Adversarial Active Learning for Sequence Labeling and Generation

Yue Deng[1], **KaWai Chen**[2], **Yilin Shen**[1], **Hongxia Jin**[1]
[1] AI Center, Samsung Research America, Mountain View, CA, USA
[2] Department of Electrical and Computer Engineering, University of California, San Diego
{y1.deng, yilin.shen, hongxia.jin}@samsung.com, w0chen@eng.ucsd.edu

Abstract

We introduce an active learning framework for general sequence learning tasks including sequence labeling and generation. Most existing active learning algorithms mainly rely on an uncertainty measure derived from the probabilistic classifier for query sample selection. However, such approaches suffer from two shortcomings in the context of sequence learning including 1) cold start problem and 2) label sampling dilemma. To overcome these shortcomings, we propose a deep-learning-based active learning framework to directly identify query samples from the perspective of adversarial learning. Our approach intends to offer labeling priorities for sequences whose information content are least covered by existing labeled data. We verify our sequence-based active learning approach on two tasks including sequence labeling and sequence generation.

这个什么意思

1 Introduction

Active learning (AL) is a traditional approach to solve supervised learning problems without sufficient labels. While there have been many existing AL works proposed for classification problems [Settles, 2010; Scheffer *et al.*, 2001; Deng *et al.*, 2013], active learning algorithms for sequences are still not widely discussed. With the growing interests in AI research, many newly emerged problems are exactly defined in the scope of sequence learning including image captioning [Vinyals *et al.*, 2015], machine translation [Luong *et al.*,] and natural language understanding [Dong and Lapata, 2016]. Compared with classification tasks that only need one label for a sample, sequence learning tasks require a series of token-level labels for a whole sequence. Precise annotations for sequences are not only labor-consuming but may also require very specific domain knowledge [Dong and Lapata, 2016] that are not easily accomplished by crowd-sourcing workers. This apparent difficulty in sequence labeling exactly motivates our explorations of more effective active learning approaches for sequences.

第一个强调token
级标注的

Existing active learning strategies mainly rely on some uncertainty measures derived from a classifier for query sample selection [Cohn *et al.*, 1994; Settles, 2010]. These uncertainty measures can be defined from various perspectives including probabilistic confidence [Culotta and McCallum, 2005], margin value [Scheffer *et al.*, 2001], entropy [Deng *et al.*, 2016], fisher information [Sutton and McCallum, 2006; Bao *et al.*, 2017] and a score voted by several base models [Seung *et al.*, 1992; Deng *et al.*, 2017b]. While these active learning algorithms work well for data classification tasks, they are unfortunately not easily extended to solving sequence learning problems due to the complexity of the label space. Consider a label sequence with p tokens and each token can belong to k possible classes, then there are k^p possible combinations of the label sequence. This complexity can grow exponentially with the length of the output.

说的好

We consider two major challenges faced by existing active learning approaches in handling sequence learning tasks: 1) cold start problem and 2) label-sampling dilemma. The first cold-start challenge is mainly due to the complexity of the learning system for structured prediction. Unlike classification tasks that just need a simple probabilistic classifier, the predictor for sequences are configured within a complex recurrent structure, e.g. a LSTM. Training a structured predictor with very limited labeled sequence can easily lead to a biased estimation. If the predictor itself is seriously biased, how can we trust the uncertain measure derived from it? This cold start problem easily happens during the initial steps of active learning when there are only insufficient labeled samples in hand. The second label sampling dilemma is ascribed to the inability of the full enumeration of all possible sequence labels. In detail, when calculating an uncertainty score e.g., the entropy, for a sequence, all possible label combinations should be taken into account (see Eq. 3) that can become impossible when the output sequences are too long. Therefore, only approximated uncertainty measures can be used as a surrogate for sequence-based active learning.

少量数据训练得到
的LSTM会过拟合，
就太bias

因为LSTM有bias就
不相信它吗？
我觉得可以部分相
信的。

可以不用熵啊，熵
的表现也不好，你
用边缘不就行了？

To overcome the aforementioned limitations, we propose a new active learning framework for sequences inspired by adversarial learning. Our approach alleviates the demands on the structured predictor for query sample selection. The proposed adversarial active learning framework incorporates a neural network to explicitly assert each sample's informativeness with regard to labeled data. The easily-induced active score avoids heavy computations in sampling the whole label space and can improve the active learning efficiency by more than 100 times on some large datasets.

我不信；
我觉得你在吹牛
。

2 Preliminaries

2.1 Active Learning for Sequences

In this part, we review some existing active learning approaches for sequence learning [Settles and Craven, 2008]. The first widely used uncertainty measure defined from the probabilistic classifier is the least confidence (LC) score:

$$\psi_{LC}(x^U) = 1 - P(y^*|x^U), \tag{1}$$

where y^* is the most likely label sequence and x^U is an unlabeled sample. While this measure is intuitive, the calculation of the most likely label sequence is not easy. It requires the dynamic programming to find the Viterbi parse. Similarly, the margin term can also be used to define the uncertainty for an unknown sample:

$$\psi_M(x^U) = P((y_2^*|x^U)), -P(y_1^*|x^U), \tag{2}$$

where y_1^* and y_2^* are the first and second best label sequences, respectively.

The sequence entropy term is also widely used as an uncertainty score [Settles and Craven, 2008]:

$$\psi_{EN}(x^U) = -\sum_{y^p} P(y^p|x^U) \log P(y^p|x^U) \tag{3}$$

where y^p ranges over all possible label sequences for input x^U. We have also noted that the number of possible labeling grows exponentially with the desired output sequence length. For the ease of computation, we follow previous work to employ an approximated N-best sequence entropy (NSE) [Kim *et al.*, 2006],

$$\psi_{NSE}(x^U) = -\sum_{y^p \in \mathcal{N}} P(y^p|x^U) \log P(y^p|x^U) \tag{4}$$

$\mathcal{N} = \{(y^1)^p, ..., (y^N)^p\}$ corresponds to N most likely parses. These N most likely parses can obtained through beam search [Koehn, 2004]. The labeling priority should be given to samples with high entropy (corresponding to low confidence).

While the definition of these terms are different, they are all closely related to the structured predictor. The calculations of these uncertainty scores are also not trivial and may require some algorithmic explorations such as the dynamic programming or the beam search. When the candidate samples' quantity is large, the calculation of such complexity uncertainty measures can take a quite long while in scoring all individual samples from the data pool. These obvious shortcomings motivate us to design more efficient and advanced active learning strategies for sequence learning. In this work, we propose such a desired framework from adversarial learning [Deng *et al.*, 2017c]. The active learning strategy in our model is not related to the structured output predictor and hence can conduct query samples scoring in light on very large dataset.

2.2 Encoder-decoder Framework

Before going to the details about our active learning model, we will first review the prevalent encoder-decoder framework for sequence learning. This generic encoder-decoder model serves as the basic building block of our active learning system. We denote $(x^L, y^L) \sim (X^L, Y^L)$ as a pair of labeled sample, where x^L is the input data that can be of any type including images, speeches and texts depending on different learning tasks; and y^L is the targeted output sequence composed of p tokens $y^L = \{y_1^L...y_p^L\}$. A feature encoder $M()$ is established to map the input x^L to a latent representation $z^L = M(x^L)$. $M()$ can be a convolution neural network for image data or a recurrent neural network for speeches and texts. Then, a decoder $C()$ adopts z^L as a conditional input and sequentially predicts each token in y^P:

我怎么感觉你这个公式有问题呢？

$$\begin{aligned} &P(y^p = \{y_1^P...y_q^P\}|x) \\ &= P(y_1^p|z^L = M(x^L)) \prod_{t=1}^{T} P(y_t^p|y_1^p...y_{t-1}^p, z^L), \end{aligned} \tag{5}$$

The above generative probability can be well modeled by a recurrent neural network, e.g., a LSTM [Deng *et al.*, 2017a]. The encoded latent representation z^L is used as the 'starting key' at step zero. Then, it sequentially outputs each token y_t based on the tth step's input and the memory vector maintained by the recurrent neural network [Sutskever *et al.*, 2014]. The training loss of this sequence learning part is obtained by counting the differences between the predicted sequence y^P and the ground truth labels y^L:

$$\mathcal{L}_s(X^L, Y^L) = \sum_{(x^L, y^L) \sim (X^L, Y^L)} L(y^L, y^P) \tag{6}$$

We noted that both y^L is the labeled sequence; and predicted sequence y^P is generated by a function of x^L (see Eq.5) ; L can be arbitrary losses defined over two sequences such as the prevalent cross-entropy. Here, we just briefly introduced this encoder-decoder framework and interested readers are refereed to [Sutskever *et al.*, 2014; Xu *et al.*, 2015] for details.

3 Adversarial Active Learning for Sequences

3.1 ALISE Model

It is conceivable that sequence learning model requires a huge amount of labeled data for robust training. We hence consider developing an active learning algorithm to facilitate the whole labeling process. In our approach, we consider defining an active score based on the informativeness of an unlabeled sample x^U with respect to all labeled samples X^L:

$$s(x^U) = sim(x^U, X^L) \tag{7}$$

where X^L is the set containing all labeled samples and $sim(\cdot, \cdot)$ defines a similarity score between a point x^U and a training set X^L composed of labeled samples. The score in Eq.7 helps to rank unlabeled samples based on their inherent informativeness similarity to existing labeled data. A small similarity score implies the certain unlabeled sample is not related to any labeled samples in training set and vice versa. The labeling priority is offered to samples with low similarity scores.

We take image captioning as an intuitive instance to explain the rational behind our active scoring approach. In the training set, most images and their corresponding descriptions are about human sports such as skating, running and swimming. Then, we have access to two extra unlabeled images that are respectively related to "swimming" and "a plate

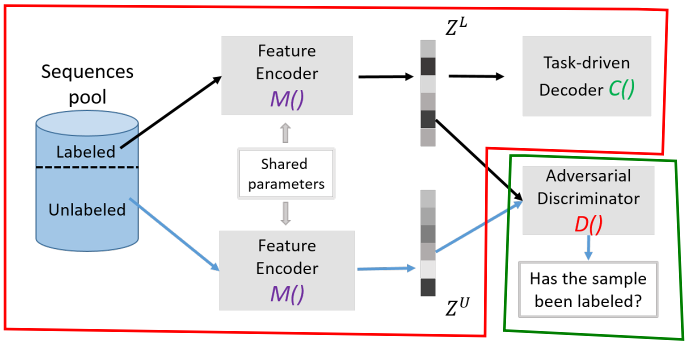

Figure 1: An overview of Adversarial Active Learning for sequences (ALISE). The black and blue arrows respectively indicate flows for labeled and unlabeled samples.

of food". With those annotated images in hand, which unlabeled image should be labeled first? There is no doubt that the image about 'food' should be sent out for captioning by human. This is because there are already some swimming images in the existing training pool and adding another similar image may not offer too much 'new' knowledge to the learning system. On the contrary, the image about food is not covered in existing training set and its captions can bring more valuable complementary information.

However, the problem still remains in how to quantitatively evaluate the informativeness similarity between an unlabeled sample with the labeled data pool. For sequence data, the similarity calculation itself is a difficult problem due to the variations in sequence lengths. Existing approaches mainly play kernel tricks to map the original sequences into a kernel space with fixed-length feature representations [Settles and Craven, 2008]. However, the selection of an appropriate kernel requires sophisticated domain knowledge and the best kernel can vary from task to task. For instance, the best kernels for Chinese and French sentences are obviously not the same.

句子变成，所以相似性很难算！

使用attention和度量学习

In this work, we propose a new adversarial active learning model for sequences (ALISE). In our ALISE, we consider designing a discriminator network $D(\cdot)$ to directly outputs the informativeness similarity scores for unlabeled samples. In Fig. 1, we pass both a labeled sample x^L and an unlabeled sample x^u trough the same feature encoder M (shared parameters), then we get $z^L = M(x^L)$ (latent representation for labeled data) and $z^U = M(x^U)$ (latent representation for unlabeled data). These two latent representations are further fed into the discriminator network (D), which is trained to classify whether the certain data is sampled from labeled or unlabeled data pool. The output of the D is a sigmoid function that indicates how likely the certain sample is from the labeled pool.

The learning objectives of M and D are involved in an adversarial process. From the aspect of encoder M, it intends to map all data to a latent space where both labeled and unlabeled data can follow very similar probabilistic distributions. In the most ideal scenario that if z^L and z^U follow exactly the same generative probability, then the decoder C trained with z^L should also seamlessly work on latent representations z^U obtained from unlabeled sample x^U. Therefore, the encoder $M()$ intends to fool the discriminator to regard all latent representations (z^L and z^U) as already labeled. Mathematically, it encourages the discriminator D to output a score 1 for both z^L and z^U. The corresponding loss is modeled by the cross-entropy in the first two terms of the following equation:

希望：有标注数据和无标注数据有相似的概率分布。
是否说明实际上是没有的？即两种数据分布对于编码器来说是不一致的。

$$\begin{aligned}\min \mathcal{L}_M = &-\mathbb{E}_{x^L \sim X^L}[(\log D(M(x^L))] \\ &-\mathbb{E}_{x^U \sim X^U}[\log D(M(x^U))] + \lambda \mathcal{L}_s(X^L, Y^L),\end{aligned} \tag{8}$$

In addition to the cross-entropy loss defined on the discriminator side, the above equation also takes the supervised loss in Eq.6 into consideration, i.e., in the third term. In all, the learning objectives of the feature encoder M are concluded as two-fold: 1) fool the discriminator and 2) improve the fitting quality on labeled data. These two learning objectives are balanced by a hyper-parameter λ.

The learning objective of the discriminator D goes against to the objective in Eq. 8. The discriminator is trained to correctly assign $z^L = M(x^L)$ to labeled category ($D(Z^L) = 1$) and $z^U = M(x^U)$ to unlabeled class ($D(Z^U) = 0$). The corresponding learning objective of D is also defined by the cross-entropy:

$$\begin{aligned}\min \mathcal{L}_D = &-\mathbb{E}_{x^L \sim X^L}[(\log D(M(x^L))] \\ &-\mathbb{E}_{x^U \sim X^U}[\log(1 - D(M(x^U)))]\end{aligned} \tag{9}$$

This adversarial discriminator D exactly serves the purpose of distribution comparisons between two set of samples. In GAN work [Deng *et al.*, 2017c], it is indicated that the adversarial discriminator implicitly compares the generative distributions between real data and fake data. Here, we borrowed the same adversarial learning idea to compare the distributions between labeled and unlabeled samples. In GAN (resp. our ALISE model), the discriminator outputs low scores for those fake (resp. unlabeled) samples that are mostly not similar to real images (resp. labeled data). Therefore, the score from this discriminator already serves as an informativeness similarity score that could be directly used for Eq.7. The feature encoder M, sequence decoder C and adversarial discriminator D can all be trained in an alternative manner by iteratively optimizing the objectives in Eq.8 and Eq.9. We have detailed the learning steps in Algorithm 1.

3.2 Active Scoring

After well training, we can pass all unlabeled samples through M and D to get their corresponding score by ALISE framework, i.e.,

$$s(x^U) = D(M(x^U)) \in (0, 1), \forall x^U \in X^U \tag{10}$$

The score $s = 1$ (resp. $s = 0$) means the information content of the certain unlabeled sample is most (resp. least) covered by the existing labeled data. Apparently, those samples with lowest scores should be sent out for labeling because they carry most valuable information in complementary to the current labeled data.

It is noted that our ALISE approach does not rely on the structured predictor (i.e. the decoder C) for uncertainty measure calculation. However, we can still consider incorporating existing predictor-dependent uncertainty scores into our

可以只用对抗得分，也可以加入解码器的预测得分

Algorithm 1: ALISE Learning

```
Input          : A data pool composed of labeled and unlabeled
                 data $X = \{X^L, X^U\}$; $X^L$ are paired with
                 sequence label $Y^L = \{y_1^L ... y_p^L\}$;
Initialization: Initialize parameters in encoder network $M$ and
                 decoder network $C$ by training an
                 encoder-decoder framework with available
                 training samples $(X^L, Y^L)$; Initialize
                 parameters in the discriminator network $D$
                 randomly.
for epoch=1...K do
    for all mini-batches $(x^L, y^L) \sim (X^L, Y^L)$ and $x^U \sim X^U$ do
        Minimize the loss $\mathcal{L}_M$ in Eq.8 to update parameters in
          the encoder network $M$ and decoder network $C$
        Minimize the loss $\mathcal{L}_D$ in Eq.9 and update parameters
          in discriminator network $D$
    end
end
Output         : The well trained $M, C$ and $D$
```

没有讲清楚整个流程，
因为没有涉及到标注数据

framework. Because the ALISE framework has already been built with a probabilistic decoder $C()$, then the calculations of uncertainty measures from it is natural and convenient. In such a combinational setting, we can first select K top samples selected by the adversarial discriminator. Then, within these K samples, we further calculate their sequence-based uncertainty scores $\psi(x^U)$ (e.g. the sequence entropy) as introduced in Section 2.1. The top k samples with highest uncertainty scores are selected as query samples for labeling. These candidate query samples are mainly determined by the adversarial discriminator and the probabilistic decoder only provides auxiliary information for fine-grained selection. Moreover, the complexity for sequence-based uncertainty measure computations have also been reduced. This is because the uncertainty measure is only required to be computed on K candidate samples selected by ALISE rather than the whole pool of unlabeled samples.

那你为什么还要计
算不确定性呢？

While there are some early works that also use the 'buzzwords' adversarial active learning, they are totally different from our ALISE. First, the work in [Zhu and Bento, 2017] used the GAN model to generate fake images and then labeling those fake images to augment training set. ALISE does not generate any fake sample and just borrows the adversarial learning objective for sample scoring. The work in [Miller *et al.*, 2014] is totally none related to adversarial learning. It just uses traditional active learning approach to solve the adversarial attract problem in security domain.

4 Experiments

In this part, we investigate the performances of ALISE on two sequence learning tasks including slot filling and image captioning.

4.1 Slot Filling

Slot filling is a basic component of spoken language understanding. It can be viewed as a sequence labeling problem, where both the input and output label sequences are of the same length. This part of experiments were mainly conducted on the ATIS (Airline Travel Information Systems) dataset [Hemphill *et al.*, 1990]. We obtained ATIS text corpus that was used in [Liu and Lane, 2016] and [Deoras and Sarikaya, 2013] for active learning. For instance, an input sentence in ATIS x^L = {business, class, fare,from, SF, to, LA} can be parsed as a label sequence y^L = {B-class-type, I-class-type, O, O, B-from-loc, B-to-loc}. This studied dataset contains 5138 utterances with annotated slot labels.

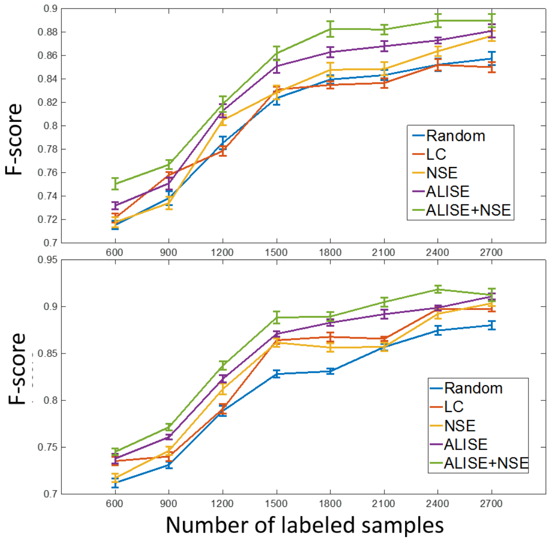

Figure 2: Slot filling F-score of different active learning approaches.

We follow the same implementation in [Liu and Lane, 2016] to use a bi-directional LSTM as the encoder network M in Fig.1. This bidirectional LSTM read the input sentence in both forward and backward directions and their hidden states at each step were concatenated as the long vector. We choose 128 for word embedding layer and 64 hidden states for the encoder LSTM. To this end, we have obtained 128 dimensions for the latent representation z. The decoder C in Fig.1 is implemented by either a standard LSTM decoder [Sutskever *et al.*, 2014] or a more advanced attention model [Liu and Lane, 2016]. Both of them are widely used in existing literatures. The adversarial network D is configured by three dense-connected layers with 128 (input layer), 64 (intermediate layer) and 1 (output layer) units, respectively. The output layer is further connected with a sigmoid function for probabilistic conversion. We use $relu$ activation among all other layers. Each token of the output sequence is coded as a one-hot vector with the hot entry indicating the underlying cateogory of the token. The whole deep learning system was trained by ADAM [Kingma and Ba, 2014]. Among all labeled training samples, we further randomly select 10% of them as validation samples. The whole training process is terminated when the loss on the validation set does not decrease or when the optimization reaches 100 epochs.

We consider comparing our ALISE approach with existing sequence-based active learning algorithms. The competitors include random sampling, least confidence score (see Eq.2), N-best sequence entropy (NSE, see Eq.4). Moreover, we further consider the combinational scoring approach as introduced in Section 3.2 that we combine both ALISE scores and NSE scores for query sample selection. To make fair comparisons, the number of optimal decoding parses (N) is chosen as five for both NSE approach and our ALISE+NSE approach. In active sequence learning, we randomly select 2130 sequences as testing samples. The remaining 3000 sequences are used for model training and active labeling.

In detail, among these 3000 data, $p = 300$ samples are randomly chosen as initial labeled data. Then, we train the ALISE model with these $p = 300$ samples and conduct active learning based on the remaining $3000 - p$ non-testing samples. $k = 300$ top samples returned by different active learning methods are selected for label query. After labeling, these k samples will be merged with the existing p labeled samples as the new labeled pool. The ALISE and other active learning models will be trained with this new labeled set and the trained model will be used to select another k unlabeled samples for the next round. Such query sample selection, labeling approach and model retraining processes will be iteratively conducted. We only report results until we have get 2700 training samples because it is the limitation for active selection. When all 3000 points are used, there are no distinctions among different active learning algorithms. The active learning results with different training sample sizes are reported in Fig.2, where both the LSTM and attention model are respectively used as the decoder C for sequence label prediction. In the figure, we do the random splitting process for five times and the average F-score with standard deviations are reported. Here, we choose the F-score as the accuracy indicator because it is widely used in existing works [Liu and Lane, 2016][Deoras and Sarikaya, 2013]

那为什么是从600开始的

From the results, we have observed that our ALISE model and its combinational extension (ALISE+NSE) both outperform existing sequence learning approaches. When the labeled number size is small, the improvements of two ALISE models are more significant. The ALISE+NSE model further improves the performances of ALISE. However, when the number of sample sizes is relatively large, the differences between ALISE and ALISE+NSE are minor. However, these two ALISE methods are still better than other sequence learning approaches. Meanwhile, we have observed that using attention model as the sequence decoder is much better than the LSTM model.

4.2 Image Captioning

We further apply ALISE model for the sequence generation task of image captioning. In this task, the input data is an image and the corresponding label is a caption sentence describing the content of the input image. We follow the same configuration and parameter settings in the work [Xu *et al.*, 2015] to implement the encoder-decoder learning framework. The structure of the adversarial discriminator in ALISE is kept the same as in the slot filling experiment. This part of active learning experiments are mainly conducted on MSCOCO dataset [Lin *et al.*, 2014], which consists of 82,783 images for training, 40,504 for validation, and 40,775 for testing. We noted that each image in MSCOCO dataset is paired with 5 ground truth captions. In our active learning setting, the query sample selection is mainly conducted at the image level. It means that if one image has been selected for labeling, its corresponding five ground-truth captions are all accessible. We follow Karpathy et al.[Karpathy and Fei-Fei, 2015] to preporcess the sentences, where all the words are converted to lower-case, and all non-alphanumeric characters are discards. We discarded all words that appear less than twice in all captions.

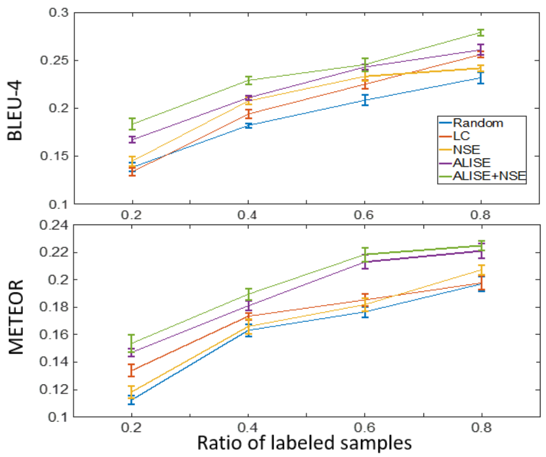

Figure 3: Image captioning results by active learning.

.

We consider all 82,783 training set as the basic data pool for active learning and query selection. We increase the labeled samples' rate from 0.2 to 0.8 with 0.2 as an incremental step. Among the first 0.2× 82,783 samples, half of them are randomly chosen as the initial labeled set and the remaining are selected by different active learning algorithms. The active selection and learning processes are iteratively conducted by adding k =0.2× 82,783 new labeled samples to the labeled pool in each round. These extra k samples are selected by different active learning algorithms. The performances of ALISE are compared with other active learning approaches in Fig.3. For result evaluations, we follow existing works to report BLEU-4 and METEOR as the accuracy indicator. These two accuracy measures can be easily calculated by the MSCOCO API. We repeat the aforementioned active learning process for 5 times with average and standard deviation reported in Fig.3. We have observed from quantitative evaluation that ALISE models (the original ALISE and ALISE+NSE) beat all existing active learning models based on these two scores. Meanwhile, the performance of ALISE can be further enhanced by combining NSE score as auxiliary indicator (ALISE+NSE).

To better understand differences among various active learning approaches, we provide some captioning results as intuitive instances in Fig.4. All these image captioning models are trained with 80% data points from the training set. Nevertheless, these same amount of training samples are selected by different active learning methods. In the figure, we provide the captioning results by NSE, ALISE and

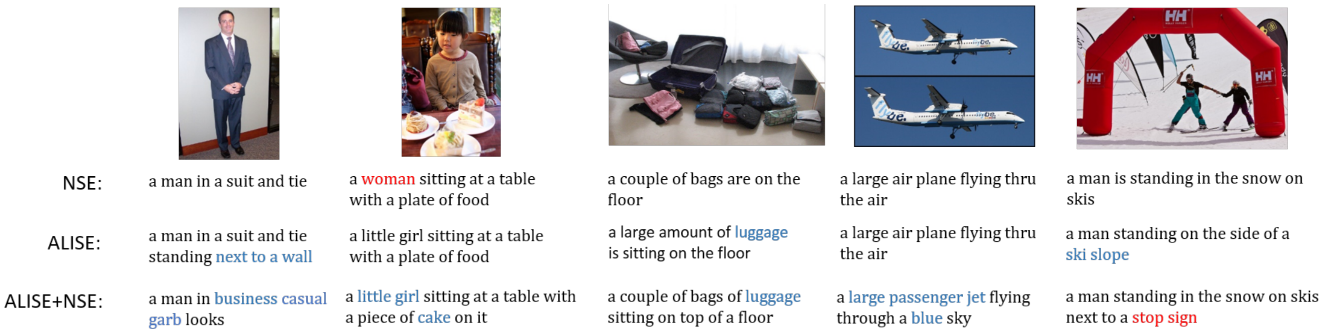
NSE:
a man in a suit and tie
a woman sitting at a table with a plate of food
a couple of bags are on the floor
a large air plane flying thru the air
a man is standing in the snow on skis
ALISE:
a man in a suit and tie standing next to a wall
a little girl sitting at a table with a plate of food
a large amount of luggage is sitting on the floor
a large air plane flying thru the air
a man standing on the side of a ski slope
ALISE+NSE:
a man in business casual garb looks
a little girl sitting at a table with a piece of cake on it
a couple of bags of luggage sitting on top of a floor
a large passenger jet flying through a blue sky
a man standing in the snow on skis next to a stop sign

Figure 4: Image captioning results in the active learning setting by ALISE, ALISE+NSE and NSE-based approaches. The novel plausible descriptions are annotated with blue color while wrong descriptions are colored in red.

.

ALISE+NSE because these three algorithms outperform others in Fig.3. From these intuitive results, we have observed that the two ALISE models tends to use more complex sentence structures for image descriptions. Moreover, these two ALISE models can cover more details about the visual information. It is mainly because the ALISE approach can actively build up a training set covering diverse information. However, there is still small chance that the ALISE model over-explains the image with wrongly recognized objects. As shown in the rightest sub-figure in Fig.4, ALISE+NSE model mistakenly describes the red finishing line as a stop sign. However, the captioning results of ALISE model are still much better than the NSE approach in producing precise captioning sentences in a more natural manner.

4.3 Computational Complexity

We also reported the computational costs of ALISE. We have observed that the computational costs of ALISE are almost the same as the original baseline model. In detail, we respectively report the training costs on the slot filling and image captioning tasks for references. The ALISE training in slot filling task (with 2700 samples) can be accomplished in just 74 seconds with 16 GPUs (Tesla K80) parallelized in optimization. The original Attention-based encoder-decoder slot-filling model costs 53 seconds when trained with the same amount of data [Liu and Lane, 2016]. For the image captioning task, the baseline attention model [Xu *et al.*, 2015] and ALISE respectively spends an average of 2.7 hours and 3.2 hours in total on 66,000 images. From these two tasks, the training cost of ALISE is not that different than the corresponding baseline encoder-decoder model. This is because ALISE has only introduced an auxiliary adversarial discriminator in the model and this discriminator neural network exhibits very simple structures (just a multi-layer neural network with a 64 nodes intermediate layer and a 1 node output layer.).

16块GPU？

However, the active learning complexity of different methods can vary significantly, especially when the candidate unlabeled pool size is large. We report the query sample selection costs on the aforementioned two datasets, that include 2,400 (i.e., the first data point in Fig.2) and 66,000 (i.e., the first data point in Fig.3) candidate samples on the slot filling

	Slot Filling	Captioning
LC	173s	2182s
NSE	245s	3956 s
ALISE	1.7s	6.9 s
ALISE+NSE	11.3s	67.4 s

Table 1: The active selection costs for different algorithms

and image captioning datasets, respectively. The correspond-ing costs of different algorithms are reported in ? omit the complexity of random sampling becau finished in real time.

感觉没这么离谱，对抗如果只是编码，不用解码，那么确实会减少时间，但是编码和解码的时间差别这么大吗

We have found that ALISE methods are much faster than existing sequence-based active learning approaches. This is because the calculation of the LC and NSE scores require the Viterbi parsing and beam search over the whole output space. Therefore, their costs are significant higher when the sample size is large (as in the image captioning dataset). However, the scoring mechanism in ALISE method just requires passing all samples through a trained neural network (i.e. the adversarial discriminator D in Fig.1). Therefore, the corresponding active scoring cost can be minor. The ALISE+NSE can also be efficiently implemented because it just performs N-best sequence entropy calculations on a selected number of samples filtered by ALISE model. Therefore, its computational costs are a bit higher than ALISE but are still far more less than other approaches.

5 Discussions

We introduced a sequence-based active learning model ALISE from the perspective of adversarial learning. It conducts query sample selections based on a well trained discriminator. Therefore, ALISE is much more efficient than existing predictor-dependent active learning approaches. Moreover, our model accomplishes both the tasks of active learning and sequence learning into a joint framework that is end-to-end trainable. Therefore, it is seamlessly applied to diverse learning tasks across different domains. Experimental verifications show that ALISE can greatly improve the performances and speed of existing models in the early active learning stages with insufficient training samples.

References

[Bao *et al.*, 2017] Feng Bao, Yue Deng, Mulong Du, Zhiquan Ren, Qingzhao Zhang, Yanyu Zhao, Jinli Suo, Zhengdong Zhang, Meilin Wang, and Qionghai Dai. Probabilistic natural mapping of gene-level tests for genome-wide association studies. *Briefings in bioinformatics*, page bbx002, 2017.

[Cohn *et al.*, 1994] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[Culotta and McCallum, 2005] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751, 2005.

[Deng *et al.*, 2013] Yue Deng, Qionghai Dai, Risheng Liu, Zengke Zhang, and Sanqing Hu. Low-rank structure learning via nonconvex heuristic recovery. *IEEE TNNLS*, 24(3):383–396, 2013.

[Deng *et al.*, 2016] Yue Deng, Feng Bao, Xuesong Deng, Ruiping Wang, Youyong Kong, and Qionghai Dai. Deep and structured robust information theoretic learning for image analysis. *IEEE TIP*, 25(9):4209–4221, 2016.

[Deng *et al.*, 2017a] Yue Deng, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE TNNLS*, 28(3):653–664, 2017.

[Deng *et al.*, 2017b] Yue Deng, Zhiquan Ren, Youyong Kong, Feng Bao, and Qionghai Dai. A hierarchical fused fuzzy deep neural network for data classification. *IEEE TFS*, 25(4):1006–1012, 2017.

[Deng *et al.*, 2017c] Yue Deng, Yilin Shen, and Hongxia Jin. Disguise adversarial networks for click-through rate prediction. In *IJCAI*, pages 1589–1595. AAAI Press, 2017.

[Deoras and Sarikaya, 2013] Anoop Deoras and Ruhi Sarikaya. Deep belief network based semantic taggers for spoken language understanding. In *Interspeech*, pages 2713–2717, 2013.

[Dong and Lapata, 2016] Li Dong and Mirella Lapata. Language to logical form with neural attention. *ACL*, 2016.

[Hemphill *et al.*, 1990] Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language Workshop*, 1990.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[Kim *et al.*, 2006] Seokhwan Kim, Yu Song, Kyungduk Kim, Jeong-Won Cha, and Gary Geunbae Lee. Mmr-based active machine learning for bio named entity recognition. In *NAACL*, pages 69–72. Association for Computational Linguistics, 2006.

[Kingma and Ba, 2014] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[Koehn, 2004] Philipp Koehn. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Association for Machine Translation in the Americas*, pages 115–124. Springer, 2004.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Liu and Lane, 2016] Bing Liu and Ian Lane. Attention-based recurrent neural network models for joint intent detection and slot filling. *Interspeech*, 2016.

[Luong *et al.*,] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation.

[Miller *et al.*, 2014] Brad Miller, Alex Kantchelian, Sadia Afroz, Rekha Bachwani, Edwin Dauber, Ling Huang, Michael Carl Tschantz, Anthony D Joseph, and J Doug Tygar. Adversarial active learning. In *Artificial Intelligent and Security Workshop*, pages 3–14. ACM, 2014.

[Scheffer *et al.*, 2001] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

[Settles and Craven, 2008] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1070–1079. Association for Computational Linguistics, 2008.

[Settles, 2010] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[Seung *et al.*, 1992] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[Sutton and McCallum, 2006] Charles Sutton and Andrew McCallum. *An introduction to conditional random fields for relational learning*, volume 2. Introduction to statistical relational learning. MIT Press, 2006.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[Zhu and Bento, 2017] Jia-Jie Zhu and Jose Bento. Generative adversarial active learning. *arXiv*, 2017.