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

Abstract

Active learning can be defined as iterations of data labeling, model training, and data acquisition, until sufficient labels are acquired. A traditional view of data acquisition is that, through iterations, knowledge from human labels and models is implicitly distilled to monotonically increase the accuracy and label consistency. Under this assumption, the most recently trained model is a good surrogate for the current labeled data, from which data acquisition is requested based on uncertainty/diversity. Our contribution is debunking this myth and proposing a new objective for distillation. First, we found example forgetting, which indicates the loss of knowledge learned across iterations. Second, for this reason, the last model is no longer the best teacher– For mitigating such forgotten knowledge, we select one of its predecessor models as a teacher, by our proposed notion of "consistency". We show that this novel distillation is distinctive in the following three aspects; First, consistency ensures to avoid forgetting labels. Second, consistency improves both uncertainty/diversity of labeled data. Lastly, consistency redeems defective labels produced by human annotators.

Introduction

Labeling data is a fundamental bottleneck in machine learning due to annotation cost and time. One practical solution is Active Learning (AL): given a limited labeling budget k, which example should I ask human annotators to label? Generally, this can be done through an *iterative* process of labeling data, model training, and data acquisition steps, until sufficient labels are obtained. At each iteration, based on the last trained model, unlabeled yet the k most desirable examples are recognized and added to the labeled dataset to train a new model. This process continues to the next iteration for selecting next k unlabeled examples based on the newly trained model. That is, a naive belief in AL is that the last trained model can be a good reference or surrogate for the distribution of the currently labeled data, which indicates what examples are desired for the next model update.

In this work, our empirical observation dispels this myth. Although the model knowledge learned from the labels is expected to be "consistently" kept or improved across AL iterations, we find that knowledge learned at some time is suddenly forgotten, which indicates that the recent model is ineligible to be treated as a good reference of the labeled dataset. More specifically, we can observe such inconsistent behaviors of the trained model during inference time, where sample i predicted correctly at iteration t is predicted incorrectly at iteration $t + \Delta t$, which is called *example forgetting* (Toneva et al. 2019).

Motivated by this, in this work, we argue that **correct-consistency** (which we call consistency for brevity) should be an essential criterion, which is the model ability to make consistent correct predictions across successive AL generations for the same input (Wang et al. 2020). In the view of consistency, prior AL methods only focusing on data acquisition steps (Dasgupta 2011; Xu et al. 2003; Bodó, Minier, and Csató 2011; Ash et al. 2019) are still sub-optimal since the three transitions among AL steps may suffer from following problems due to inconsistency (reverse phenomenon of consistency), which we empirically analyze later:

- **From labeling to model training:** Despite successful data acquisition, the subsequent labeling efforts can be negated by forgetting the learned knowledge later, which wastes annotation cost. We argue that consistency is key to make a label-efficient AL (Figure 3).
- **From model training to data acquisition:** Inconsistent data acquisition models cannot serve as a good reference for the current data distribution, which leads to contaminating the next data acquisition step. Improving consistency may be synergetic to either uncertainty- or diversity-based acquisition strategies (Figure 5).
- **From data acquisition to labeling:** Human annotators who act as oracles are usually subject to accidental mislabeling (Bouguelia et al. 2018) which degrades traditional AL methods. Learning to keep consistency enables to mitigate the confusion from the noisy labels (Figure 6).

To overcome these drawbacks and thus make all the three transitions in AL more **trustworthy**, we propose a label-efficient AL framework, called **Trust**worthy **AL** (**TrustAL**), for bridging the knowledge discrepancy between labeled data and model. In TrustAL, our key idea is to add a new step in the iterative process of AL to learn the forgotten knowledge, which is orthogonally applicable to state-of-the-art data acquisition strategies in a synergistic manner. Specifi-

*Corresponding author

cally, at each iteration, TrustAL first searches for an expert model for the forgotten knowledge among the predecessor models. Then, TrustAL transfers the model knowledge (*e.g.*, logits) to the current model training step by leveraging the knowledge distillation technique (Hinton, Vinyals, and Dean 2015). By optimizing the dual goals of following both human labels and machine labels of the expert predecessor, the newly trained model can relieve forgotten knowledge and thus be more consistent, keeping its correct predictions.

For the purpose of identifying which predecessor is the most desired teacher to relieve the forgotten knowledge, we further explore the teacher selection problem. To resolve this, we present two teacher selection strategies, (1) **TrustAL-MC**: monotonic choice of the most recent model (*i.e.*, a proxy of the most accurate model), and (2) **TrustAL-NC**: non-monotonic choice of the well-balanced model with accuracy and consistency, which we thoroughly design as analysis/evaluation measures in this paper.

Our experiments show that the TrustAL framework significantly improves performance with various data acquisition strategies while preserving the valuable knowledge from the labeled dataset. We validate the pseudo labels from the predecessor models are not just approximate/weak predictions - It can be viewed as knowledge from the previous generation, and can be used as consistency regularization for conventional AL methods solely aiming at higher accuracy.

Preliminaries & Related Work

Active Learning for Classification

Given an arbitrary classification task, assume that there is a large unlabeled dataset $\mathcal{U} = \{x_i\}_{i=1}^n$ of n data samples. The goal of AL is to sample a subset $\mathcal{Q} \subset \mathcal{U}$ to efficiently label so that newly training a deep neural network parameters θ for the classifier $f(x;\theta)$ improves test accuracy. Algorithm 1 describes the conventional procedure in AL. On each iteration t, the learner uses strategy $\mathcal{A}$ (*e.g.*, uncertainty or diversity) to acquire k samples $\mathcal{Q}_t$ from dataset $\mathcal{U}$. Generally, data acquisition model M_t is used for evaluating unlabeled samples according to $\mathcal{A}$. Then, for sample x_i, the learner queries for its oracle label $y_i \in 1, ..., c$, where c is the number of classes. We denote the predicted label of trained model θ_t for x_i by $\hat{y}_i^t = argmax_c f(y_{ic}|x_i;\theta_t)$.

In most AL approaches, a data acquisition model at time t corresponds to the trained classification model at time t-1, *i.e.*, $M_t = \theta_{t-1}$. We call this **monotonic acquisition**, since a naive belief would be assuming the last trained model θ_{t-1} is effective enough to not only provide a good representation for the entire labeled data $\mathcal{L}$ but also estimate acquisition factors (*e.g.*, confidence) for remaining unlabeled data $\mathcal{U}$.

Algorithm 1: Conventional AL procedure

Input: Initial labeled data pool $\mathcal{L}$, unlabeled data pool $\mathcal{U}$, number of queries per iteration (budget) k, number of iterations T, sampling algorithm $\mathcal{A}$
Output: Model parameters θ_T
$\theta_0 \leftarrow$ Train a seed model on $\mathcal{L}$
for *iteration* $t = 1, ..., T$ **do**
  $M_t(x) = f(x;\theta_{t-1})$
  $\mathcal{Q}_t \leftarrow$ Apply $\mathcal{A}(x, M_t, k)$ for $\forall x \in \mathcal{U}$
  $\bar{\mathcal{Q}}_t \leftarrow$ Label queries $\mathcal{Q}_t$ by oracles
  $\mathcal{L} \leftarrow \mathcal{L} \cup \bar{\mathcal{Q}}_t$
  $\mathcal{U} \leftarrow \mathcal{U} \setminus \bar{\mathcal{Q}}_t$
  $\theta_t \leftarrow$ Train a new model on $\mathcal{L}$
end
return θ_T

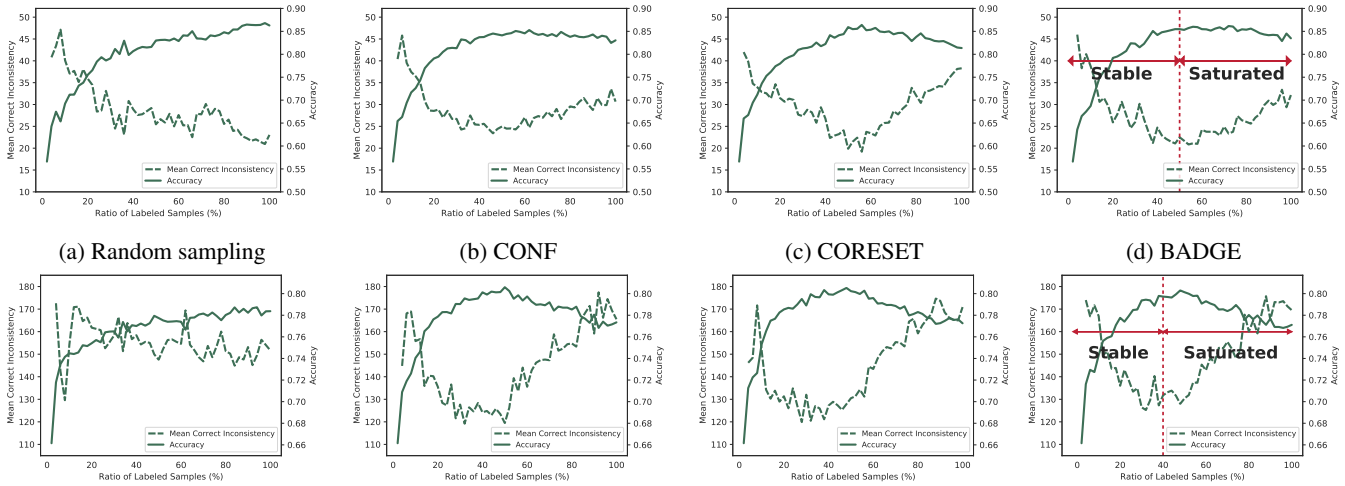
Data Acquisition Strategies in AL

The ultimate goal of AL is to improve the classification accuracy with a fixed annotation budget (Settles 2009; Lowell, Lipton, and Wallace 2018). Existing research efforts on pool based active learning (Lewis and Gale 1994) achieve this goal by focusing on data acquisition based on query strategy and data strategy (Ren et al. 2020). As a query strategy, there are two general approaches to recognize the most appropriate samples (Dasgupta 2011) with monotonic acquisition: uncertainty sampling and diversity sampling. While uncertainty sampling efficiently searches the hypothesis space by finding difficult examples to label (Asghar et al. 2017; He et al. 2019; Ranganathan et al. 2017), diversity sampling exploits heterogeneity in the feature space (Hu, Mac Namee, and Delany 2010; Bodó, Minier, and Csató 2011). Recently, hybrid approaches are proposed (Zhdanov 2019; Ash et al. 2019). Particularly, BADGE (Ash et al. 2019) successfully integrates both aspect by clustering hallucinated gradient vectors based on monotonic acquisition scheme.

Data Acquisition Models in AL

Despite the remarkable success in query strategies, recent research has concerned several limitation of AL. Yun, Kim, and Kim (2020); Wang et al. (2016) point out the difficulty of learning good representation across AL iterations since insufficient annotations may lead to the instability of training models. This indicates that the monotonic acquisition does not ensure the last trained model as a good surrogate of the currently labeled data to identify the informative samples for data acquisition. As a result, Karamcheti et al. (2021); Farquhar, Gal, and Rainforth (2020); Prabhu, Dognin, and Singh (2019) reveal that the acquired samples are vulnerable to sampling bias, and especially, Karamcheti et al. (2021) have presented Dataset Maps (Swayamdipta et al. 2020) of AL, which visualizes harmful outliers preferred by AL methods. Despite these facts, Lowell, Lipton, and Wallace (2018) suggest that the monotonic acquisition is still promising as another remedy using external models (*e.g.*, SVM out of the AL iterations) for data acquisition extremely hampers accuracy of AL.

Motivated by this line of research, in this work, we explore how the limitation of monotonic acquisition can be addressed, in particular, considering consistency as a solution to mitigate the instability of AL iterations. Similar to (Lowell, Lipton, and Wallace 2018) reporting unreliable performance of AL in the NLP field, we choose text classification tasks as our testbed.



(a) Random sampling (b) CONF (c) CORESET (d) BADGE

(e) Random sampling (f) CONF (g) CORESET (h) BADGE

Figure 1: The accuracy and consistency (MCI) of Bi-LSTM model under three acquisition strategies on two text classification test datasets (TREC (a-d) and SST-2 (e-h)); x-axis represents the ratio of labeled samples and y-axis represents the corresponding metrics; We report the average values with five random seeds.

Accuracy-Consistency Dynamics

In this section, we analyze the training dynamics of AL in terms of consistency along with accuracy, observing (catastrophic) *example forgetting event* (Toneva et al. 2019) on the AL iterations: these occur when examples that have been "learnt" (*i.e.*, correctly classified) at some time t in the optimization process are subsequently misclassified – or in other terms "forgotten" – at a time $t{+}\Delta t > t$. Formally,

Definition 1 *(Forgetting and Learning Events) Given a predicted label $\hat{y}_i^t$, let $acc_i^t = \mathbb{1}_{\hat{y}_i^t = y_i}$ be a binary variable indicating whether the example is correctly classified by θ_t. Sample i undergoes a forgetting event when acc_i^t decreases between two different iterations, i.e., $acc_i^t > acc_i^{t+\Delta t}$, where $\Delta t > 0$ (misclassified). Conversely, a learning event has occurred if $acc_i^t < acc_i^{t+\Delta t}$.*

While learning new knowledge is also one of the important factors for generalization ability, our focus is on measuring how well models in AL preserve the learned knowledge. For further analysis, we introduce *Correct Inconsistency* of a model as a measure of how inconsistent the target model is with its predecessor models for a sample. That is, correct inconsistency counts the forgetting events between the model and each of the predecessor models.

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 $0 < \Delta t \leq t$:*

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

As the number of predecessor models are different per AL iteration, to fairly show the degree of inconsistency, we use mean of correct inconsistency for every sample in development split *e.g.*, MCI $=\sum_i \mathbb{C}_i^{(t)}/t$.

In Figure 1, we present both accuracy and MCI of trained models through the full progress of AL. We analyze three data acquisition strategies that are carefully chosen by considering the uncertainty-diversity dichotomy (Yuan, Lin, and Boyd-Graber 2020) along with random strategy. CONF (Wang and Shang 2014), CORESET (Sener and Savarese 2018), and BADGE (Ash et al. 2019) represent three lines of acquisition strategies in AL: uncertainty, diversity, and their hybrid. Across all dataset and acquisition strategies, accuracy and MCI follow the anti-correlated relationship. For convenience of analysis, we divide the training progress into two phases based on the transition of tendency in terms of accuracy: *stable* and *saturated* phases.

In the stable phase, more data leads to more accurate model. Validation accuracy increases on 0-50% of TREC and 0-40% of SST-2, while MCI decreases, where newly labeled samples improve generalization of the trained models. In this phase, AL strategies seek to achieve label efficiency, reaching higher accuracy with a given amount of labeled samples or reversely achieving the same accuracy with less amount of labeled samples. What is notable here is that dramatic improvement of accuracy mostly involves the rapid drop in MCI. These analytical results provide a guide towards the idealistic property of AL methods, which is preserving existing knowledge and simultaneously learning new knowledge. Thus, we hypothesize that relieving forgetting events may contribute to faster and better (*i.e.*, higher accuracy) convergence of AL.

In the saturated phase, the monotonic trends observed in stable phase do not hold. Validation accuracy converges or decreases with the rapid increase of MCI, suggesting that the generalization performance of model deteriorates as even more labeled samples are fed into the trained models.

That is, more data does not always lead to more accurate model, which indicates labeling efforts may be negated in this phase. Although such an extremely undesirable situation in AL is barely addressed by stopping AL iterations in prior work (Ishibashi and Hino 2020), an idealistic AL framework would avoid this phase so that the models can be learned in a more label-efficient manner.

TrustAL: Trustworthy Active Learning

Based on the prior findings on training dynamics of AL procedure, we aim to relieve forgotten knowledge to train better acquisition model that serves as a good surrogate for labeled dataset. A naive way to obtain more generalized models is simply using enough labeled data. However, this approach is not always applicable since budget is limited in AL. Another line of work is using multiple equivalent models (*e.g.*, ensemble) based on complementary nature across different generations. However, this approach is also not always affordable since querying on the huge pool of unlabeled data using multiple models is computationally too expensive.

We now present **TrustAL** (Trustworthy AL) that enables the training of consistent acquisition model that serves as a good reference for labeled dataset in smart and resource-efficient way. **TrustAL** utilizes additional machine generated labels for the purpose of mitigating the forgotten knowledge. Especially, among predecessor models, **TrustAL** identifies a proper expert model that can efficiently contribute to mitigating forgotten samples, which is a novel way of tackling the possible knowledge loss during AL procedure.

Distillation-based Consistency Regularization

Knowledge distillation (Hinton, Vinyals, and Dean 2015) is originally proposed to transfer knowledge from one model (*i.e.*, teacher model) to another (*i.e.*, student model) to compress the size of model. Inspired by recent approaches to transfer knowledge between equivalent models (Furlanello et al. 2018), we propose using the inferior (*e.g.*, less accurate) predecessor model as a teacher model to mitigate example forgetting of the student model (*i.e.*, last trained, superior model) by learning from pseudo labels (*i.e.*logits). This distillation method can be interpreted as a type of consistency regularizer to alleviate forgotten knowledge.

Formally, Algorithm 2 describes the overall procedure of the TrustAL framework on the AL iterations. When given a labeled data pool $\mathcal{L} = \{x_i, y_i\}_{\forall i}$ at the t-th iteration, let L_{CE} be a typical cross-entropy loss with oracle labeled examples, *i.e.*, $\sum_{(x_i,y_i)\in\mathcal{L}} CrossEntropy(y_i, f(x_i;\theta_t))$, and let L_{KL} be a knowledge distillation loss with the pseudo labels of a predecessor model from t-Δt, *i.e.*, $\sum_{(x_i,y_i)\in\mathcal{L}} KL\text{-}Divergence(f(x_i;\theta_{t-\Delta t}), f(x_i;\theta_t))$. On the top of an arbitrary data acquisition method (*e.g.*, CORESET and BADGE), model parameter θ_t produced by TrustAL is obtained by the following optimization:

$$\theta_t = argmin_{\theta_t} L_{CE}(\theta_t) + \alpha \cdot L_{KL}(\theta_{t-\Delta t}, \theta_t) \quad (1)$$

where α is a preference weight. We empirically analyze the effect of varying α in Appendix D.

Algorithm 2: TrustAL

Input: Initial labeled data pool $\mathcal{L}_0$, unlabeled data pool $\mathcal{U}$, number of queries per iteration (budget) k, number of iterations T, sampling algorithm $\mathcal{A}$, fixed development dataset $\mathcal{D}_{dev}$ with size m
Output: Model parameters θ_T
$\theta_0 \leftarrow$ Train a seed model on $\mathcal{L}$
for *iteration* $t = 1, ..., T$ **do**
  $M_t(x) = f(x;\theta_{t-1})$
  $\mathcal{Q}_t \leftarrow$ Apply $\mathcal{A}(x, M_t, k)$ for $\forall x \in \mathcal{U}$
  $\theta_{t-\Delta t} \leftarrow$ TeacherSelection$(\theta_0, ..., \theta_{t-1}, \mathcal{D}_{dev})$
  $\bar{\mathcal{Q}}_t \leftarrow$ Label queries $\mathcal{Q}_t$ by oracles and $f(x;\theta_{t-\Delta t})$
  $\mathcal{L} \leftarrow \mathcal{L} \cup \bar{\mathcal{Q}}_t$
  $\mathcal{U} \leftarrow \mathcal{U} \setminus \bar{\mathcal{Q}}_t$
  $\theta_t \leftarrow \text{argmin}_{\theta_t} L_{CE}(\theta_t) + \alpha \cdot L_{KL}(\theta_{t-\Delta t}, \theta_t)$
end
return θ_T

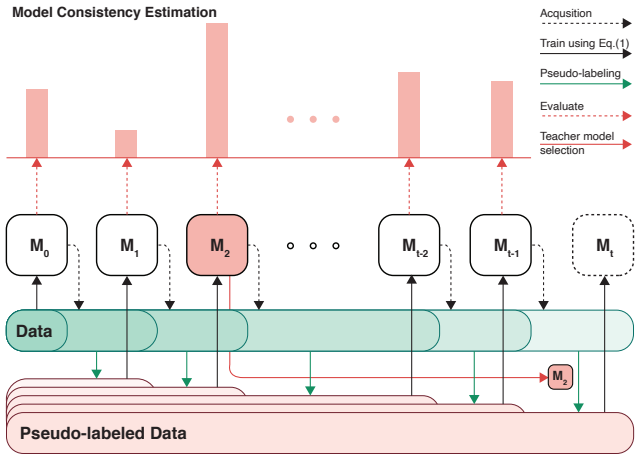


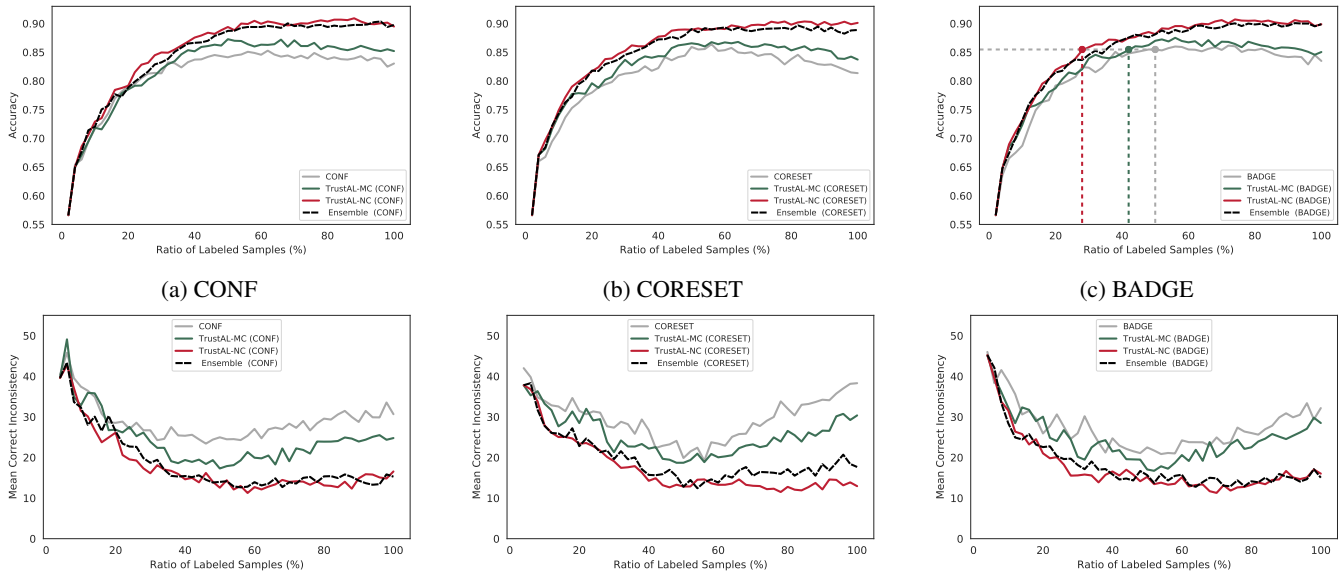
Figure 2: Illustration of TrustAL-NC

This framework motivates us to leverage more sophisticated techniques for knowledge distillation, such as (Yuan et al. 2020; Park et al. 2019). We leave such exploration for future work, as using Eq. (1) works quite well for multiple AL methods in our experiments. Instead, as any predecessor model can be a teacher, we extend this framework to further exploration of teacher selection.

Teacher Selection Strategies

The key factor of TrustAL framework is considering a predecessor model as a specialist model for the forgettable knowledge. As reported in Figure 1, specific to data increments across multiple generations, predecessor models have different status of learned and forgotten knowledge. Therefore, the distillation effects are different in how to select teacher models. Here, we introduce two strategies with TrustAL: monotonic and non-monotonic consistency.

Monotonic Consistency (TrustAL-MC) Basically, we can inherit the monotonic approach not only for data acqui-



(a) CONF (b) CORESET (c) BADGE

(d) CONF (e) CORESET (f) BADGE

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

sition but also for teacher selection, synchronizing both, *i.e.*, always $\theta_{t-\Delta t} = \theta_{t-1} = M_t$. This allows to iteratively transfer the learned knowledge generation by generation.

Non-monotonic Consistency (TrustAL-NC) *Correct Inconsistency* (Definition 2) can be a strong signal to indicate which sample is forgettable for the current acquisition model. Using such sample level inconsistency, we aim at choosing the predecessor model with the learned knowledge especially of those forgettable samples. This allows to transfer the easily forgettable knowledge from one of the predecessor model, not always from $\theta_{t-\Delta t} = \theta_{t-1} = M_t$ as described in Figure 2.

Specifically, given a development dataset $\mathcal{D}_{dev}$ with m samples, let $\mathbb{C}^t$ be a vector of correct inconsistency values of M_t $(= \theta_{t-1})$ at the t-th iteration for all m samples, *i.e.*, $\langle \mathbb{C}_1^{(t-1)}, ..., \mathbb{C}_m^{(t-1)} \rangle \in \mathbb{R}^m$. For the purpose of using this vector as importance weights for samples, we normalize $\mathbb{C}^t$ into $\tilde{\mathbb{C}}^t$ where $\sum_{\forall i} \tilde{\mathbb{C}}_i^t = 1$ by a softmax function. We note that the sample x_i with high importance weight $\tilde{\mathbb{C}}_i^t$ means easily forgettable sample for M_t. Based on such consistency-aware sample importance, we define a function $g(\theta_{t-\Delta t}, M_t)$ of measuring how reliably a predecessor model $\theta_{t-\Delta t}$ can be a synergetic teacher with the data acquisition of M_t, by an weighted accuracy as:

$$g(\theta_{t-\Delta t}, M_t) = \tilde{\mathbb{C}}^{t\top} \langle acc_1^{t-\Delta t}, ..., acc_m^{t-\Delta t} \rangle / m \quad (2)$$

High $g(\theta_{t-\Delta t}, M_t)$ implies that the teacher model $\theta_{t-\Delta t}$ tends to have the knowledge of forgettable examples for the current data acquisition model M_t, and vice versa. Based on this, we can select a predecessor having the maximum value, as a teacher model to teach a new model θ_t:

$$\underset{1<\Delta t \leq t}{\operatorname{argmax}}\, g(\theta_{t-\Delta t}, M_t) \quad (3)$$

Development Set Strategies

One of the plausible tools to estimate the learning status of AL generations is development set as it is often used for validation process. In fact, TrustAL-NC catches forgetting signals as a by-product of the validation process. The experiment on the robustness of TrustAL-NC on the size of development set shows marginal performance decrease even when halving development set size, which resolve concerns about keeping development set in label-scarce situation. Full empirical results are presented in Appendix D.

Experiments

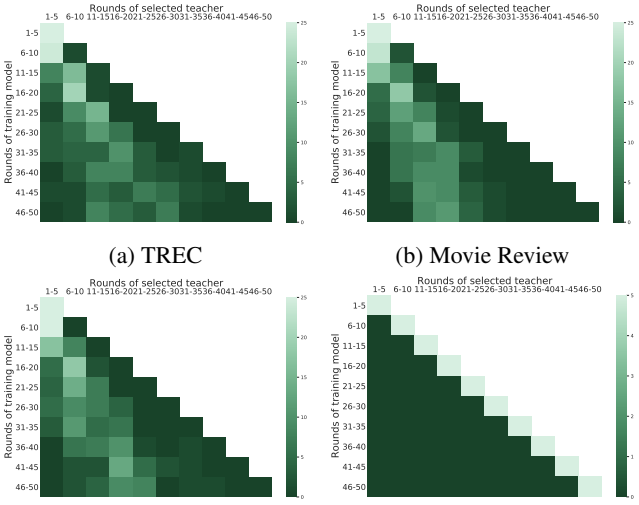
Experimental Setup

Dataset We use three text classification datasets, TREC (Roth et al. 2002), Movie review (Pang and Lee 2005) and SST-2 (Socher et al. 2013), which are widely used in AL (Lowell, Lipton, and Wallace 2018; Siddhant and Lipton 2018; Yuan, Lin, and Boyd-Graber 2020) and statistically diverse. The data statistics are presented in Appendix A for more details.

Baselines As TrustAL is orthogonally applicable to any data acquisition strategy, for the purpose of better analysis, we use the following three acquisition methods as baselines.

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

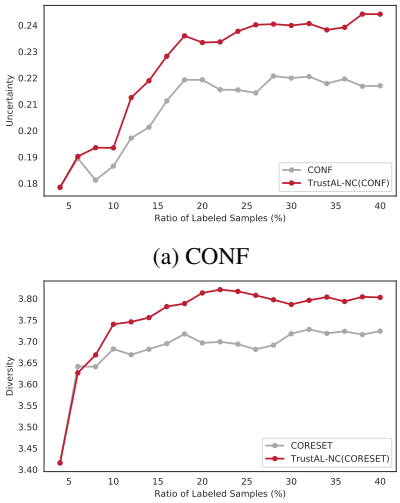
More details on the baselines are discussed in Appendix B.



(a) TREC (b) Movie Review

(c) SST-2 (d) Trust-MC on all dataset

Figure 4: Heatmap showing frequency of the model selection in TrustAL-NC with BADGE on three datasets. The x- and y-axis indicate the bins of the selected teacher model and the training model, respectively. In each round the size of labeled sample increases by 2%. Each bin consists of 5 rounds. Brighter cell indicates more frequently selected bin.



(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.

Implementation For all three datasets, we follow the commonly used default settings in AL for text classification (Liu et al. 2021; Zhou et al. 2021; Lowell, Lipton, and Wallace 2018; Siddhant and Lipton 2018): Bi-LSTM (Hochreiter and Schmidhuber 1997) is adopted as a base model architecture; In each iteration of AL, training a classification model from scratch (not by incremental manner) with the entire labeled samples gathered, to avoid the training issues with warm-starting (Ash and Adams 2020). Note that the development set is held out in every experiments so that it is not used for training models. We describe our implementation details in Appendix C.

Results and Discussion

We present the empirical findings for the following three research questions:
RQ1: Does TrustAL outperform AL baselines?
RQ2: How does TrustAL help data acquisition?
RQ3: Does TrustAL make consistent and robust models?
Additional experiments on hyperparameter sensitivity are presented in Appendix D.

Overall Performance (RQ1) First, we compare the performance of AL methods across AL iterations with and without TrustAL. Figure 3 shows accuracy and MCI of AL methods on TREC. The empirical results in SST-2 and Movie Review are presented in Appendix F.

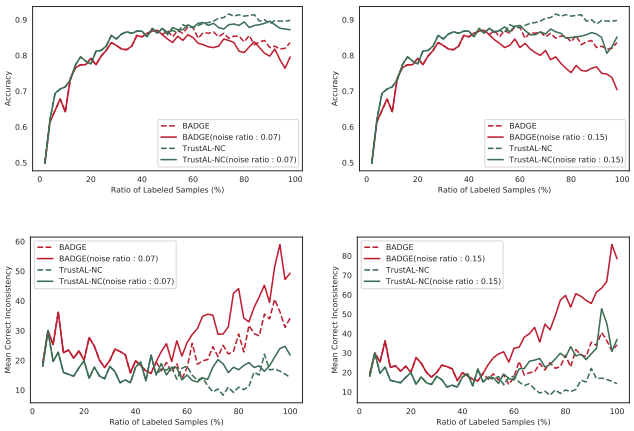
Overall, AL strategies combined with TrustAL-NC/MC show improved label efficiency and relieved MCI compared to stand-alone baselines in all datasets. The models trained with TrustAL framework require much smaller number of labeled samples to achieve the same level of accuracy than baselines. To facilitate the comparison of label efficiency of TrustAL and a baseline, we draw the horizontal reference line where the baseline starts to show convergence in Figure 3 (c). As a result, we find that TrustAL-MC and TrustAL-NC require only 40% and 30% of the training data pool, respectively, while baselines requires 50% of total training data to reach the same level of accuracy. This result suggests that keeping consistency of model in AL is an essential criterion, and TrustAL successfully satisfies the ultimate goal of AL, *i.e.*, improving the label efficiency.

Further, TrustAL-NC performs comparably to ensemble based distillation method (Fukuda et al. 2017) which aims to distill the ensembled (i.e. averaged) probability distribution of multiple models. This indicates TrustAL-NC selects a teacher model that can effectively relieve forgotten knowledge, even without using all predecessor models. Figure 4 visualizes the behavior of teacher selection by TrustAL. The figure shows that TrustAL-MC selects the most recent model as its definition and TrustAL-NC chooses the teacher model based on the consistency guidance. While preferring the more generalized teacher models from the end of the stable learning stages (16-20) rather than earlier stages, TrustAL-NC also selects earlier generation that might be inferior but professional in terms of forgotten knowledge. That is, TrustAL-NC can select complementary models for forgotten knowledge in an automatic manner.

Data Acquisition Quality (RQ2) Having tested for the overall accuracy and MCI of TrustAL, we evaluate the quality of data acquisition results when using TrustAL. To discuss how TrustAL affects data acquisition, we analyze TrustAL-NC based on the two distinctive strategies on data acquisition: uncertainty and diversity. Note that, we choose to compare acquisition quality of stable phase only since the label efficiency of saturated phase is negative for traditional

		TREC	Movie review	SST-2
A	baseline	0.726	0.637	0.686
	Ensemble	0.770	0.669	0.727
	TrustAL-MC	0.743	0.654	0.705
	TrustAL-NC	**0.774**	**0.676**	**0.730**
B	baseline	0.727	0.627	0.697
	Ensemble	0.777	0.658	0.724
	TrustAL-MC	0.753	0.654	0.707
	TrustAL-NC	**0.785**	**0.665**	**0.724**
C	baseline	0.735	0.636	0.681
	Ensemble	0.773	0.668	0.722
	TrustAL-MC	0.748	0.653	0.711
	TrustAL-NC	**0.780**	**0.670**	**0.729**

Table 1: Correct consistency of TrustAL-NC with (A) CONF (B) CORESET and (C) BADGE



(a) Noise ratio : 7% (b) Noise ratio : 15%

Figure 6: Robustness analysis varying the ratio of noise. Accuracy and MCI are shown in pair for each noise ratio.

data acquisition strategies in AL.

For uncertainty, following (Yuan, Lin, and Boyd-Graber 2020), we first obtain a reference model trained on the full training data of a target task, then measure the uncertainty of samples selected on each iteration. Specifically, by using Shannon Entropy (Shannon 2001), we compute the entropy of the predicted probability distribution of individual samples, and report their average values for each AL iteration in Figure 5a, where a higher value implies each iteration successfully acquires uncertain samples.

For diversity, we reuse the reference model to encode the full training data into a feature space, then obtain the k disjoint sets of the all training data by K-means algorithm where we set k as the number of samples acquired per AL iterations. Then, based on these k groups, we measure the diversity of a sample set selected on each iteration, by computing the entropy of a cluster distribution of the selected samples, which we report in Figure 5b. The measure shows whether the samples are uniformly picked among the clusters, since diversely acquired samples would belong to different clusters (Ash et al. 2019).

As shown in Figure 5, providing more labeled data leads to improvement of uncertainty and diversity for both baseline and TrustAL. Considering the reported increase of generalization performance in RQ1, this suggests that better model training leads to better acquisition, strengthening model's ability to identify more informative samples. For CONF and CORESET each representing uncertainty and diversity based strategies, we observe that TrustAL largely improves the quality of acquisition across the AL procedure. Since TrustAL aims to enhance the model's ability of surrogating the labeled dataset, we believe that TrustAL can be orthogonally and effectively applicable to any acquisition strategy with its synergetic nature.

Model Consistency and Robustness (RQ3) The correct consistency (Wang et al. 2020), $i.e., avg_{\forall i} \mathbb{1}_{\hat{y}_i^m = \hat{y}_i^n = y_i}$, is a measure for the consistency between m^{th} and n^{th} generation models. By measuring the correct consistency between any two models, we demonstrate that our framework contributes not only to the label efficiency, but also to the overall consistency of model generations. Shown in Table 1, TrustAL-NC shows better correct consistency than baselines. This reveals that models in AL iterations accord with each other for the correctly classified samples, which is also related to user's trust on system (Wang et al. 2020).

To demonstrate the robustness of TrustAL on the careless transition from human labeling, we deliberately corrupt the acquired samples by randomly flipping certain ratio of labels. Specifically, after stable phase, we corrupt 7% and 15% of the labels. In Figure 6, the result of BADGE and TrustAL(BADGE) are shown. Since other strategies show similar behaviors, we only present the result of BADGE. When the stable phase ends, the noisy labels cause the rapid increase of forgotten knowledge. Based on this observation, we believe that one of the possible suspects of performance degradation in the saturated phase might be noisy examples. Despite such degradation, TrustAL performs more robustly which is in strike contrast to the baseline. With 7% of noise, TrustAL even shows comparable result to the one trained without noise. This result shows that TrustAL is robust to accidental noise in labels produced by human annotators since TrustAL regularizes the negative impact of such labels by pursuing consistency as an additional objective in training.

Conclusion

In this paper, we debunk the monotonicity assumption which is a common belief in conventional AL methods by empirical observation of example forgetting. For that, we present TrustAL, an effective and robust framework that uses the predecessor model as an expert model for knowledge distillation to compensate the loss of knowledge between data and model. Especially, our framework can be orthogonally applicable to existing data acquisition in a highly efficient way. Further, we present multi-pronged analysis for our method through extensive experiments.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University)). And this research was partly supported by the MSIT(Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program)(2021-11-1603) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). Jinyoung Yeo is a corresponding author.

References

Asghar, N.; Poupart, P.; Jiang, X.; and Li, H. 2017. Deep Active Learning for Dialogue Generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, 78–83.

Ash, J.; and Adams, R. P. 2020. On Warm-Starting Neural Network Training. *Advances in Neural Information Processing Systems*, 33.

Ash, J. T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; and Agarwal, A. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *ICLR*.

Bodó, Z.; Minier, Z.; and Csató, L. 2011. Active learning with clustering. In *JMLR*.

Bouguelia, M.-R.; Nowaczyk, S.; Santosh, K.; and Verikas, A. 2018. Agreeing to disagree: Active learning with noisy labels without crowdsourcing. *International Journal of Machine Learning and Cybernetics*, 9(8): 1307–1319.

Dasgupta, S. 2011. Two faces of active learning. *Theoretical computer science*, 412(19): 1767–1781.

Farquhar, S.; Gal, Y.; and Rainforth, T. 2020. On Statistical Bias In Active Learning: How and When to Fix It. In *International Conference on Learning Representations*.

Fukuda, T.; Suzuki, M.; Kurata, G.; Thomas, S.; Cui, J.; and Ramabhadran, B. 2017. Efficient Knowledge Distillation from an Ensemble of Teachers. In *INTERSPEECH*.

Furlanello, T.; Lipton, Z.; Tschannen, M.; Itti, L.; and Anandkumar, A. 2018. Born again neural networks. In *International Conference on Machine Learning*, 1607–1616. PMLR.

He, T.; Jin, X.; Ding, G.; Yi, L.; and Yan, C. 2019. Towards better uncertainty sampling: Active learning with multiple views for deep convolutional neural network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 1360–1365. IEEE.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hochreiter, S.; and Schmidhuber, J. 1997. LSTM can solve hard long time lag problems. *NIPS*.

Hu, R.; Mac Namee, B.; and Delany, S. J. 2010. Off to a good start: Using clustering to select the initial training set in active learning. In *Twenty-Third International FLAIRS Conference*.

Ishibashi, H.; and Hino, H. 2020. Stopping criterion for active learning based on deterministic generalization bounds. In *International Conference on Artificial Intelligence and Statistics*, 386–397. PMLR.

Karamcheti, S.; Krishna, R.; Fei-Fei, L.; and Manning, C. 2021. Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Lewis, D. D.; and Gale, W. A. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, 3–12. Springer.

Liu, Q.; Zhu, Y.; Liu, Z.; Zhang, Y.; and Wu, S. 2021. Deep Active Learning for Text Classification with Diverse Interpretations. *arXiv preprint arXiv:2108.10687*.

Lowell, D.; Lipton, Z. C.; and Wallace, B. C. 2018. Practical obstacles to deploying active learning. *EMNLP-IJCNLP*.

Pang, B.; and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *ACL*.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3967–3976.

Prabhu, A.; Dognin, C.; and Singh, M. 2019. Sampling Bias in Deep Active Classification: An Empirical Study. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4058–4068.

Ranganathan, H.; Venkateswara, H.; Chakraborty, S.; and Panchanathan, S. 2017. Deep active learning for image classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, 3934–3938. IEEE.

Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Chen, X.; and Wang, X. 2020. A Survey of Deep Active Learning. *arXiv preprint arXiv:2009.00236*.

Roth, D.; Cumby, C.; Li, X.; Morie, P.; Nagarajan, R.; Rizzolo, N.; Small, K.; and Yih, W.-t. 2002. Question-answering via enhanced understanding of questions. In *TREC*. Citeseer.

Sener, O.; and Savarese, S. 2018. Active learning for convolutional neural networks: A core-set approach. *ICLR*.

Settles, B. 2009. Active learning literature survey.

Shannon, C. E. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1): 3–55.

Siddhant, A.; and Lipton, Z. C. 2018. Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2904–2909.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models

for semantic compositionality over a sentiment treebank. In *EMNLP*.

Swayamdipta, S.; Schwartz, R.; Lourie, N.; Wang, Y.; Hajishirzi, H.; Smith, N. A.; and Choi, Y. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9275–9293.

Toneva, M.; Sordoni, A.; Combes, R. T. d.; Trischler, A.; Bengio, Y.; and Gordon, G. J. 2019. An empirical study of example forgetting during deep neural network learning. *ICLR*.

Wang, D.; and Shang, Y. 2014. A new active labeling method for deep learning. In *2014 International joint conference on neural networks (IJCNN)*, 112–119. IEEE.

Wang, K.; Zhang, D.; Li, Y.; Zhang, R.; and Lin, L. 2016. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12): 2591–2600.

Wang, L.; Ghosh, D.; Gonzalez Diaz, M.; Farahat, A.; Alam, M.; Gupta, C.; Chen, J.; and Marathe, M. 2020. Wisdom of the Ensemble: Improving Consistency of Deep Learning Models. *Advances in Neural Information Processing Systems*, 33.

Xu, Z.; Yu, K.; Tresp, V.; Xu, X.; and Wang, J. 2003. Representative sampling for text classification using support vector machines. In *ECIR*, 393–407. Springer.

Yuan, L.; Tay, F. E.; Li, G.; Wang, T.; and Feng, J. 2020. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3903–3911.

Yuan, M.; Lin, H.-T.; and Boyd-Graber, J. 2020. Cold-start active learning through self-supervised language modeling. *EMNLP*.

Yun, J.; Kim, B.; and Kim, J. 2020. Weight Decay Scheduling and Knowledge Distillation for Active Learning. In *European Conference on Computer Vision*, 431–447. Springer.

Zhdanov, F. 2019. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*.

Zhou, Y.; Renduchintala, A.; Li, X.; Wang, S.; Mehdad, Y.; and Ghoshal, A. 2021. Towards Understanding the Behaviors of Optimal Deep Active Learning Algorithms. In *International Conference on Artificial Intelligence and Statistics*, 1486–1494. PMLR.