International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Faster Model Improvements through Weakly Supervised Labels

Ashish Bansal

Submitted:15/03/2024 **Revised**: 27/04/2024 **Accepted**: 04/05/2024

Abstract*:* Deep neural networks are becoming omnipresent in natural language applications (NLP). How- ever, they require large amounts of labeled training data, which is often only available for English. This is a big challenge for many languages and domains where labeled data is limited. In recent years, a variety of methods have been proposed to tackle this situation. This paper gives an overview of these approaches that help you train NLP models in resource-lean scenarios. This includes both ideas to increase the amount of labeled data as well as methods following the popular pre-train and fine-tune paradigm.

Supervised learning techniques construct predictive models by learning from a large number of training examples, where each training example has a *label* indicating its ground-truth output. Though current techniques have achieved great success, it is noteworthy that in many tasks it is difficult to get strong supervision information like fully ground-truth labels due to the high cost of the data-labeling process. Thus, it is desirable for machine-learning techniques to work with weak supervision.

This paper outlines the advantages of *weakly supervised learning* in collecting more robust data fastly and using less resource, focusing on three typical types of weak supervision: incomplete supervision, where only a subset of training data is given with labels; inexact supervision, where the training data are given with only coarse-grained labels; and inaccurate supervision, where the given labels are not always ground-truth.

The main focus will be on the weak supervision technique where we will explain how a smaller dataset is used to train a classifier model and then that model is used to label the new data having weak labels which might be accurately predicting those labels to some extent. This method in- volves human-in-loop where human would reviews those predicted labels and correct the wrong predictions which create an additional data points to train a new weak labeler model. Using this technique iteratively it helped researchers in creating more ground truth data that can be used to train better performing models very fast.

Key Word: *machine learning, weakly supervised learning, supervised learning, NLP,*

Introduction

There is a catch to training state-of-the-art NLP models: their reliance on massive hand-labeled training sets. That's why data labeling is usually the bottleneck in developing NLP applications and keeping them up-to-date. For example, imagine how much it would cost to pay medical specialists to label thousands of electronic health records. In general, having domain experts label thousands of examples is too expensive. To solve the issue of limited domain specific labelled training data weak supervision labeler are utilized to collect more accurate and robust training data.

Getting labeled training data has become *the* key development bottleneck in supervised ma- chine learning. We provide a broad, high-level overview of recent weak supervision approaches, where noisier or higher-level supervision is used as a more expedient and flexible way to get su- pervision signal, in particular from subject matter experts (SMEs). We provide a simple, broad definition of weak supervision as being comprised of one or more noisy conditional distributions over unlabeled data, and focus on the key technical challenge of unifying and modeling these sources. In recent years, the real-world impact of machine learning has grown in leaps and bounds. In large part, this is due to the advent of deep learning models, which allow practitioners to get state- of-the-art scores on benchmark datasets without any hand-engineered features. The reliance of these models on massive sets of hand-labeled training data. This dependence of machine learning on labeled training sets is nothing new, and arguably has been the primary driver of new advances for many years. But deep learning models are massively more complex than most traditional mod- els–many standard deep learning models today have hundreds of millions of free parameters–and thus require commensurately more labeled training data.

These hand-labeled training sets are expensive and time-consuming to create–taking months or years for large benchmark sets, or when domain expertise is required–and cannot be practically repurposed for new objectives. In practice, the cost and inflexibility of hand-labeling such training sets is the key bottleneck to actually deploying machine learning.

That's why in practice today, most large ML systems actually use some form of weak supervi- sion: noisier, lower-quality, but larger-scale training sets constructed

NYU

ab6995@nyu.edu

via strategies such as using cheaper annotators, programmatic scripts, or more creative and high-level input from domain ex- perts, to name a few common techniques.

0.1 Weak Supervision

Weak Supervision helps us alleviate the **data bottleneck** problem by enabling us to cheaply leverage subject matter expertise to pro-grammatically label millions of data points. More specif- ically, it's a framework that helps subject matter experts (SMEs) infuse their knowledge into an AI system in the form of hand-written heuristic rules or distant supervision. In the weak supervision setting, our objective is the same as in the supervised setting, however instead of a ground-truth labeled training set we have:

- Unlabeled data \(X_u = x_1, . . . , x_N\);

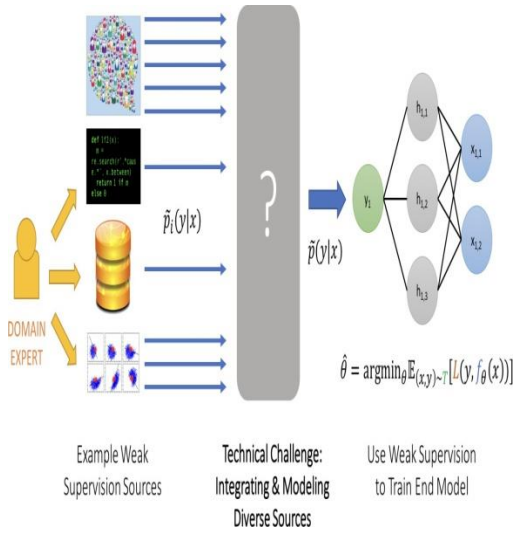

Figure 1: A high-level schematic of the basic weak supervision "pipeline": We start with one or more weak supervision sources: for example crowdsourced data, heuristic rules, distant supervision, and/or weak classifiers provided by an SME. The core technical challenge is to unify and model these disparate sources, which we discuss in the next section. Then, this must be used to train the end model–in the standard ERM context, we can imagine changing either the training set *T*, loss function *L*, or model *f* to accomplish this.

- One or more weak supervision sources \(\tilde{p}i(y \vert x), i=1:M\) provided by a human subject matter expert (SME), such that each one has:
  - A coverage set \(C_i\), which is the set of points \(x\) over which it is defined
  - An accuracy, defined as the expected probability of the true label \(y*^\) over its cover- age set, which we assume is \(\lt 1.0\)

In general, we are motivated by the setting where these weak label distributions serve as a way for human supervision to be provided more cheaply and efficiently: either by providing higher- level, less precise supervision (e.g. heuristic rules, expected label distributions), cheaper, lower- quality supervision (e.g. crowd-sourcing), or taking opportunistic advantage of existing resources (e.g. knowledge bases, pre-trained models). These weak label distributions could thus take one of many well-explored forms:

- Weak Labels: The weak label distributions could be deterministic functions–in other words, we might just have a set of noisy labels for each data point in \(C_i\). These could come from crowd workers, be the output of heuristic rules \(f_i(x)\), or the result of distant supervision (Mintz et al. 2009), where an external knowledge base is heuristically mapped onto \(X_u\). These could also be the output of other classifiers which only yield MAP estimates, or which are combined with heuristic rules to output discrete labels.

- Constraints: We can also consider constraints represented as weak label distributions. Though straying outside of the simple categorical setting we are considering here, the structured prediction setting leads to a wide range of very interesting constraint types, such as physics- based constraints on output trajectories (Stewart and Ermon 2017) or output constraints on execution of logical forms (Clarke et al. 2010; Guu et al. 2017), which encode various forms of domain expertise and/or cheaper supervision from e.g. lay annotators.

- Distributions: We might also have direct access to a probability distribution. For example, we could have the posterior distributions of one or more weak (i.e. low accuracy/coverage) or biased classifiers, such as classifiers trained on different data distributions as in the transfer learning setting. We could also have one or more user-provided label or feature expecta- tions or measurements (Mann and McCallum 2010; Liang, Jordan, and Klein 2009), i.e. an expected distribution \(p_i(y)\) or \(p_i(y\vert f(x))\) (where \(f(x)\) is some feature of \(x\)) provided by a domain expert as in e.g. (Druck, Settles, and McCallum 2009).

- In-variances: Finally, given a small set of labeled data, we can express functional in-variances as weak label distributions–e.g., extend the coverage of the labeled distribution to all trans- formations of \(t(x)\) or \(x\), and set \(p_i(y\vert t(x)) = p_i(y\vert x)\). In this way we view techniques such as data augmentation as a form of weak supervision as well.

Given a potentially heterogeneous set of such weak supervision sources, we can conceptu- ally break the technical challenges of weak supervision into two components. First, we need to deal with the fact that our weak sources are noisy and conflicting–we view this as the core lurking technical challenge of weak supervision, and discuss it more further on. Second, we need to then modify the traditional empirical risk minimization (ERM) framework to accept our weak supervi- sion.

Realted Work

Weak supervision aims to replace hand-annotated 'ground truths' with labelling functions that are pro- grammatically applied to data points – in our case, texts – from the target domain (Ratner et al., 2017, 2019; Lison et al., 2020; Safranchik et al., 2020b; Fu et al., 2020). Those functions may take the form of rule-based heuristics, gazetteers, annotations from crowd-workers, external databases, data-driven models trained from related domains, or linguistic constraints. A particular form of weak supervision is distant supervision, which relies on knowledge bases to automatically label documents with entities (Mintz et al., 2009; Ritter et al., 2013; Shang et al., 2018). Weak supervision is also related to models for aggregating crowd-sourced annotations (Kim and Ghahra- mani, 2012; Hovy et al., 2013; Nguyen et al., 2017).

Crucially, labelling functions do not need to provide a prediction for every data point and may "abstain" whenever certain conditions are not met. They may also rely on external data sources that are unavailable at runtime, as is the case for labels obtained by crowd-workers. After being applied to a dataset, the results of those labelling functions are aggregated into a single, probabilis- tic annotation layer. This aggregation is often implemented with a generative model connecting the latent (unobserved) labels to the outputs of each labelling function (Ratner et al., 2017; Lison et al., 2020; Safranchik et al., 2020a). Based on those aggregated labels, a discriminative model (often a neural architecture) is then trained for the task.

Weak supervision shifts the focus away from collecting manual annotations and concentrates the effort on developing good labelling functions for the target domain. This approach has been shown to be much more efficient than traditional annotation efforts (Ratner et al., 2017). Weak supervision allows domain experts to directly inject their domain knowledge in the form of various heuristics. Another benefit is the possibility to modify/extend the label set during development, which is a common situation in industrial R&D projects.

Several software frameworks for weak supervision have been released in recent years. One such framework is Snorkel (Ratner et al., 2017, 2019) which combines various supervision sources using a generative model. However, Snorkel requires data points to be independent, making it difficult to apply to sequence labelling tasks as done in skweak. Swellshark (Fries et al., 2017) is another framework optimized for biomedical NER. Swellshark, is however, limited to classifying already segmented entities, and relies on a separate, ad-hoc mechanism to generate candidate spans.

FlyingSquid (Fu et al., 2020) presents a novel approach based on triplet methods, which is shown to be fast enough to be applicable to structured prediction problems such as sequence labelling. However, compared to skweak, the aggregation model of FlyingSquid focuses on esti- mating the accuracies of each labelling function, and is therefore difficult to apply to problems where labelling sources may exhibit very different precision/recall trade-offs. A labelling function may for instance rely on a pattern that has a high precision but a low recall, while the opposite may be true for other labelling functions. Such difference is lost if accuracy is the only metric as- sociated for each labelling function. FinallySafranchik et al. (2020b) describe a weak supervision model based on an extension of HMMs called linked hidden Markov models. Although their ag- gregation model is related to skweak, they provide a more limited choice of labelling functions, in particular regarding the inclusion of document-level constraints or under-specified labels. skweak is also another toolkit which is more distantly related to ensemble methods (Sagi and Rokach, 2018), as those meth- 339 ods also rely on multiple estimators whose results are combined at prediction time. Our method was to make the weak supervision easier and to make it very less compute inten- sive and as well as involve human in the loop to validate the labels being generated from the weak labeler models to fast track model training and collecting more better domain specific data easily.

Methodology

There are many different methods to employ weak supervisors. These supervisors are picked depending on the use case and availability of the resources and looking for variety of constraint's that organizations they have. Here are some common types:

- **Hard-coded heuristics**: usually regular expressions (regexes)
- **Syntactics:** for instance, Spacy's dependency trees
- **Distant supervision:** external knowledge bases
- **Noisy manual labels:** crowd-sourcing

- **Weak labeler models:** A small model trained with limited training data

In this paper our focus is more on how to use Weak labeler model to employ and weak super- vision method. Weak labeler models are used when there is an availability of very limited training data and is sufficient to train a classification model which can have an accuracy of more that 65% F1 Score. Given this scenario and having a small dataset this can be developed by following the below steps:

1. Utilize the data to train a weak labeler classification model
2. Gather all the unlabelled data-points which can be labelled using the weak labeler model
3. Once those data-points are labelled, Humans are used to validate the data-points predicted and make correction on the wrong predictions.
4. Once the data is verified, follow the steps 1-3 again till the point user have a good amount of data which can produce the best performing model.
5. Finally, collect all the data labelled through this process and train a final classifier model.

This is a simple way of employing a smaller model that can be used to label a million of data- points using programmatic approach and then utilizing human to iterate through them fastly as they give prefilled labels and human labelers just need to validate the predicted labels.

Conclusion

In addition to the above-mentioned methods, there exist other exciting approaches to tackle low-resource NLP, with many more to come in the future. Discussed method fastracks the genera- tion of more labels and helps in generating a more robust model. We have seen the development time to reduce to 70% and saw model performance improve by 5-7% but that comes with a caveat of human labelers involved in this method. These human labelers should understand the labels and the ability to understand the data they are working with. There would be need of unified labelling rules that would be followed by all the labelers involved in the process to be very consistent.

Future Work

Given the methodology involves human to validate the prediction done by the weak labeler model we would like to automate this process using some statistical model where in the process we can identify the labels that are predicted wrong and discard them in the next training process. This would enable users to automatically label more data for task such as classification.

References;

[1] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Re. 2017. ´ Snorkel: Rapid training data creation with weak supervision. Proc. VLDB Endow., 11(3):269–282.

[2] Pierre Lison, Jeremy Barnes, AliaksandrHubin, and SamiaTouileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1518–1533, Online. Association for Computational Linguistics.

[3] Esteban Safranchik, Shiying Luo, and Stephen Bach. 2020b. Weakly supervised sequence tagging from noisy rules. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04):5570–5578.

[4] Daniel Y. Fu, Mayee F. Chen, Frederic Sala, Sarah M. Hooper, KayvonFatahalian, and Christo- pher Re.´ 2020. Fast and three-rious: Speeding up weak supervision with triplet methods. In Proceedings of the 37th International Conference on Machine Learning (ICML 2020).

[5] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for re- lation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

[6] Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction. Transactions of the Association for Computa- tional Linguistics, 1:367–378.

[7] Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.

[8] Hyun-Chul Kim and ZoubinGhahramani. 2012. Bayesian classifier combination. In Proceed- ings of the Fifteenth International Conference on Artificial Intelligence and Statistics, volume 22 of

Proceedings of Machine Learning Research, pages 619–627, La Palma, Canary Islands. PMLR.

[9] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In Proceedings of the 2013 Conference of the North American Chap- ter of the Association for Computational Linguistics: Human Language Technologies, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

[10] An Thanh Nguyen, Byron Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 299–309, Vancouver, Canada. Association for Computational Linguistics.

[11] Jason Fries, Sen Wu, Alex Ratner, and Christopher Re.´ 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data.

[12] Omer Sagi and LiorRokach. 2018. Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery, 8(4):e1249.

[13] Clarke, James, Dan Goldwasser, Ming-Wei Chang, and Dan Roth. 2010. "Driving Semantic Parsing from the World's Response." In Proceedings of the Fourteenth Conference on Com- putational Natural Language Learning, 18–27. Association for Computational Linguistics.

[14] Guu, Kelvin, PanupongPasupat, Evan Zheran Liu, and Percy Liang. 2017. "From Language to Programs: Bridging Reinforcement Learning and Maximum Marginal Likelihood." arXiv Preprint arXiv:1704.07926.

[15] Stewart, Russell, and Stefano Ermon. 2017. "Label-Free Supervision of Neural Networks with Physics and Domain Knowledge." In AAAI, 2576–82.

[16] Takamatsu, Shingo, Issei Sato, and Hiroshi Nakagawa. 2012. "Reducing Wrong Labels in Distant Supervision for Relation Extraction." In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 721–29. Association for Computational Linguistics.

[17] Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society. Series B (Methodological). JSTOR, 267–88.

[18] Varma, Paroma, Rose Yu, Dan Iter, Christopher De Sa, and Christopher Ré. 2016. "Socratic Learning: Empowering the Generative Model." arXiv Preprint arXiv:1610.08123.

[19] Xiao, Tong, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. "Learning from Mas- sive Noisy Labeled Data for Image Classification." In Proceedings of the Ieee Conference on Computer Vision and Pattern Recognition, 2691–9.

[20] Zaidan, Omar F, and Jason Eisner. 2008. "Modeling Annotators: A Generative Approach to Learning from Annotator Rationales." In Proceedings of the Conference on Empirical Meth- ods in Natural Language Processing, 31–40. Association for Computational Linguistics.