Abductive Learning with Ground Knowledge Base *

Le-Wen Cai[1] , **Wang-Zhou Dai**[2] , **Yu-Xuan Huang**[1] ,
Yu-Feng Li[1] , **Stephen Muggleton**[2] , **Yuan Jiang**[1]
[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
{cailw, huangyx, liyf, jiangy}@lamda.nju.edu.cn
[2]Department of Computing, Imperial College London, London SW7 2AZ, UK
{w.dai, s.muggleton}@imperial.ac.uk

Abstract

Abductive Learning is a framework that combines machine learning with first-order logical reasoning. It allows machine learning models to exploit complex symbolic domain knowledge represented by first-order logic rules. However, it is challenging to obtain or express the ground-truth domain knowledge *explicitly* as first-order logic rules in many applications. The only accessible knowledge base is *implicitly* represented by groundings, i.e., propositions or atomic formulas without variables. This paper proposes *Grounded Abductive Learning* (GABL) to enhance machine learning models with abductive reasoning in a ground domain knowledge base, which offers inexact supervision through a set of logic propositions. We apply GABL on two weakly supervised learning problems and found that the model's initial accuracy plays a crucial role in learning. The results on a real-world OCR task show that GABL can significantly reduce the effort of data labeling than the compared methods.

1 Introduction

To address current limitations of data-driven machine learning, the next generation of Artificial Intelligence asks for a strong integration of machine learning with knowledge-driven reasoning such as logic inference [Bengio, 2017]. Recent years have witnessed a vast growth in this area, representative progress includes Neuro-Symbolic Learning (NeSy) [Garcez *et al.*, 2019] and Statistical Relational AI (StarAI) [Getoor and Taskar, 2007; Raedt *et al.*, 2016]. However, most of them are trying to build an end-to-end learning pipeline by subsuming logical calculus into differentiable modules in deep learning or statistical inference, in which first-order logical formulas are utilized as the basic relational topology for belief propagation and message passing.

Abductive Learning (ABL) [Zhou, 2019; Dai *et al.*, 2019] is a novel framework for combining machine learning with pure first-order logical reasoning in a mutually beneficial way. In ABL, the machine learning model learns to convert raw data into primitive logic facts serving as input to symbolic reasoning; while logical reasoning can infer the truth-value of the facts, which are named as pseudo-labels, for training the machine learning model. The integration of the two systems is realized by abduction, i.e., abductive reasoning, which can selectively infer particular predicted facts based on existing background knowledge [Magnani, 2009]. Therefore, ABL allows machine learning to utilize complex domain knowledge such as first-order logic theories [Dai *et al.*, 2019; Huang *et al.*, 2020].

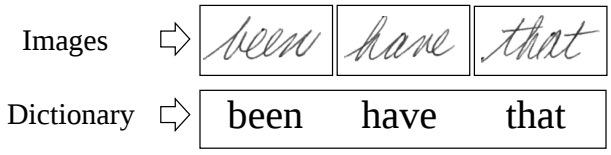


Figure 1: Example of the OCR Dictionary

Nevertheless, in many real-world applications, accessible knowledge bases only consist of a finite number of groundings (i.e., propositions or atomic logical formulas without variables). To give an example, for Optical Character Recognition (OCR) tasks, it is difficult to *explicitly* represent the underlying structure of words and characters with first-order theories, while the set of correct spellings can be easily obtained from a dictionary. As shown in figure 1, the dictionary of OCR can be represented as a *ground knowledge base* consisting of ground atoms of a predicate `valid_word(Y)`, such as `valid_word(['h', 'a', 'v', 'e'])`, etc.

For StarAI and NeSy systems, the lack of first-order logic formulas means there is no relational structure to establish the paths for belief propagation and message passing in probabilistic reasoning; for abduction-based approaches, the lack of logic clauses makes logical abduction impossible. A possible workaround is formulating this type of problem as multi-class learning [Li *et al.*, 2018], in which each ground atom (proposition) corresponds to a category. However, the lack of instances for each category brings class-imbalance issues [Japkowicz and Stephen, 2002], which makes machine learning even harder. Moreover, treating the groundings as independent categories neglects the *implicit* relational structure

*Our work is supported by the National Key Research and Development Program of China No.2020AAA0109400 and the National Natural Science Foundation of China (61772262). Yuan Jiang is the corresponding author.

among them (e.g., in English, the probability of "s" followed by "e" is much higher than that of "s" followed by "z").

This paper presents *Grounded Abductive Learning* (GABL) to solve the above deficiency, which allows machine learning models to exploit ground domain knowledge base within first-order logic context. Abduction in GABL is accomplished by augmenting the ground knowledge base with a default abductive logic program, which contains some general assumptions for abducing the pseudo-labels. For example, in OCR tasks, given an incorrect recognition result from an under-trained machine learning model, the augment abductive logic program could be "finding the word in the dictionary with the highest recognition confidence". The strong expressiveness power of first-order logic allows GABL to exploit various complex assumptions in different applications.

GABL provides a way to study in ABL, as the abductive reasoning in GABL is explicitly grounded. From the empirical study in a synthetic dataset, we find that the initial accuracy of the machine learning model is crucial for GABL. When model's prediction accuracy is higher than a certain threshold, GABL could improve model performance with unlabeled data. Furthermore, we verify the performance of GABL in a real-world weakly supervised OCR task. Results show that GABL can use unlabeled data and ground knowledge base to improve model performance and significantly decrease data labeling effort.

2 Related work

In recent years, many approaches have been proposed to deal with the lack of labeled data in machine learning. Semi-supervised learning is a powerful technique that attempts to exploit unlabeled data to improve model performance without human intervention. One category of semi-supervised learning methods related to this work is the proxy-label methods, which leverages the pre-trained model to produce pseudo-labels for unlabeled data based on some heuristics. The representatives of them are Self-training [Yarowsky, 1995] and Tri-training [Zhou and Li, 2010]. The self-training method predicts the label of input data, and then uses the predicted examples with probability higher than a pre-defined threshold or the top N confident predicted samples to retrain the model. Tri-training is a disagreement-based method based on ensemble, it uses diverse models to vote for the pseudo-labels for retraining the model.

This work is also related to multi-class learning. In multi-class classification, the model is required to classify instances into one of many categories. Error-Correcting Output Codes (ECOC) [Dietterich and Bakiri, 1994] is an ensemble method that transforms multi-class classification task into multi-label learning by encoding each class with an error-correcting code, which introduces sub-labels and can model a certain degree of label correlation. The performance of ECOC is highly related to the label encoding, which is difficult to construct without domain knowledge. Meanwhile, there are few studies about ECOC under the semi-supervised setting.

Some methods consider using symbolic domain knowledge to help model training. Neuro-Symbolic Learning (NeSy) [Garcez *et al.*, 2019] targets at combining machine learning with symbolic reasoning. It tries to integrate the ability to learn from the environment (for perception and pattern recognition) and reason from what has been learned (for reasoning and explanation). In most NeSy systems, learning and reasoning are both realized by a neural network, in which the external domain knowledge is used for building an explainable neural structure. Statistical Relational Learning (SRL) [Getoor and Taskar, 2007; Raedt *et al.*, 2016] shares the same motivation with NeSy, but it attempts to use domain knowledge to construct or initialize a probabilistic graphical model structure for statistical inference.

Abductive Learning (ABL) [Dai *et al.*, 2019; Zhou, 2019] is a framework that combines machine learning with pure first-order logical reasoning in a mutually beneficial way. ABL focuses on using first-order logic rules to revise model predicted labels and using these revised labels for training machine learning models. However, in order to perform abduction, a first-order abductive logic theory is required.

Unlike previous approaches, GABL exploits both unlabeled data and a ground knowledge base to improve model performance. It is different from NeSy, SRL, and ABL, which require first-order logic rules as domain knowledge.

3 Grounded Abductive Learning

This section presents problem setting and the Ground Abductive Learning (GABL) approach.

3.1 Problem Setting

The main target of this paper is to improve a pre-trained model, whose labeled training data are unavailable, with a set of unlabeled instances together with a ground knowledge base (GKB) that constrains the model's output space.

Formally, the input of the task contains a set of unlabeled training data $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ and a ground knowledge base GKB. Each $\boldsymbol{x}_i \in \mathcal{X}$ is corresponding to an unknown label $\boldsymbol{y}_i \in \mathcal{Y}$, where $\mathcal{X}$ is the feature space and $\mathcal{Y}$ is the label space. $GKB \subseteq \mathcal{Y}$ is a subset of the label space. In this paper, we consider classification problems, so $\mathcal{Y}$ is discrete and symbolic, i.e., each point $\boldsymbol{y}_i \in \mathcal{Y}$ can be considered as a ground atom or a proposition in Herbrand universe. For example, for the OCR task in figure 1, $\boldsymbol{x}$ is an image of a hand-written word, $\boldsymbol{y}$ is a string composed of the 26 English characters. As we can see, $\mathcal{Y}$ could be infinitely large (e.g., there could be an infinite number of possible strings). GKB is the set of ground atoms that lists all valid candidates for $\boldsymbol{y}$, e.g., the dictionary of correct spellings in English. Hence, for each $\boldsymbol{x}_i \in D$ with a corresponding $\boldsymbol{y}_i$, we have $\boldsymbol{y}_i \in GKB \subseteq \mathcal{Y}$. From the aspect of first-order logic, GKB is the answer set [Lifschitz, 2008] of an unknown first-order logic theory.

We denote the pre-trained machine learning model to be improved as $M : \mathcal{X} \mapsto \mathcal{Y}$. Given an input $\boldsymbol{x}_i \in D$, this model can output $\hat{\boldsymbol{y}}_i = M(\boldsymbol{x}_i)$ as its prediction.

3.2 Abductive Learning

In Abductive Learning (ABL), the machine learning model learns to convert raw data into primitive logic facts, which are regarded as pseudo-labels $\hat{\boldsymbol{y}}$ for logical reasoning. Meanwhile, abduction can selectively infer particular predicted

facts based on first-order logic rules. The inferred facts, which are regarded as abduced-labels $\bar{\boldsymbol{y}}$, will be utilized like ground-truth labels for training the machine learning model.

Abduction [Josephson and Josephson, 1996] is a basic form of logical inference for seeking the best explanation for observations based on implication. For example, when there is a formula "$wet\_ground \leftarrow rain$" (rain causes wet ground). When we observed the ground is wet, we could guess that it has rained.

Challenge It is challenging for GABL to realize abduction based on a ground knowledge base without any first-order logic rules. The ground knowledge base could only offer predicates such as "`valid_word(Y)`" to judge whether Y belongs to the knowledge base. However, abduction requires that there are some first-order logic rules for abduction. Thus, it is impossible to perform abduction when there is only a ground knowledge base.

3.3 Implementation of GABL

In order to perform abduction in a ground knowledge base, we propose to include an augmenting GKB with an abductive logic program that contains some very general assumptions that can constrain the search for $\bar{\boldsymbol{y}}$—the revised pseudo-labels for the predicted $\hat{\boldsymbol{y}}$ by M.

Considering the motivating OCR task that tries to map images to strings, we assume each label $\boldsymbol{y}_i$ as a sequence $[y_{i,1}, \ldots, y_{i,L_i}]$, where L_i is the length of sequence and $y_{i,l}$ is l-th *sub-label* of $\boldsymbol{y}_i$.

For this type of problem, we could include a general abductive program that uses GKB to constrain the search of pseudo-labels by string distance (e.g., edit distances). The program can be represented in first-order logic by following definitive clause:

$$\begin{aligned} program(\hat{\boldsymbol{y}}, \bar{\boldsymbol{y}}, GKB) \quad \leftarrow \quad & between(1, m, D) \\ \wedge \quad & distance(\hat{\boldsymbol{y}}, \bar{\boldsymbol{y}}, D) \\ \wedge \quad & \bar{\boldsymbol{y}} \in GKB \\ \wedge \quad & confidence(\bar{\boldsymbol{y}}, C) \\ \wedge \quad & C \geq threshold. \end{aligned} \tag{1}$$

Here m constrains the maximum allowed distance between model M's predicted-label $\hat{\boldsymbol{y}}$ and abduced-label $\bar{\boldsymbol{y}}$; $threshold$ is used to exclude results with low confidence; C is the confidence of $\bar{\boldsymbol{y}}$ that calculated by M. According to this first-order logic rule, GABL can automatically find out the $\bar{\boldsymbol{y}}$, which is close enough to $\hat{\boldsymbol{y}}$ and has high confidence. When there is more than one solution after abduction, GABL can include another rule to pick out the most confident one.

In fact, $distance(\hat{\boldsymbol{y}}, \bar{\boldsymbol{y}}, D)$ and $confidence(\bar{\boldsymbol{y}}, C)$ could be combined as that how similar is the model predicted result to groundings. We use the model training loss function to represent the similarity because model training loss function is carefully designed for learning task. In other words, when we consider the confidence in the abduction, we will directly select abduced-labels based on the loss function. Otherwise, we will select abduced-labels based only on the distance.

The Grounded Abductive Learning algorithm is described in algorithm 1. GABL will repeat E epochs and for every epoch, GABL uses model M to generate the result $\hat{\boldsymbol{r}}$ (labels with confidence) of input data $\boldsymbol{x}$. GABL selects abduced-labels based on Eq. (1). When the $\hat{\boldsymbol{y}}$ exits, GABL accepts it as training data and puts it into $\bar{D}$. At the end of the epoch, we use the training database $\bar{D}$ to update model M.

Algorithm 1 Grounded Abductive Learning

Input: Unlabeled Dataset $\mathcal{D}_u$, Pre-trained Model M, Ground Knowledge Base GKB, Augment Abductive Logic Program $\mathcal{P}$
Parameter: Epoch E
Output: Fine-tuned Model M

```
for e = 1 to E do
    D̄ = []
    for x ∈ D_u do
        r̂ = M(x)
        ȳ = abduce(r̂, GKB, P)
        if ȳ is not None then
            D̄.append((x, ȳ))
        end if
    end for
    Updating model M via D̄
end for
return M
```

4 Empirical Study

This section discusses why GABL can improve the machine learning model performance by leveraging unlabeled data and ground knowledge base. Firstly, we illustrate the mechanism of GABL through an intuitive example. Secondly, We construct experiments and aim to address: 1) how model accuracy would impact abduction learning when given domain knowledge; 2) How domain knowledge affects abductive learning.

4.1 Mechanism of GABL

Intuitively, GABL is similar to the classical self-training method [Yarowsky, 1995] for semi-supervised learning, which is a pseudo-label based method that uses model-predicted labels for further model training. Besides, GABL can exclude invalid pseudo-labels and even correct inaccurate pseudo-sub-labels, which could be more efficient for exploiting the unlabeled data than self-training methods. For the input data $\boldsymbol{x}$, when the model predicted label $\hat{\boldsymbol{y}}$ is inconsistent with the ground knowledge base, abduction needs to revise $\hat{\boldsymbol{y}}$ into $\bar{\boldsymbol{y}}$ that belongs to ground knowledge base. It assumes that only a few parts of the predicted label are incorrect. The pseudo-label $\hat{\boldsymbol{y}}$ might be distorted from a neighboring label $\bar{\boldsymbol{y}}$ that belongs to GKB. On the contrary, each $\hat{\boldsymbol{y}}$, which is in the neighborhood of $\bar{\boldsymbol{y}}$, should be revised to $\bar{\boldsymbol{y}}$.

Figure 2 better illustrates the mechanism of abduction in GABL. As shown in figure 2, every $\bar{\boldsymbol{y}}$ in GKB covers a part of the space like a Vonoroi diagram [Edelsbrunner and Seidel, 1986] under some special distance measurement (Hamming distance or other distance). Because of some disturbance, the predicted results float from points into their neighborhood, e.g., a ball surrounding it. The radius of balls depends on

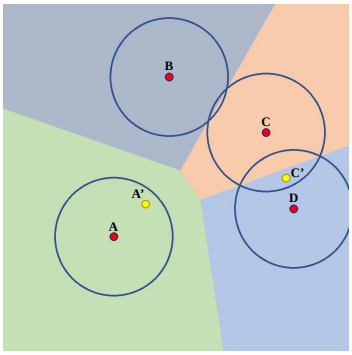


Figure 2: The labels space $\mathcal{Y}$ divided by neighborhoods of groundings in abduction. The red points in the center of circles are groundings in GKB. Space is divided into four parts according to the distance measure in the label space $\mathcal{Y}$. The predicted result (the yellow points) would appear in the blue circles with high probability. For example, an input data, whose ground truth label is point A, is predicted as A' and we can use abduction method to re-annotate point A' as A and fine-tune the model. However, when two points in GKB are too closed, or the model prediction error is too large, the abduced result could be wrong. For example, an input data, whose ground-truth label is point C, is predicted as C' and wrongly abduced as point D because of its distance to D is closer than C.

the accuracy of the pre-trained model, and higher accuracy leads to a smaller radius. When the model performance is ideal, almost all pseudo-labels fall into ground-truth covered space and can be classified correctly. When disturbance becomes more prominent, the radius of circles is bigger, more predicted results would be wrongly classified and may damage model performance.

We can use uncertainty to explain the above phenomenon. According to information theory [Cover, 1999], the uncertainty about the ground-truth label based on the pseudo-label can be decomposed into the entropy of the ground-truth label and the mutual information of the pseudo-label about the ground-truth label. When the task is given, the uncertainty of model prediction depends on the model's prediction accuracy. Therefore, the model's prediction accuracy plays a crucial role in abductive learning.

Moreover, GABL is a multi-epoch method. The model performance is boosted by repeatedly executing prediction and abduction. The model's accuracy after each epoch of training depends on the accuracy of the previous epoch. It can be seen that the initial accuracy of the model is crucial. Through some assumptions, we find that the accuracy threshold exists when given a domain knowledge base. First of all, we assume that after training, the generalization accuracy of the model is infinitely close to the accuracy of the data set. Second, we assume the higher the model's prediction accuracy leads to the higher the accuracy of the abduced result. Third, we assume that the prediction accuracy of all categories is almost the same. We note the accuracy of the label predicted by the model in epoch i as $\hat{p}_i^c$, and the accuracy of the sub-label after abduced is $\bar{p}_i^c$. In particular, $\hat{p}_0^c$ represents the sub-label prediction accuracy predicted by the initial model, and $\bar{p}_0^c$ represents the abduced sub-label accuracy predicted by the initial model. This means that if $\hat{p}_0^c < \bar{p}_0^c$, then $\hat{p}_0^c < \bar{p}_0^c \approx \hat{p}_1^c < \hat{p}_1^c ... \approx \hat{p}_E^c < \hat{p}_E^c$, where E means training epochs. But if $\hat{p}_0^c > \bar{p}_0^c$, then $\hat{p}_0^c > \bar{p}_0^c \approx \hat{p}_1^c > \hat{p}_1^c ... \approx \hat{p}_E^c > \hat{p}_E^c$. Therefore, there is a threshold p_{t1}. When $\hat{p}_0^c > p_{t1}$, the GABL can be help the model improve performance. At the same time, there is another threshold p_{t2}. When $\hat{p}_0^c < p_{t2}$ the GABL would hurt model performance, and the accuracy rate will gradually decrease. Therefore, there are accuracy thresholds p_{t1} and p_{t2}, and we note them as p_t for convenience.

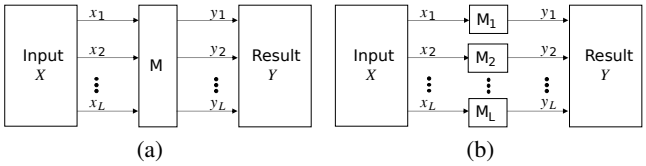


Figure 3: Sequence prediction setting. There are two settings when the length of the label is L. (a) Every sub-label has a unique machine learning model for classification; (b) All sub-labels share the same perception model.

Hamming	Equation	Random
0000000	0+0=0	0010110
0001011	0+1=1	0100100
...	...	...
1110100	98+1=99	1010100
1111111	99+0=99	1101011

Figure 4: Examples of GKBs.

4.2 Experiment on Synthetic Data

In this experiment, we verify whether there exists an initial accuracy threshold p_t that improves the model during abductive learning with the augment program in Eq. (1), and explore what would impact the accuracy threshold p_t based on the synthesized dataset. The code is available for download[1].

Does the accuracy threshold exist?

Dataset The dataset includes two parts, a ground knowledge base GKB represented by a set of groundings (as shown in figure 4)) and unlabeled training data $\mathcal{D}_u = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$. The groundings are generated by different domain knowledge base which includes hamming code of length 7 (experimental results note as "hamming-*") and decimal addition equation of length between 5 and 7 (experimental results note as "addition-*"). Every unlabeled data $\boldsymbol{x}_i$ contains $L_i d$ features, L_i represents the length of label $\boldsymbol{y}_i = [y_{i,1}, y_{i,2}, ..., y_{i,L_i}] \in GKB$, d represents every sub-label $y_{i,k}$ corresponds to d features, such as $[x_{i,(k-1)d+1}, ..., x_{i,kd}]$. $[x_{i,(k-1)d+1}, ..., x_{i,kd}]$ is sampled from basic data (MNIST images [LeCun *et al.*, 1995], CIFAR-10 images [Krizhevsky and Hinton, 2009] or synthetic data (as shown in figure 5)) according to sub-label $y_{i,k}$. Images of plus and equal signs are additionally added to MNIST images and CIFAR-10 images.

[1]https://github.com/AbductiveLearning/GABL

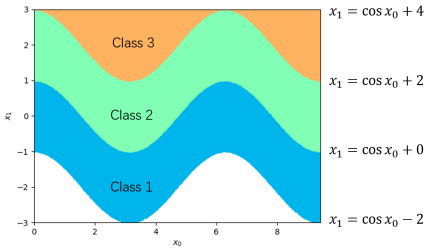


Figure 5: The distribution of synthetic data.

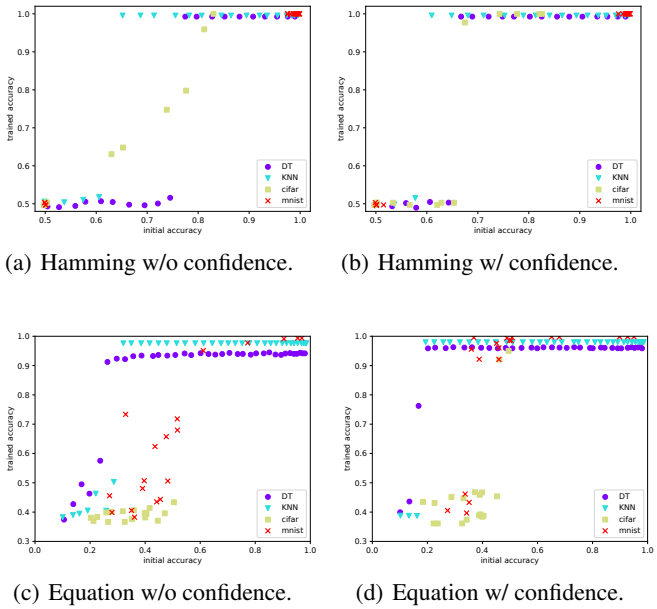
Experiment Setting Mimicking the perception-and-reasoning pipeline of NeSy and ABL models, we use one model for predicting the sub-labels and then feed them to GKB for reasoning. Specifically, as shown in figure 3(a), model M converts $[x_{i,(k-1)d+1}, ..., x_{i,kd}]$ into sub-label $y_{i,k}$. Ideally, the final result $\hat{\boldsymbol{y}}_i$, which includes $[\hat{y}_{i,1}, \hat{y}_{i,2}, ..., \hat{y}_{i,L}]$, belongs to GKB. When basic data are MNIST images (or CIFAR-10 images), we use CNN as perception model M and note experiment result as "mnist" (or "cifar"). When basic data are synthetic data, we use KNN (or decision tree) as perception model M and note experiment result as "KNN" (or "DT"). We control the model's initial accuracy through noisy data or controlling the number of pre-training data. Abduction considering model prediction confidence or not, are both tested. When abduction does not consider confidence, GABL rejects the sample with multiple solutions. Additionally, we set $k = 3$ in KNN and let each leaf node of the decision tree at least three samples in training.

Experimental Results Figure 6 shows there is an accuracy threshold in experiments. The experiments are conducted based on different GKB when models with different initial accuracy. As shown in figure 6, there are boundaries (accuracy threshold) in all experiments. When model initial prediction accuracy is high enough, GABL improves model performance via unlabeled data. It is worth noting that when we utilize model prediction confidence in abduction, the accuracy threshold is lower than the accuracy of abduction only based on model classification results.

How does domain knowledge affect the threshold?

Dataset & Setting We use random binary fixed-length code as GKB whose size N and code's length L can be controlled. The basic data are sampled from synthetic data. We use a decision tree as the base model and allow only one sample in the leaf node in training. Abduction does not consider model prediction confidence and only uses model classification results. As shown in figure 3(b), when the code length of GKB is L, we use L classifiers as the perception model and the k-th classifier response to predict k-th sub-label. It can avoid some special situations. For example, when there is only one model for predicting all sub-labels and $GKB = \{$"101","100"$\}$, the training process is almost supervised learning.

Experimental Results Figure 7 illustrates the relationship between accuracy threshold p_t and experiments' parameters (the length L of groundings and the size N of GKB). The result in figure 7 shows that when GKBs have same size N,



(a) Hamming w/o confidence. (b) Hamming w/ confidence.

(c) Equation w/o confidence. (d) Equation w/ confidence.

Figure 6: Trained accuracy of different model's initial accuracy experiments. (a) and (b) use hamming code as GKB. (c) and (d) use addition equation as GKB. (a) and (c) use classification result of sub-label in abduction. (b) and (d) consider confidence generated by the model in abduction.

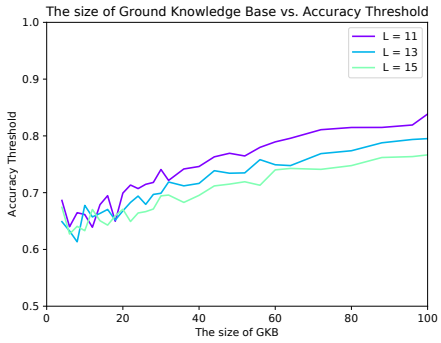


Figure 7: Accuracy threshold varies with different lengths of sub-labels and different sizes of the ground knowledge base (GKB). Every accuracy threshold p_t in the figure is the maximum value in 10 experiments in which the GKBs have the same label's length L and GKB's size N.

different label lengths have different threshold p_t. Moreover, the longer code length (number of sub-labels) requires lower threshold p_t. It is like information transmission, which uses longer code to overcome noise in the channel.

Summary In these experiments, we empirically verify that in Ground Abductive Learning (GABL), there is a threshold p_t, such that when the model accuracy is higher than p_t, GABL with the abductive program in Eq. (1) will improve model performance. Furthermore, the p_t is related to the sparsity of label space $\mathcal{Y}$, which is mainly influenced by the

Figure 8: A handwriting example in OCR task.

size N of GKB and the length L of the label. In short, a sparser space offers more tolerance of model errors and requires lower pre-trained model accuracy of task.

5 Optical Character Recognition Experiments

This section describes an experiment that applies GABL to a handwritten Optical Character Recognition (OCR) task. The experiment's main objective is to verify whether GABL can be applied in real-world applications with noisy input data.

Optical Character Recognition is an important application in the real world. For example, many handwriting archival materials are not transcribed into text. It is not friendly for amateurs to read and not easy for information retrieval. Therefore, it is meaningful to transcribe these handwritten documents into text. In practice, there exist two kinds of handwriting recognition tasks, lexicon-free and lexicon-based. Lexicon-based handwriting recognition offers a lexicon for model inference, which means that it should pick words in the lexicon.

Dataset We use IAM-database [Marti and Bunke, 2002] as the test benchmark. IAM-database contains 115,320 isolated word-level English handwriting images which are not pre-segmented. An example is shown in figure 8. Moreover, this is the first time that abductive learning is applied to tasks with unsegmented raw inputs. We only reserve words whose length is longer than 3 because there are too many short words, which causes the long-tail problem and is beyond this article's scope. We split the dataset into labeled data, unlabeled data, and test data in experiments. We leave 10% of the data for testing and randomly pick out different number data as labeled data. We collect all labels of the IAM database's images as the ground knowledge base (GKB).

Experimental Setting The setting in this experiment is like figure 3(a), where the same model predicts each sub-labels. We use CRNN [Shi *et al.*, 2017] as the basic machine learning model. During the prediction, the CRNN greedy selects the highest probability letters of each position and then merges the repeating letters. We use Burkhard-Keller-tree [Burkhard and Keller, 1973] (BK-tree) to select similar candidates in GKB and use edit distance to measure the similarity between candidates in BK-tree. At last, we pick the abduced pseudo-labels ranked by the CTC loss [Graves *et al.*, 2006]. We test our method in a semi-supervised setting. We use labeled data to train the model and then combine labeled data and abduced data for the training model.

We compare GABL with three types of semi-supervised baselines which all use CRNN as the basic model. 1) ST: Self-training methods [Yarowsky, 1995]; 2) Tri: Tri-training [Zhou and Li, 2010]; 3) VAT: Virtual Adversarial Training [Miyato *et al.*, 2019]. We also test CRNN's performance in fully supervised learning.

Table 1: Accuracy in handwriting experiments.

	5%	10%	15%	20%	25%	100%
CRNN	0.262	0.434	0.515	0.561	0.592	0.742
ST	0.461	0.572	0.636	0.660	0.677	-
Tri	0.222	0.484	0.588	0.638	0.647	-
VAT	0.301	0.476	0.567	0.594	0.627	-
GABL	**0.615**	**0.674**	**0.713**	**0.717**	**0.720**	-

Experimental Results We use the model's best performance on the test set as the experiment result. Because the performance of the comparison method will deteriorate rapidly as the number of training epochs increases, and it is difficult to determine the optimal performance through the number of training epochs. As shown in table 1, Grounded Abductive Learning has achieved the best performance in the handwritten Optical Character Recognition tasks. Although it is not a fair comparison, GABL utilizes a ground knowledge base to improve model performance and reduce the number of labeled data.

We also discover that insufficient unlabeled data could limit GABL's performance due to model overfitting on insufficient training data during the abduction process. Using all unlabeled data in one epoch may trap the model in a local optimum. When we subsample a batch of data for training in every epoch, models achieve better performance in the OCR.

The experiments are run on a single V100S GPU. GABL takes about 48 hours to train a CRNN model. In each epoch, GABL takes twice as much time as self-training. GABL is faster than the Tri-training and slower than the VAT.

6 Conclusion

This paper presents Ground Abductive Learning (GABL) to exploit the logical domain knowledge base represented by groundings. By augmenting the ground knowledge base with a program that exploits edit distance to abduce pseudo-labels, GABL can significantly outperform the compared supervised and semi-supervised learning approaches given the same amount of labeled data. Empirical study shows that the augment logic program can improve the performance of model when the accuracy of the pre-trained model exceeds a threshold. From the results of our experiments with synthetic data, we show that the threshold depends on the size of the ground knowledge base and the sparsity of the space covered by groundings. In general, a ground knowledge base can be regarded as an answer set of a first-order logic theory [Lifschitz, 2008]. Thus GABL is suitable for combining machine learning with any type of logic background knowledge.

Acknowledgment

The authors thank the Nanjing University-Imperial College London Machine Learning Joint Research Hub and the Office of International Cooperation & Exchanges of Nanjing University for their financial support.

References

[Bengio, 2017] Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

[Burkhard and Keller, 1973] Walter A. Burkhard and Robert M. Keller. Some approaches to best-match file searching. *Communications of the ACM*, 16(4):230–236, 1973.

[Cover, 1999] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[Dai *et al.*, 2019] Wang-Zhou Dai, Qiu-Ling Xu, Yang Yu, and Zhi-Hua Zhou. Bridging machine learning and logical reasoning by abductive learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 2811–2822, 2019.

[Dietterich and Bakiri, 1994] Thomas G Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1994.

[Edelsbrunner and Seidel, 1986] Herbert Edelsbrunner and Raimund Seidel. Voronoi diagrams and arrangements. *Discrete & Computational Geometry*, 1(1):25–44, 1986.

[Garcez *et al.*, 2019] Artur S. d'Avila Garcez, Marco Gori, Luís C. Lamb, Luciano Serafini, Michael Spranger, and Son N. Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Journal of Applied Logic*, 6(4):611–632, 2019.

[Getoor and Taskar, 2007] Lise. Getoor and Ben Taskar, editors. *Introduction to statistical relational learning*. MIT Press, Cambridge, Massachusetts, 2007.

[Graves *et al.*, 2006] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning (ICML)*, pages 369–376, 2006.

[Huang *et al.*, 2020] Yu-Xuan Huang, Wang-Zhou Dai, Jian Yang, Le-Wen Cai, Shaofen Cheng, Ruizhang Huang, Yu-Feng Li, and Zhi-Hua Zhou. Semi-supervised abductive learning and its application to theft judicial sentencing. In *Proceedings of the 20th IEEE International Conference on Data Mining (ICDM)*, pages 1070–1075, 2020.

[Japkowicz and Stephen, 2002] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

[Josephson and Josephson, 1996] John R Josephson and Susan G Josephson. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press, 1996.

[Krizhevsky and Hinton, 2009] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.

[LeCun *et al.*, 1995] Yann LeCun, Lawrence D Jackel, Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Urs A Muller, Eduard Sackinger, Patrice Simard, et al. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261(276):2, 1995.

[Li *et al.*, 2018] Jian Li, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, and Weiping Wang. Multi-class learning: From theory to algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1586–1595, 2018.

[Lifschitz, 2008] Vladimir Lifschitz. What is answer set programming? In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1594–1597, 2008.

[Magnani, 2009] Lorenzo Magnani. *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*. Springer, Berlin, 2009.

[Marti and Bunke, 2002] Urs-Viktor Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46, 2002.

[Miyato *et al.*, 2019] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019.

[Raedt *et al.*, 2016] Luc De Raedt, Kristian Kersting, Sriraam Natarajan, and David Poole. Statistical relational artificial intelligence: Logic, probability, and computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 10(2):1–189, 2016.

[Shi *et al.*, 2017] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2017.

[Yarowsky, 1995] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 189–196, 1995.

[Zhou and Li, 2010] Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.

[Zhou, 2019] Zhi-Hua Zhou. Abductive learning: towards bridging machine learning and logical reasoning. *Science China Information Sciences*, 62(7):76101, 2019.