EZLearn: Exploiting Organic Supervision in Automated Data Annotation

Maxim Grechkin[1], **Hoifung Poon**[2], **Bill Howe**[1]
[1] University of Washington
[2] Microsoft Research
grechkin@uw.edu, hoifung@microsoft.com, billhowe@uw.edu

Abstract

Many real-world applications require automated data annotation, such as identifying tissue origins based on gene expressions and classifying images into semantic categories. Annotation classes are often numerous and subject to changes over time, and annotating examples has become the major bottleneck for supervised learning methods. In science and other high-value domains, large repositories of data samples are often available, together with two sources of *organic supervision*: a lexicon for the annotation classes, and text descriptions that accompany some data samples. Distant supervision has emerged as a promising paradigm for exploiting such indirect supervision by automatically annotating examples where the text description contains a class mention in the lexicon. However, due to linguistic variations and ambiguities, such training data is inherently noisy, which limits the accuracy of this approach. In this paper, we introduce an auxiliary natural language processing system for the text modality, and incorporate co-training to reduce noise and augment signal in distant supervision. Without using any manually labeled data, our *EZLearn* system learned to accurately annotate data samples in functional genomics and scientific figure comprehension, substantially outperforming state-of-the-art supervised methods trained on tens of thousands of annotated examples.

1 Introduction

The confluence of technological advances and the open data movement [Molloy, 2011] has led to an explosion of publicly available datasets, heralding an era of data-driven hypothesis generation and discovery in high-value applications [Piwowar and Vision, 2013]. A prime example is *open science*, which promotes open access to scientific discourse and data to facilitate broad data reuse and scientific collaboration [Friesike *et al.*, 2015]. In addition to enabling reproducibility, this trend has the potential to accelerate scientific discovery, reduce the cost of research, and facilitate automation [Rung and Brazma, 2013; Libbrecht and Noble, 2015].

However, progress is hindered by the lack of consistent and high-quality annotations. For example, the NCBI Gene Expression Omnibus (GEO) [Clough and Barrett, 2016] contains over two million gene expression profiles, yet only a fraction of them have explicit annotations indicating the tissue from which the sample was drawn, information that is crucial to understanding cell differentiation and cancer [Hanahan and Weinberg, 2011; Gutierrez-Arcelus *et al.*, 2015]. As a result, only 20% of the datasets have ever been reused, and tissue-specific studies are still only performed at small scales [Piwowar and Vision, 2013].

Annotating data samples with standardized classes is the canonical multi-class classification problem, but standard supervised approaches are difficult to apply in these settings. Hiring experts to annotate examples for thousands of classes such as tissue types is unsustainable. Crowd-sourcing is generally not applicable, as annotation requires domain expertise that most crowd workers do not possess. Moreover, the annotation standard is often revised over time, incurring additional cost for labeling new examples.

While labeled data is expensive and difficult to create at scale, unlabeled data is usually in abundant supply. Many methods have been proposed to exploit it, but they typically still require labeled examples to initiate the process [Blum and Mitchell, 1998; McClosky and Charniak, 2008; Fei-Fei *et al.*, 2006]. Even zero-shot learning, where the name implies learning with no labeled examples for *some* classes, still requires labeled examples for related classes [Palatucci *et al.*, 2009; Socher *et al.*, 2013].

In this paper, we propose *EZLearn*, which makes annotation learning easy by exploiting two sources of *organic supervision*. First, the annotation classes generally come with a lexicon for standardized references (e.g., “liver”, “kidney”, “acute myeloid leukemia cell” for tissue types). While labeling individual data samples is expensive and time-consuming, it takes little effort for a domain expert to provide a few example terms for each class. In fact, in sciences and other high-value applications, such a lexicon is often available from an existing ontology. For example, the Brenda Tissue Ontology specifies 4931 human tissue types, each with a list of standard names [Gremse *et al.*, 2011]. Second, data samples are often accompanied by a free-text description, some of which directly or indirectly mention the relevant classes (e.g., the caption of a figure, or the description for a gene expression sample). Together with the lexicon, these descriptions present an opportunity for exploiting distant supervision by generating (noisy) labeled examples at scale [Mintz *et al.*, 2009]. We call such indirect supervision "organic" to emphasize that it is readily available as an integral part of a given domain.

In practice, however, there are serious challenges to enact this learning process. Descriptions are created for general human consumption, not as high-quality machine-readable annotations. They are provided voluntarily by data owners and lack consistency; ambiguity, typos, abbreviations, and non-standard references are common [Lee *et al.*, 2013; Rung and Brazma, 2013]. Multiple samples may share a text description that mentions several classes, introducing uncertainty as to which class label is associated with which sample. Additionally, annotation standards evolve over time, introducing new terms and evicting old ones. As a result, while there are potentially many data samples whose descriptions contain class information, only a fraction of them can be correctly labeled using distant supervision. This problem is particularly acute for domains with numerous classes and frequent updates, such as the life sciences.

To best exploit indirect supervision using all instances, *EZLearn* introduces an auxiliary text classifier for handling complex linguistic phenomena. This auxiliary classifier first uses the lexicon to find exact matches to teach the main classifier. In turn, the main classifier helps the auxiliary classifier improve by annotating additional examples with non-standard text mentions and correcting errors stemming from ambiguous mentions. This co-supervision continues until convergence. Effectively, *EZLearn* represents the first attempt in combining distant supervision and co-training, using text as the auxiliary modality for learning (Figure 1).

To investigate the effectiveness and generality of *EZLearn*, we applied it to two important applications: functional genomics and scientific figure comprehension, which differ substantially in sample input dimension and description length. In functional genomics, there are thousands of relevant classes. In scientific figure comprehension, prior work only considers three coarse classes, which we expand to twenty-four. In both scenarios, *EZLearn* successfully learned an accurate classifier with zero manually labeled examples.

While standard co-training has labeled examples from the beginning, *EZLearn* can only rely on distant supervision, which is inherently noisy. We investigate several ways to reconcile distant supervision with the trained classifier's predictions during co-training. We found that it generally helps to "remember" distant supervision while leaving room for correction, especially by accounting for the hierarchical relations among classes. We also conducted experiments to evaluate the impact of noise on *EZLearn*. The results show that *EZLearn* can withstand a large amount of simulated noise without suffering substantial loss in annotation accuracy.

2 Related Work

A perennial challenge in machine learning is to transcend the supervised paradigm by making use of unlabeled data. Standard unsupervised learning methods cluster data samples by explicitly or implicitly modeling similarity between them. It cannot be used directly for classification, as there is no direct relation between learned clusters and annotation classes.

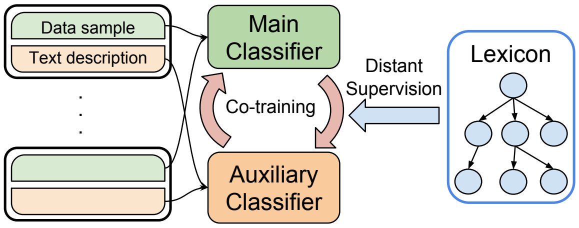


Figure 1: The *EZLearn* architecture: an auxiliary text-based classifier is introduced to bootstrap from the lexicon (often available from an ontology) and co-teaches the main classifier until convergence.

In semi-supervised learning, direct supervision is augmented by annotating unlabeled examples using either a learned model [Nigam and Ghani, 2000; Blum and Mitchell, 1998] or similarity between examples [Zhu and Ghahramani, 2002]. It is an effective paradigm to refine learned models, but still requires initialization with sufficient labeled examples for all classes. Zero-shot learning or few-shot learning relax the requirement of labeled examples for some classes, but still need to have sufficient labeled examples for *related* classes [Palatucci *et al.*, 2009; Socher *et al.*, 2013]. In this regard, they bear resemblance with domain adaptation [Blitzer *et al.*, 2007; Daumé III, 2007] and transfer learning [Pan and Yang, 2010; Raina *et al.*, 2007]. Zero-shot learning also faces additional challenges such as novelty detection to distinguish between known classes and new ones.

An alternative approach is to ask domain experts to provide example annotation functions, ranging from regular expressions [Hearst, 1991] to general programs [Ratner *et al.*, 2016]. Common challenges include combating low recall and semantic drifts. Moreover, producing useful annotation functions still requires domain expertise and substantial manual effort, and may be impossible when predictions depend on complex input patterns (e.g., gene expression profiles).

EZLearn leverages domain lexicons to annotate noisy examples from text, similar to distant supervision [Mintz *et al.*, 2009]. However, distant supervision is predominantly used in information extraction, which considers the single view on text [Quirk and Poon, 2017; Peng *et al.*, 2017]. In *EZLearn*, the text view is introduced to support the main annotation task, resembling co-training [Blum and Mitchell, 1998]. The original co-training algorithm annotates unlabeled examples in batches, where *EZLearn* relabels all examples in each iteration, similar to co-EM [Nigam and Ghani, 2000].

3 *EZLearn*

Let $X = \{x_i : i\}$ be the set of data samples and C be the set of classes. Automating data annotation involves learning a multi-class classifier $f : X \to C$. For example, x_i may be a vector of gene expression measurements for an individual, where C is the set of tissue types. Additionally, we denote t_i as the text description that accompanies x_i. If the description is not available, t_i is the empty string.

Algorithm 1 shows the *EZLearn* algorithm. By default, there are no available labeled examples (x, y^*) where $y^* \in C$ is the true class for annotating $x \in X$. Instead, *EZLearn* as-

Algorithm 1 *EZLearn*

```
Input: Data samples $X$, text descriptions $T$, annotation classes $C$, and lexicon $L$ containing example terms $L_c$ for each class $c \in C$.
Output: Trained classifiers $f : X \rightarrow C$ (main) and $f_T : T \rightarrow C$ (auxiliary).
Initialize: Generate initial training data $D^0$ as all $(x_i, t_i, c)$ for $x_i \in X, t_i \in T$, where $t_i$ mentions some term in $L_c$.
for $k = 1 : N_{iter}$ do
    $f \leftarrow \texttt{Train}_{\texttt{main}}(D^{k-1})$; $D_T^k \leftarrow \texttt{Resolve}(f(X), D^0)$
    $f_T \leftarrow \texttt{Train}_{\texttt{aux}}(D_T^k)$; $D^k \leftarrow \texttt{Resolve}(f_T(T), D^0)$
end for
```

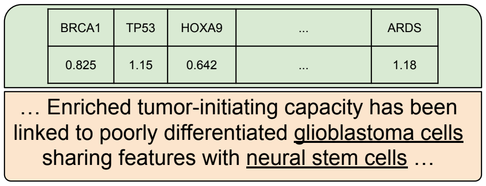
sumes that a lexicon L is available with a set of *example terms* L_c for each $c \in C$. We do not assume that L_c contains every possible synonym for c, nor that such terms are unambiguous. Rather, we simply require that L_c is non-empty for any c of interest. We use L_c's for distant supervision in *EZLearn*, by creating an initial labeled set D^0, which consists of all (x_i, t_i, c) where the text description t_i explicitly contains at least one term in L_c.

To handle linguistic variations and ambiguities, *EZLearn* introduces an auxiliary classifier $f_T : T \rightarrow C$, where $T = \{t_i : i\}$. At iteration k, we first train a new main classifier f^k using D^{k-1}. We then apply f^k to X and create a new labeled set D_T^k, which contains all (t_i, c) where $f^k(x_i) = c$. We then train a new text classifier f_T^k using D_T^k, and create the new labeled set D^k with all (x_i, c) where $f_T^k(t_i) = c$. This process continues until convergence, which is guaranteed given independence of the two views conditioned on the class label [Blum and Mitchell, 1998]. Empirically, it converges quickly.

We can use any classifier for $\texttt{Train}_{\texttt{main}}$ and $\texttt{Train}_{\texttt{aux}}$. Typically, the classifiers will take a parametric form (e.g., $f(x) = f(x, \theta)$) and training with a labeled set D amounts to minimize some loss function L (i.e., $\theta^* = \arg\min_\theta \sum_{(x,y^*) \in D} L(f(x, \theta), y^*)$). In this paper, we opted for simple, standard choices. In particular, we used logistic regression and fastText [Joulin *et al.*, 2017] as the main and auxiliary classifiers, respectively.

Generally, a classifier will output a score for each class rather than predicting a single class. The score reflects the confidence in predicting the given class. *EZLearn* generates the labeled set by adding all (sample, class) pairs for which the score crosses a hyperparameter threshold. We used 0.3 in this paper, which allows up to 3 classes to be assigned to a sample. The performance of *EZLearn* was not sensitive to this parameter: values in (0.2, 0.6) yielded similar results. In all iterations, a labeled set might contain more than one class for a sample, which is not a problem for the learning algorithm and is useful when there is uncertainty about the correct class.

For samples with distant-supervision labels, a classifier (main or auxiliary) might predict different labels in an iteration. Since distant supervision is noisy, reconciling it with the classifier's prediction could help correct its errors. The `Resolve`$(\cdot)$ function is introduced for this purpose. The direct analog of standard co-training returns distant-supervision labels if they are available (`Standard`). Conversely, `Resolve` could ignore distant supervision and always return the classifier's prediction (`Predict`). Alternatively, `Resolve` may return all labels (`Union`) or the common ones (`Intersect`).



BRCA1	TP53	HOXA9	...	ARDS
0.825	1.15	0.642	...	1.18

… Enriched tumor-initiating capacity has been linked to poorly differentiated glioblastoma cells sharing features with neural stem cells …

Figure 2: Example gene expression profile and its text description in Gene Expression Omnibus (GEO). Description is provided voluntarily and may contain ambiguous or incomplete class information.

However, none of the above approaches consider the hierarchical relations among the label classes. Suppose that the text mentions both `neuron` and `leukemia`, whereas the classifier predicts `leukocyte` with high confidence. Our confidence in `leukemia` being the correct label should increase since `leukemia` is a subtype of `leukocyte`, and our confidence in `neuron` should decrease. We thus propose a more sophisticated variant of `Resolve` that captures such reasoning (`Relation`). Let c_1, c_2 be the two labels from distant supervision and classifier prediction, respectively. If c_1 and c_2 are the same, `Relation` returns $c = c_1 = c_2$. If they have a hierarchical relation, `Relation` will return the more specific one (i.e., the subtype). Otherwise, `Relation` returns none. If distant supervision or the classifier prediction assigns multiple labels to a sample, `Relation` will return results from all label pairs. (In domains with no hierarchical relations among the classes, `Relation` is the same as `Intersect`.)

4 Application: Functional Genomics

Different tissues, from neurons to blood, share the same genome but differ in gene expression. Annotating gene expression data with tissue type is critical to enable data reuse for cell-development and cancer studies [Rung and Brazma, 2013]. Lee et al. manually annotated a large dataset of 14,510 expression samples to train a state-of-the-art supervised classifier [Lee *et al.*, 2013]. However, their dataset only covers 176 tissue types, or less than 4% of classes in BRENDA Tissue Ontology. In this section, we applied *EZLearn* to learn a far more accurate classifier that can in principle cover all tissue types in BRENDA. (In practice, the coverage is limited by the available unlabeled gene expression samples; in our experiments *EZLearn* learned to predict 601 tissue types.)

Annotation task The goal is to annotate gene expression samples with their tissue type. The input is a gene expression profile (a 20,000-dimension vector with a numeric value signifying the expression level for each gene). The output is a tissue type. We used the standard BRENDA Tissue Ontology [Gremse *et al.*, 2011], which contains 4931 human tissue types. For gene expression data, we used the Gene Expression Omnibus (GEO) [Edgar *et al.*, 2002], a popular repository run by the National Center for Biotechnology Information. Figure 2 shows an example gene expression profile with text description in GEO. We focused on the most common data-generation platform (Affymetrix U133 Plus 2.0), and obtained a dataset of 116,895 human samples. Each sample was

Method	# Labeled	# All	AUPRC	Prec@0.5	Use Expression	Use Text	Use Lexicon	Use EM
URSA	14510	0	0.40	0.52	yes	no	no	no
Co-EM	14510	116895	0.51	0.61	yes	yes	no	yes
Dist. Sup.	0	116895	0.59	0.63	yes	yes	yes	no
EZLearn	0	116895	**0.69**	**0.86**	yes	yes	yes	yes

Table 1: Comparison of test results between *EZLearn* and state-of-the-art supervised, semi-supervised, and distantly supervised methods on the CMHGP dataset. We reported the area under the precision-recall curve (AUPRC) and precision at 0.5 recall. *EZLearn* requires no manually labeled data, and substantially outperforms all other methods. Compared to URSA and co-EM, *EZLearn* can effectively leverage unlabeled data by exploiting organic supervision from text descriptions and lexicon. *EZLearn* amounts to initializing with distant supervision (first iteration) and continuing with an EM-like process as in co-training and co-EM, which leads to further significant gains.

processed using Universal exPression Codes (UPC) [Piccolo *et al.*, 2013] to minimize batch effects and normalize the expression values to [0,1]. Text descriptions were obtained from GEOmetadb [Zhu *et al.*, 2008].

Main classifier We implemented $\mathtt{Train}_{\mathtt{main}}$ using a deep denoising auto-encoder (DAE) with three LeakyReLU layers to convert the gene expression profile to a 128-dimensional vector [Vincent *et al.*, 2008], followed by multinomial logistic regression, trained end-to-end in Keras [Chollet, 2015], using L2 regularization with weight $1e-4$ and RMSProp optimizer [Tieleman and Hinton, 2012].

Auxiliary classifier We implemented $\mathtt{Train}_{\mathtt{aux}}$ using fastText with their recommended parameters (25 epochs and starting learning rate of 1.0) [Joulin *et al.*, 2017]. In principle, we can continue the alternating training steps until neither classifier's predictions change significantly. In practice, the algorithm converges quickly [Nigam and Ghani, 2000], and we simply ran all experiments with five iterations.

Systems We compared *EZLearn* with URSA [Lee *et al.*, 2013], the state-of-the-art supervised method that was trained on a large labeled dataset of 14,510 examples and used a sophisticated Bayesian method to refine SVM classification based on the tissue ontology. We also compared it with co-training [Blum and Mitchell, 1998] and co-EM [Nigam and Ghani, 2000], two representative methods for leveraging unlabeled data that also use an auxiliary view to support the main classification. Unlike *EZLearn*, they require labeled data to train their initial classifiers. After the first iteration, high-confidence predictions on the unlabeled data are added to the labeled examples. In co-training, once a unlabeled sample is added to the labeled set, it is not reconsidered again, whereas in co-EM, all of them are re-annotated in each iteration. We found that co-training and co-EM performed similarly, so we only report the co-EM results.

Evaluation The BRENDA Tissue Ontology is a directed acyclic graph (DAG), with nodes being tissue types and directed edges pointing from a parent tissue to a child, such as `leukocyte` → `leukemia cell`. We evaluated the classification results using *ontology-based precision and recall*. We expand each singleton class (predicted or true) into a set that includes all ancestors except the root. We then measure precision and recall as usual: precision is the proportion of correct predicted classes among all predicted classes, and recall is the proportion of correct predicted classes among true classes, with ancestors included in all cases. This metric closely resembles the approach by Verspoor et al. [Verspoor *et al.*, 2006], except that we are using the "micro" version (i.e., the predictions for all samples are first combined before measuring precision and recall). If the system predicts an irrelevant class in a different branch under the root, the intersection of the predicted and true sets is empty and the penalty is severe. If the predicted class is an ancestor (more general) or a descendent (more specific), the intersection is non-empty and the penalty is less severe, but overly general or overly specific predictions are penalized more than close neighbors. We tested on the Comprehensive Map of Human Gene Expression (CMHGP), the largest expression dataset with manual tissue annotations [Torrente *et al.*, 2016]. CMHPG used tissue types from the Experimental Factor Ontology (EFO) [Malone *et al.*, 2010], which can be mapped to the BRENDA Tissue Ontology. To make the comparison fair, 7,209 CMHGP samples that were in the supervised training set for URSA were excluded from the test set. The final test set contains 15,129 samples of 628 tissue types.

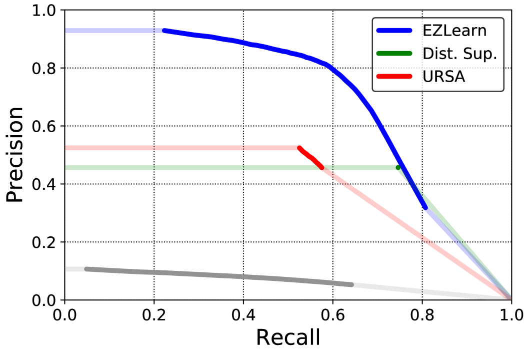
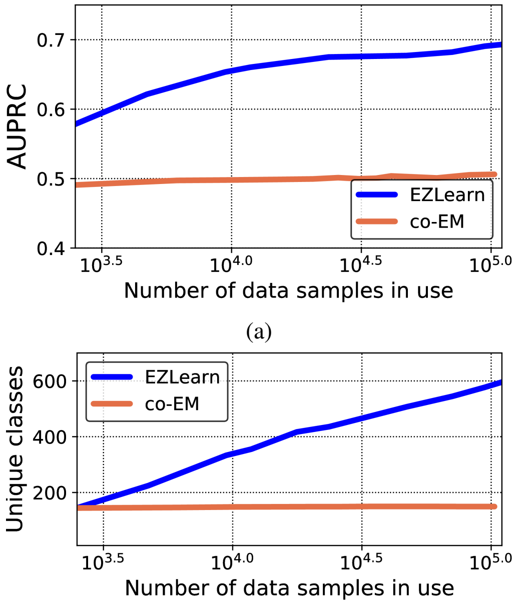


Figure 3: Ontology-based precision-recall curves comparing *EZLearn*, distant supervision, URSA, and the random baseline (gray). Extrapolated points are shown in transparent colors.

Results We report both the area under the precision-recall curve (AUPRC) and the precision at 0.5 recall. Table 1 shows the main classification results (with `Resolve` $=$ `Relation` in *EZLearn*). Remarkably, without using any manually labeled data, *EZLearn* outperformed the state-of-the-art supervised method by a wide margin, improving AUPRC by an absolute 27 points over URSA, and over 30 points in precision at 0.5 recall. Compared to co-EM, *EZLearn* improves AUPRC by 18 points and precision at 0.5 recall by 25 points. Figure 3 shows the precision-recall curves.

To investigate why *EZLearn* attained such a clear advantage even against co-EM, which used both labeled and unlabeled data and jointly trained an auxiliary text classifier, we compared their performance using varying amount of unlabeled data (averaged over fifteen runs). Figure 4(a) shows the



(a)

(b)

Figure 4: (a) Comparison of test accuracy with varying amount of unlabeled data, averaged over fifteen runs. *EZLearn* gained substantially with more data, whereas co-EM barely improves. (b) Comparison of number of unique classes in high-confidence predictions with varying amount of unlabeled data. *EZLearn*'s gain stems in large part from learning to annotate an increasing number of classes, by using organic supervision to generate noisy examples, whereas co-EM is confined to classes in its labeled data.

results. Note that the x-axis (number of unlabeled examples in use) is in log-scale. Co-EM barely improves with more unlabeled data, whereas *EZLearn* improves substantially from 2% to 100% of unlabeled data.

To understand why this is the case, we further compare the number of unique classes predicted by the two methods. See Figure 4(b). Co-EM is confined to the classes in its labeled data and its use of unlabeled data is limited to the extent of improving predictions for those classes. In contrast, by using organic supervision from the lexicon and text descriptions, *EZLearn* can expand the classes in its purview with more unlabeled data, in addition to improving predictive accuracy for individual classes. The gain seems to gradually taper off (Figure 4(a)), but we suspect that this is an artifact of the current test set. Although CMHGP is large, the number of tissue types in it (628) is still a fraction of that in the BRENDA Tissue Ontology (4931). Indeed, Figure 4(b) shows that the number of its predicted classes keeps climbing. This suggests that with additional unlabeled data *EZLearn* can improve even further, and with additional test classes, the advantage of *EZLearn* might become even larger.

We also evaluated on the subset of CMGHP with tissue types confined to those in the labeled data used by URSA and co-EM, to perfectly match their training conditions. Unsurprisingly, URSA and co-EM performed much better, attaining 0.53 and 0.67 in AUPRC, respectively (though URSA's accuracy is significantly lower than its training accuracy, suggesting overfitting). Remarkably, by exploiting organic supervision, *EZLearn* still outperformed both URSA and co-EM, attaining 0.71 in AUPROC in this setting.

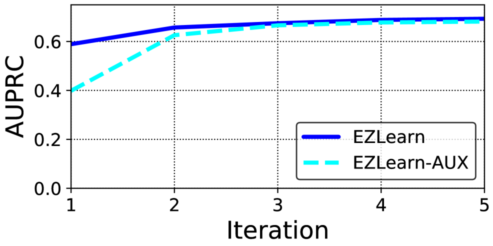


Figure 5: Comparison of test accuracy of the main and auxiliary classifiers at various iterations during learning.

Resolve	Stand.	Pred.	Union	Inter.	Relat.
# Classes	623	329	603	351	601
AUPRC	0.59	0.64	0.59	0.66	**0.69**

Table 2: Comparison of test results and numbers of unique classes in high-confidence predictions on the Comprehensive Map of Human Gene Expression by *EZLearn* with various strategies in resolving conflicts between distant supervision and classifier prediction.

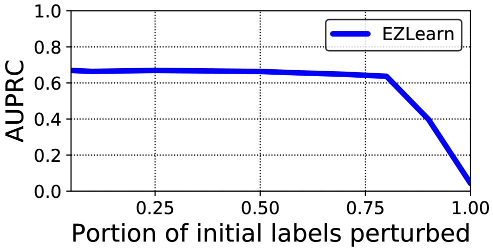


Figure 6: *EZLearn*'s test accuracy with varying portion of the distant-supervision labels replaced by random ones in the first iteration. *EZLearn* is remarkably robust to noise, with its accuracy only starting to deteriorate significantly after 80% of labels are perturbed.

EZLearn amounts to initializing with distant supervision (first iteration) and continuing with an EM-like process as in co-training and co-EM. This enables the main classifier and the auxiliary text classifier to improve each other during learning (Figure 5). Overall, compared to distant supervision, adding co-training led to further significant gains of 10 points in AUPRC and 23 points in precision at 0.5 recall (Table 1).

If labeled examples are available, *EZLearn* can simply add them to the labeled sets at each iteration. After incorporating the URSA labeled examples [Lee *et al.*, 2013], the AUPRC of *EZLearn* improved by two absolute points, with precision at 0.5 recall increasing to 0.87 (not shown in Table 1).

Compared to direct supervision, organic supervision is inherently noisy. Consequently, it is generally beneficial to reconcile classifier prediction with distant supervision when they are in conflict, as Table 2 shows. `Standard` (always choosing distant supervision when available) significantly trailed the alternative approach that always picks classifier's prediction (`Predict`). `Union` predicted more classes than `Intersect` but suffered large precision loss. By taking into account of hierarchical relations in the class ontology, `Relation` substantially outperformed all other methods in accuracy, while also covering a large number of classes.

To evaluate *EZLearn*'s robustness, we simulated noise by replacing a portion of the initial distant-supervision labels

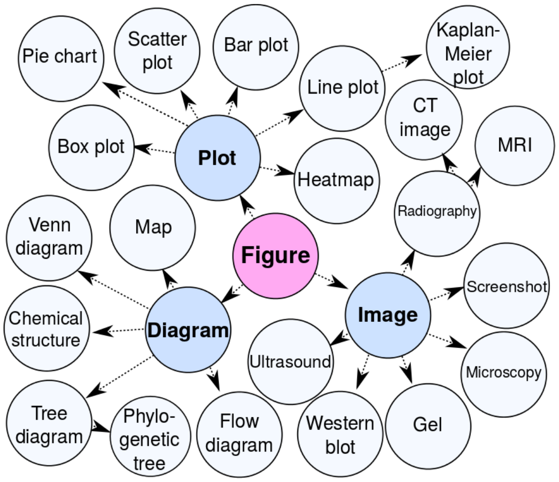


Figure 7: The Viziometrics project only considers three coarse classes `Plot`, `Diagram`, and `Image` for figures due to high labeling cost. We expanded them into 24 classes, which *EZLearn* learned to accurately predict with zero manually labeled examples.

with random ones. Figure 6 shows the results. Interestingly, *EZLearn* can withstand a significant amount of label perturbation: test performance only deteriorates drastically when more than 80% of initial labels are replaced by random ones. This result suggests that *EZLearn* can still perform well for applications with far more noise in their organic supervision.

5 Application: Figure Comprehension

Figures in scientific papers communicate key results and provide visual explanations of complex concepts. However, while text understanding has been intensely studied, figures have received much less attention in the past. A notable exception is the Viziometrics project [Lee *et al.*, 2017], which annotated a large number of examples for classifying scientific figures. Due to the considerable cost of labeling examples, they only used five coarse classes: `Plot`, `Diagram`, `Image`, `Table` and `Equation`. We exclude the last two as they do not represent true figures. In practice, figure-comprehension projects would be much more useful if they include larger set of specialized figure types. To explore this direction, we devised an ontology where `Plot`, `Diagram`, and `Image` are further refined into a total of twenty-four classes, such as `Boxplot`, `MRI` and `PieChart` (Figure 7). *EZLearn* naturally accommodates a large and dynamic ontology since no manually labeled data is required.

Annotation task The goal is to annotate figures with semantic types shown in Figure 7. The input is the image of a figure with varying size. The output is the semantic type. We obtained the data from the Viziometrics project [Lee *et al.*, 2017] through its open API. For simplicity, we focused on the non-composite subset comprising single-pane figures, yielding 1,174,456 figures along with free-text captions for use as distant supervision. As in the gene expression case, captions might be empty or missing.

System Each figure image was first resized and converted to a 2048-dimensional real-valued vector using a convolutional neural network [He *et al.*, 2016] trained on ImageNet [Deng *et al.*, 2009]. We follow [Howe *et al.*, 2017] and use the ResNet-50 model with pre-trained weights provided by Keras [Chollet, 2015]. We used the same classifiers and hyperparameters as in the functional genomics application. We used a lexicon that simply comprises of the names of the new classes, and compared *EZLearn* with the Viziometrics classifier. We also compared with a lexicon-informed baseline that annotates a figure with the most specific class whose name is mentioned in the caption (or root otherwise).

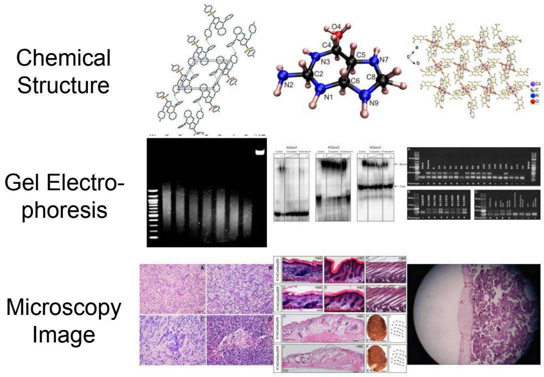


Figure 8: Example annotations by *EZLearn*, all chosen among figures with no class information in their captions.

Evaluation We followed the functional genomics application and evaluated on ontology-based precision and recall. Since the new classes are direct refinement of the old ones, we can also evaluate the Viziometrics classifier using this metric. To the best of our knowledge, there is no prior dataset or evaluation for figure annotation with fine-grained semantic classes as in Figure 7. Therefore, we manually annotated an independent test set of 500 examples.

	Lexicon	Vizio.	Dist. Sup.	*EZLearn*
AUPRC	0.44	0.53	0.75	**0.79**
Prec@0.5	0.31	0.43	0.87	**0.92**

Table 3: Comparison of test results between *EZLearn*, the lexicon baseline, the Viziometrics classifier, and distant supervision.

Results *EZLearn* substantially outperformed both the lexicon-informed baseline and the Viziometrics classifier (Table 3). The state-of-the-art Viziometrics classifier was trained on 3271 labeled examples, and attained an accuracy of 92% on the coarse classes. So the gain attained by *EZLearn* reflects its ability to extract a large amount of fine-grained semantic information missing in the coarse classes. Figure 8 shows example annotations by *EZLearn*, all chosen from figures with no class mention in the caption.

6 Conclusion

We propose *EZLearn* for automated data annotation, by combining distant supervision and co-training. *EZLearn* is well suited to high-value domains with numerous classes and frequent update. Experiments in functional genomics and scientific figure comprehension show that *EZLearn* is broadly applicable, robust to noise, and capable of learning accurate classifier without manually labeled data, even outperforming state-of-the-art supervised systems by a wide margin.

References

[Blitzer *et al.*, 2007] John Blitzer, Mark Dredze, Fernando Pereira, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[Chollet, 2015] François Chollet. Keras. https://github.com/fchollet/keras, 2015.

[Clough and Barrett, 2016] Emily Clough and Tanya Barrett. The gene expression omnibus database. *Methods Mol Biol*, 2016.

[Daumé III, 2007] Hal Daumé III. Frustratingly easy domain adaptation. *ACL 2007*, 2007.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*, pages 248–255. IEEE, 2009.

[Edgar *et al.*, 2002] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *NAR*, 2002.

[Fei-Fei *et al.*, 2006] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE TPAMI*, 2006.

[Friesike *et al.*, 2015] Sascha Friesike, Bastian Widenmayer, et al. Opening science: towards an agenda of open science in academia and industry. *J. of Tech. Transfer*, 2015.

[Gremse *et al.*, 2011] Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Research*, 2011.

[Gutierrez-Arcelus *et al.*, 2015] Maria Gutierrez-Arcelus, Halit Ongen, Tuuli Lappalainen, et al. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet*, 2015.

[Hanahan and Weinberg, 2011] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hearst, 1991] Marti Hearst. Noun homograph disambiguation using local context in large text corpora. *Using Corpora*, pages 185–188, 1991.

[Howe *et al.*, 2017] Bill Howe, Po-shen Lee, Maxim Grechkin, Sean T Yang, and Jevin D West. Deep mapping of the visual literature. In *WWW Companion*, pages 1273–1277, 2017.

[Joulin *et al.*, 2017] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *EACL 2017*, page 427, 2017.

[Lee *et al.*, 2013] Young-suk Lee, Arjun Krishnan, Qian Zhu, and Olga G. Troyanskaya. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics*, 29(23):3036–3044, 2013.

[Lee *et al.*, 2017] Po-shen Lee, Jevin D West, and Bill Howe. Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data*, 2017.

[Libbrecht and Noble, 2015] Maxwell W. Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nat. Rev. Genetics*, 2015.

[Malone *et al.*, 2010] James Malone, Ele Holloway, Tomasz Adamusiak, et al. Modeling sample variables with an experimental factor ontology. *Bioninf.*, 2010.

[McClosky and Charniak, 2008] David McClosky and Eugene Charniak. Self-training for biomedical parsing. In *ACL*, 2008.

[Mintz *et al.*, 2009] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL*, pages 1003–1011, 2009.

[Molloy, 2011] Jennifer C Molloy. The open knowledge foundation: open data means better science. *PLoS Biol*, 2011.

[Nigam and Ghani, 2000] Kamal Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *CIKM*, 2000.

[Palatucci *et al.*, 2009] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2010.

[Peng *et al.*, 2017] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. Cross-sentence n-ary relation extraction with graph LSTMs. *TACL*, 5:101–115, 2017.

[Piccolo *et al.*, 2013] Stephen Piccolo, Michelle Withers, Owen Francis, Andrea Bild, and Evan Johnson. Multiplatform single-sample estimates of transcriptional activation. *PNAS*, 2013.

[Piwowar and Vision, 2013] Heather Piwowar and Todd Vision. Data reuse and the open data citation advantage. *PeerJ*, 2013.

[Quirk and Poon, 2017] Chris Quirk and Hoifung Poon. Distant supervision for relation extraction beyond the sentence boundary. *EACL-2017*, 2017.

[Raina *et al.*, 2007] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML*, 2007.

[Ratner *et al.*, 2016] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data Programming: Creating large training sets, quickly. In *NIPS*, 2016.

[Rung and Brazma, 2013] Johan Rung and Alvis Brazma. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, 2013.

[Socher *et al.*, 2013] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[Tieleman and Hinton, 2012] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *Coursera: NN for ML*, 2012.

[Torrente *et al.*, 2016] Aurora Torrente, Margus Lukk, et al. Identification of cancer related genes using a comprehensive map of human gene expression. *PLOS ONE*, 2016.

[Verspoor *et al.*, 2006] Karin Verspoor, Judith Cohn, Susan Mniszewski, and Cliff Joslyn. A categorization approach to automated ontological function annotation. *Prot. Sc.*, 2006.

[Vincent *et al.*, 2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.

[Zhu and Ghahramani, 2002] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. *Technical Report CMU-CALD-02-107*, 2002.

[Zhu *et al.*, 2008] Yuelin Zhu, Sean Davis, Robert Stephens, Paul S. Meltzer, and Yidong Chen. GEOmetadb: powerful alternative search engine for the gene expression omnibus. *Bioinf.*, 2008.