|  |  |
| --- | --- |
| Title | Snorkel: Rapid Training Data Creation  with Weak Supervision |
| Author | Alexander Ratner Stephen H. Bach Henry Ehrenberg  Jason Fries Sen Wu Christopher Re´  Stanford University  Stanford, CA, USA |
| URL | https://arxiv.org/pdf/1711.10160.pdf |
| Content | **What is Snorkel:**  Snorkel is a first-of-its-kind system that enable users to train state-of-the-art models without hand labeling any training data.(labeling data is a bottleneck work), the first end-to-end system for combining weak supervision sources to rapidly create training data.  **Motivation:**  Training sets for ML are expensive to create which makes weak supervision neccessary for practice use.  Deep learning is a kind of machine learning which is effective in tasks like NLP and image analysis. However, deep learning is the most time-consuming method and it has a major upfront cost: these methods need massive training sets of labeled examples to learn from.(often tens of thousands to millions to reach peak predictive performance)  **Techniques:**  Snorkel has 3 principles:   1. Bring all sources to Bear 2. Training data as the interface to ML 3. Supervision as Interactive Programming   The architecture of snorkel:   1. writing labeling function 2. Modeling accuracies and correlations 3. Training a Discriminative Model   Contributions:   1. flexible interface for Sources 2. Tradeoffs in Modeling of Sources 3. First end-to-end system for data programming   Setup: make a label matrix, these training labels are used to train a discriminative model  Data model: how to manage complex unstructured data in a proper way that enables SMEs to write labeling functions over it. In Snorkel, input data is stored in a context hierarchy.  **Experiment:**  Evaluate Snorkel by drawing on deployments developed in collarboration with users.  Snorkel outperforms distant supervision baselines  Snorkel approaches hand supervision  Snorkel enables a new interaction paradigm  **Challenge and Related work:**  How to combine different sources: multi-instance learning to reduce the noise in weak supervision sources  (R. Hoffffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S.  Weld. Knowledge-based weak supervision for information  extraction of overlapping relations. In Meeting of the  Association for Computational Linguistics (ACL), 2011.)  (S. Riedel, L. Yao, and A. McCallum. Modeling relations and their mentions without labeled text. In European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), 2010.)  How to estimate the accuracy of labell sources without a gold standard to compare:  (A. P. Dawid and A. M. Skene. Maximum likelihood  estimation of observer error-rates using the EM algorithm.  Journal of the Royal Statistical Society C, 28(1):20–28,   1. )   Crowdsourcing in which workers have unknown accuracy  N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi.  Aggregating crowdsourced binary ratings. In International  World Wide Web Conference (WWW), 2013.  M. Joglekar, H. Garcia-Molina, and A. Parameswaran.  Comprehensive and reliable crowd assessment algorithms.  In International Conference on Data Engineering (ICDE),  2015.  Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. Spectral  methods meet EM: A provably optimal algorithm for  crowdsourcing. Journal of Machine Learning Research,  17:1–44, 2016.  Use generative models on with hand-specified dependency structure to label data for specific modalities  E. Alfonseca, K. Filippova, J.-Y. Delort, and G. Garrido.  Pattern learning for relation extraction with a hierarchical  topic model. In Meeting of the Association for  Computational Linguistics (ACL), 2012.  B. Roth and D. Klakow. Combining generative and  discriminative model scores for distant supervision. In  Conference on Empirical Methods on Natural Language  Processing (EMNLP), 2013.  S. Takamatsu, I. Sato, and H. Nakagawa. Reducing wrong  labels in distant supervision for relation extraction. In  Meeting of the Association for Computational Linguistics  (ACL), 2012.  Spectral methods  F. Parisi, F. Strino, B. Nadler, and Y. Kluger. Ranking and  combining multiple predictors without labeled data.  Proceedings of the National Academy of Sciences of the  USA, 111(4):1253–1258, 2014.  Semi-supervised  O. Chapelle, B. Sch¨olkopf, and A. Zien, editors.  Semi-Supervised Learning. Adaptive Computation and  Machine Learning. MIT Press, 2009.  Active learning  B. Settles. Active Learning. Synthesis Lectures on Artifificial  Intelligence and Machine Learning. Morgan & Claypool  Publishers, 2012.  Transfer learning  S. J. Pan and Q. Yang. A survey on transfer learning.  IEEE Transactions on Knowledge and Data Engineering,  22(10):1345–1359, 2010.  Self training  A. K. Agrawala. Learning with a probabilistic teacher.  IEEE Transactions on Infomation Theory, 16:373–379,  1970.  H. J. Scudder. Probability of error of some adaptive  pattern-recognition machines. IEEE Transactions on  Infomation Theory, 11:363–371, 1965.  Co trining  A. Blum and T. Mitchell. Combining labeled and unlabeled  data with co-training. In Workshop on Computational  Learning Theory (COLT), 1998.  Data fusion  X. L. Dong and D. Srivastava. Big Data Integration.  Synthesis Lectures on Data Management. Morgan &  Claypool Publishers, 2015.  T. Rekatsinas, M. Joglekar, H. Garcia-Molina,  A. Parameswaran, and C. R´e. SLiMFast: Guaranteed  results for data fusion and source reliability. In ACM  SIGMOD International Conference on Management of  Data (SIGMOD), 2017.  Truth discovery  Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and  J. Han. A survey on truth discovery. SIGKDD Explor.  Newsl., 17(2), 2015.  The latent truth model  B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A  Bayesian approach to discovering truth from conflflicting  sources for data integration. PVLDB, 5(6):550–561, 2012.  how to model user-specifified correlations among data  sources  R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, and  D. Srivastava. Fusing data with correlations. In ACM  SIGMOD International Conference on Management of  Data (SIGMOD), 2014. |
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|