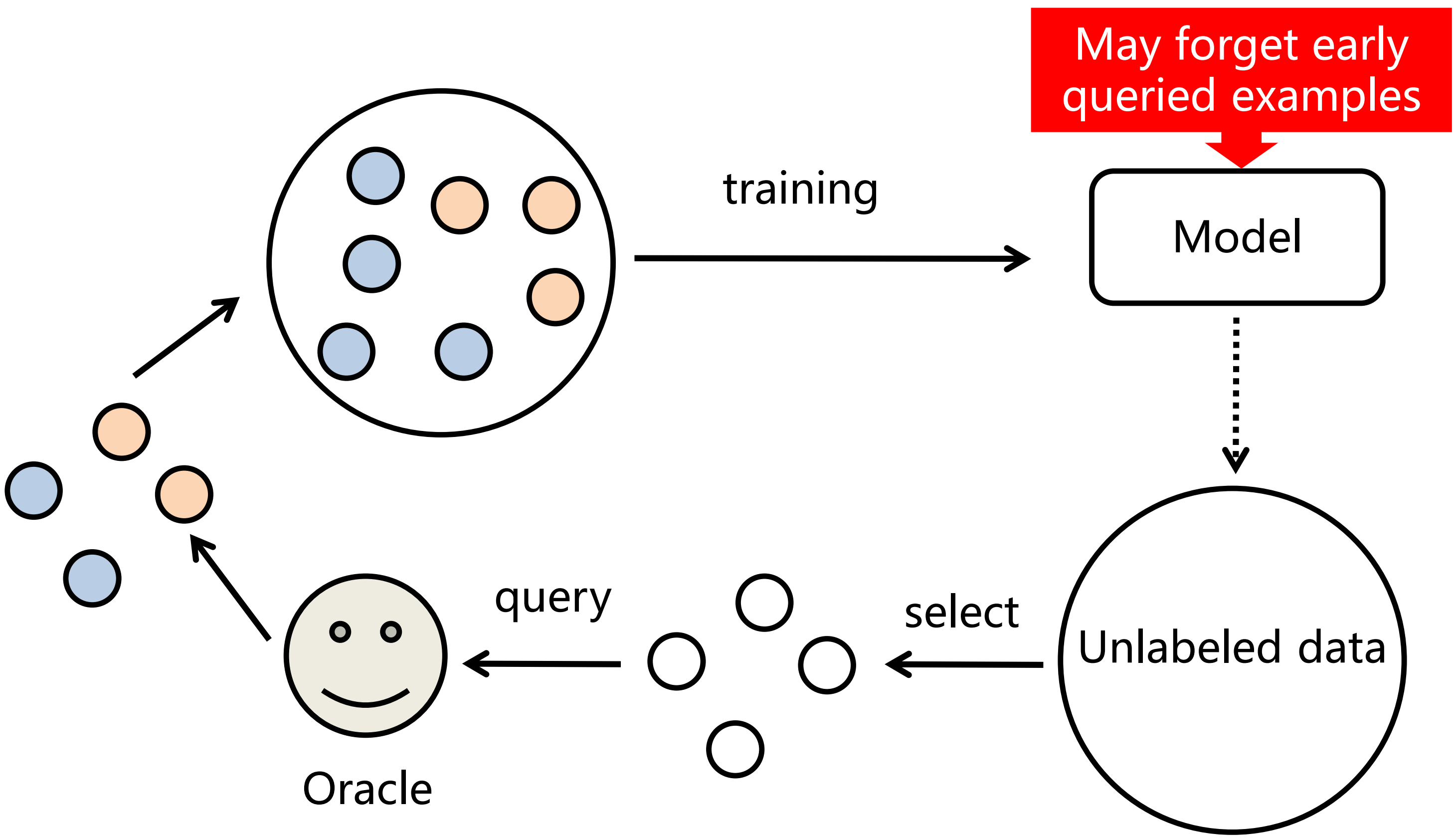
Beong-woo Kwak[1], Youngwook Kim[2], Yu Jin Kim[1], Seung-won Hwang[3], Jinyoung Yeo[1]*

[1] Department of Artificial Intelligence, Yonsei University
[2] Department of Computer Science, Yonsei University
[3] Department of Computer Science and Engineering, Seoul National University

{beongwoo.kwak, youngwook, yujin000731}@yonsei.ac.kr
seungwonh@snu.ac.kr
jinyeo@yonsei.ac.kr

AAAI 2022

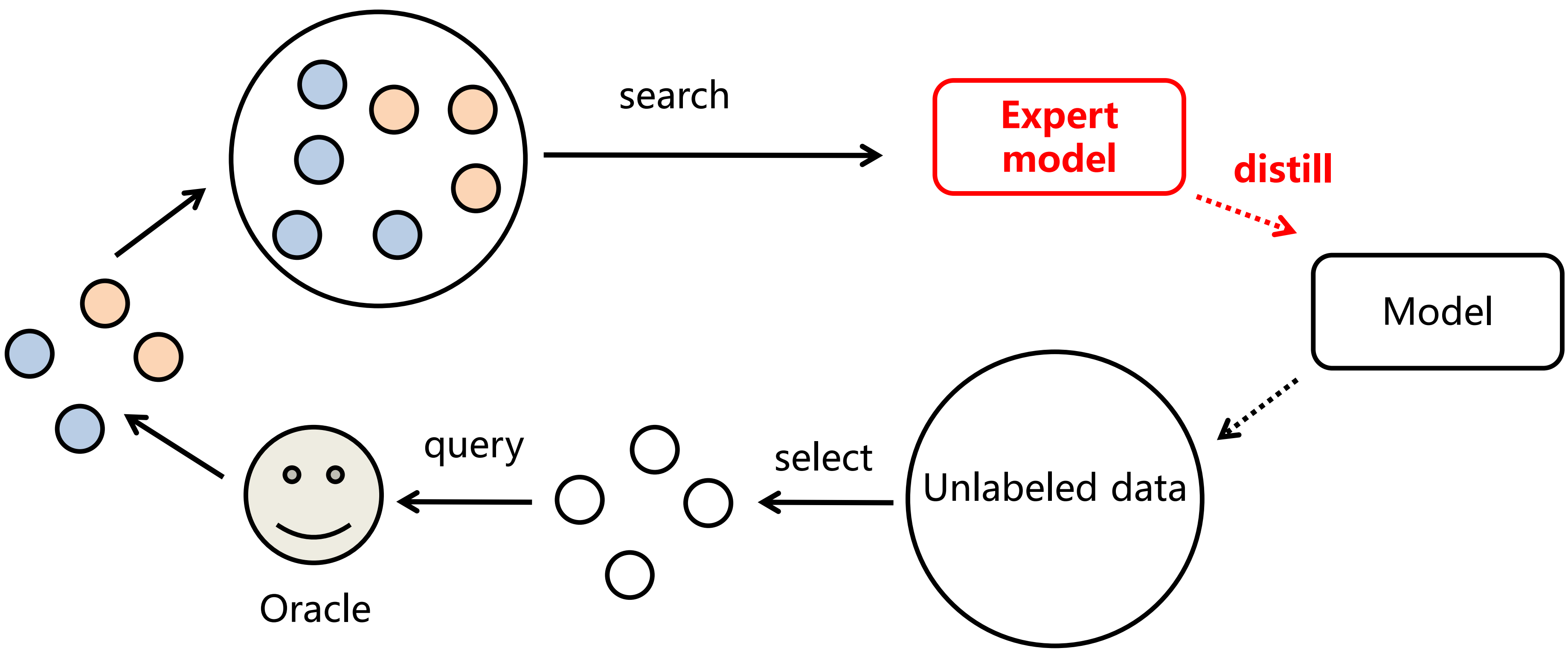
Active Learning



Active learning

Main idea

- At each iteration, **search an expert model** for the forgotten knowledge and **distill to the current model**



Model Learning

$$\theta_t = argmin_{\theta_t} L_{CE}(\theta_t) + \alpha \cdot \underline{L_{KL}(\boxed{\theta_{t-\Delta t}}, \theta_t)}$$

Knowledge distillation loss

$$\boxed{i.e., \sum_{(x_i, y_i) \in \mathcal{L}} KL\text{-}Divergence(f(x_i; \theta_{t-\Delta t}), f(x_i; \theta_t)).}$$

Key Problem: How to select the expert model for knowledge distillation?

1. Monotonic Consistency (TrustAL-MC)

   Use the last round model, i.e., $\theta_{t-\Delta t} = \theta_{t-1} = M_t.$

2. Non-monotonic Consistency (TrustAL-NC)

   Find a model which has the knowledge of forgettable examples for the current model.

TrustAL-NC

- Given a development dataset $\mathcal{D}_{dev}$ with m examples, calculate the forgotten event for each example i in $\mathcal{D}_{dev}$

Definition 2 *(Correct Inconsistency) The degree of correct inconsistency of θ_t for sample x_i is measured as the number of occurrences of forgetting events for sample x_i from any predecessor model $\theta_{t-\Delta t}$, where $0 < \Delta t \leq t$:*

$$\mathbb{C}_i^{(t)} = \sum_{\Delta t=1}^{t} \mathbb{1}_{(acc_i^{t-\Delta t} > acc_i^t)}$$

$$acc_i^t = \mathbb{1}_{\hat{y}_i^t = y_i}$$

Higher value means easily forgettable

- Select the expert model based on the following weighted accuracy on $\mathcal{D}_{dev}$.

$$g(\theta_{t-\Delta t}, M_t) = \tilde{\mathbb{C}}^{t\top} \langle acc_1^{t-\Delta t}, \ldots, acc_m^{t-\Delta t} \rangle / m$$

Higher value means the expert model θt−Δt tends to have the knowledge of forgettable examples for the current model

✓ Datasets

- TREC (Roth et al. 2002).
- Movie review (Pang and Lee 2005).
- SST-2 (Socher et al. 2013).

✓ Baselines

- **CONF** (Wang and Shang 2014): An uncertainty-based method that selects samples with least confidence..
- **CORESET** (Sener and Savarese 2018): A diversitybased method that selects coreset of remaining samples.
- **BADGE** (Ash et al. 2019): A hybrid method that selects samples considering both uncertainty and diversity.

Results

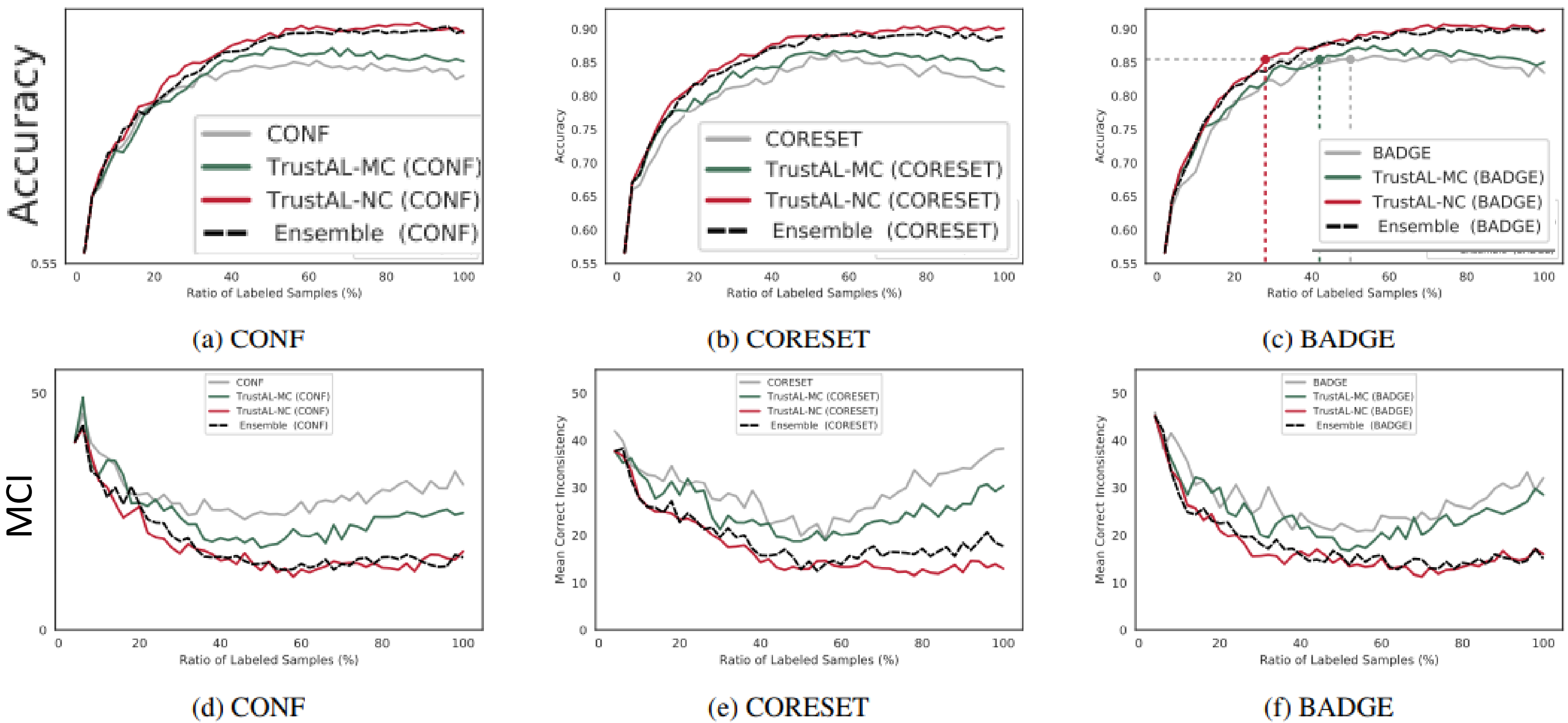
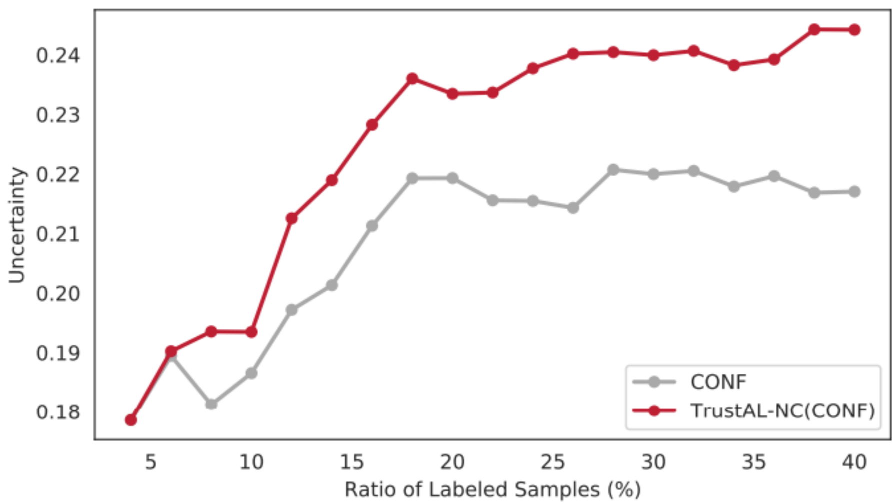


Figure 3: Accuracy (a-c) and MCI (d-f) versus the ratio of labeled samples

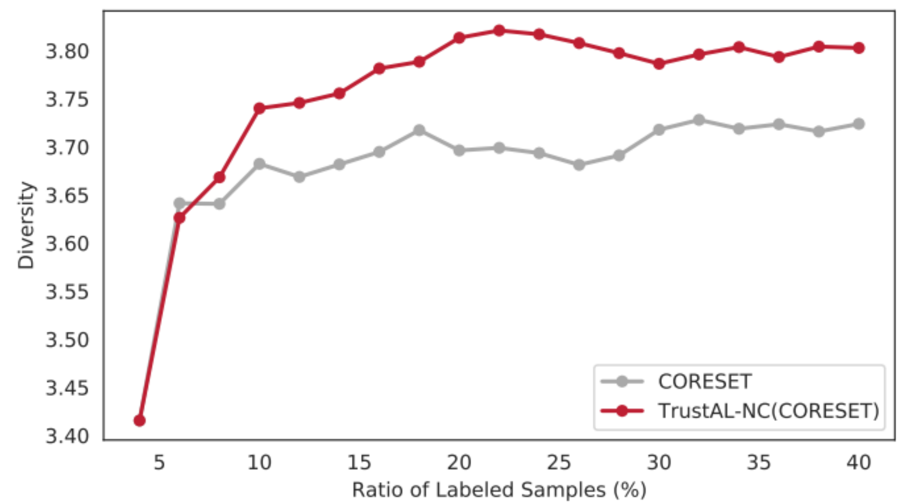
$$\mathrm{MCI} = \sum_i \mathbb{C}_i^{(t)} / t. \qquad \mathbb{C}_i^{(t)} = \sum_{\Delta t=1}^{t} \mathbb{1}_{(acc_i^{t-\Delta t} > acc_i^{t})}$$

Results

- How does TrustAL help data acquisition?



(a) CONF



(b) CORESET

Figure 5: Data acquisition analysis in stable phase on TREC; x-axis represents the ratio of labeled samples and y-axis represents the corresponding metrics.

- better model training leads to better acquisition, strengthening models ability to identify more informative samples

Conclusion

- Traditional AL framework may suffer from knowledge forgetting.
- Introducing the knowledge distillation technique can mitigate this problem by properly selecting the expert model.
- Better model learning scheme also strengthen the subsequent query quality.

THANKS