Data Augmentation for ML-driven Data Preparation and Integration

Yuliang Li, Xiaolan Wang
Megagon Labs
{yuliang,xiaolan}@megagon.ai

Zhengjie Miao
Duke University
zjmiao@cs.duke.edu

Wang-Chiew Tan
Facebook AI
wangchiew@fb.com

ABSTRACT

In recent years, we have witnessed the development of novel data augmentation (DA) techniques for creating additional training data needed by machine learning based solutions. In this tutorial, we will provide a comprehensive overview of techniques developed by the data management community for data preparation and data integration. In addition to surveying task-specific DA operators that leverage rules, transformations, and external knowledge for creating additional training data, we also explore the advanced DA techniques such as interpolation, conditional generation, and DA policy learning. Finally, we describe the connection between DA and other machine learning paradigms such as active learning, pre-training, and weakly-supervised learning. We hope that this discussion can shed light on future research directions for a holistic data augmentation framework for high-quality dataset creation.

PVLDB Reference Format:
Yuliang Li, Xiaolan Wang, Zhengjie Miao, and Wang-Chiew Tan. Data Augmentation for ML-driven Data Preparation and Integration. PVLDB, 14(12): 3182-3185, 2021.
doi:10.14778/3476311.3476403

1 INTRODUCTION

Machine learning (ML), particularly deep learning, is revolutionizing almost all fields of computer science. Since the last decade, this trend has extended to classical data management tasks in data preparation and integration [3, 7, 8, 34, 40, 41, 45, 54] achieving promising results. For example, in Entity Matching (EM), ML-based solutions [3, 34] achieved the state-of-the-art matching quality across EM benchmarks by fine-tuning pre-trained language models (LMs) such as BERT. However, just like in NLP or CV, these ML-based solutions are *data-hungry*: they typically require to be trained on a large, high-quality labeled dataset to achieve the best results. For example, in EM, the size of an ideal training set can be up to tens of thousands of labeled pairs of match/not-match entity records. Such high label requirements prevent the adoption of machine learning methods to a wider range of new domains and applications in practice.

To this end, data augmentation (DA) has become a common practice in ML to address the challenge of insufficient training data. The goal of data augmentation is to *create synthetic training examples*. For example, in image classification, simple transformations such as rotation, cropping, or flipping are shown to be effective in generating semantics-preserving modified images to boost the performance of an image classifier. It has been an active line of research in NLP and CV exploring the space of possible data augmentation operators as well as techniques for tuning and composing these operators to form more effective *data augmentation policies*.

In this tutorial, we aim at providing a comprehensive overview of data augmentation techniques for ML-driven data preparation and integration tasks. More specifically, we focus on information extraction, data cleaning, and schema/entity matching where ML-based solutions heavily rely on labeled examples. Apart from surveying existing DA techniques that commonly leverage rules, transformations, or external knowledge, this tutorial also covers advanced topics including interpolation [44, 64], conditional generation [32, 50], and Auto-ML [9, 38]. These are techniques that have been shown to be successful in related NLP/CV tasks which we believe have a high potential also in data management tasks. We will also draw the connections between DA and other machine learning methods such as active learning, pre-training, and weakly-supervised learning that interest the DB community at large.

Scope, target audience, and outline. We plan for a 3-hour tutorial but are also flexible with a 1.5-hour arrangement. The tutorial targets both data management researchers and practitioners who are interested in learning about any of these topics: data integration, cleaning, extraction, data augmentation, and ML. However, there are no pre-requisites for this tutorial apart from basic data management background.

This tutorial will start with a general introduction of the aforementioned data management tasks, their recent ML-based solutions, and DA (Section 2). Next, we provide a survey of existing DA techniques for each task (Section 3). We will also overview advanced ML techniques on how to further improve the effectiveness of DA (Section 4). Finally, we will connect the existing approaches with other learning paradigms to shed light on potential future research directions (Section 5). In the 1.5-hour version, we will shorten Section 3 and 4 as well as keeping Section 5 a brief discussion.

Recent related tutorials. This will be the first tutorial focusing on data augmentation for data management tasks. There were two related tutorials presented at recent data-centric research venues ([17] at SIGMOD 2018 and VLDB 2018 that covered ML-based data integration and [59] at VLDB 2020 that referred to DA as part of data acquisition that integrates training data with additional data).

2 BACKGROUND

Machine learning, especially supervised learning models, has been used for solving data management tasks, including data integration, data cleaning, and information extraction, for years [17]. Techniques used for these problems also evolve from Naïve Bayes [60],

Proceedings of the VLDB Endowment, Vol. 14, No. 12 ISSN 2150-8097.
doi:10.14778/3476311.3476403

decision trees [5], to deep neural networks [29, 45, 46, 59] and recently pre-trained language models (LMs) [34, 43, 65].

To harness the power ML, many supervised ML models, especially deep learning models, require large amounts of annotated training data to avoid over-fitting, increase robustness and improve quality. However, acquiring training data is time consuming, expensive, and oftentimes error-prone [53] process. Therefore, data augmentation is used for automatically enriching and diversifying the existing training dataset without collecting new labels. Data augmentation refers to **the process of automatically enriching and diversifying examples in the training dataset**.

In computer vision, data augmentation operators, such as rotating, cropping, padding, and flipping the original image, are widely used and proved to be very effective [9, 49]. Recent years, data augmentation also has received increasing attention recently in the natural language processing community [18, 30, 58]. Likewise, data augmentation also significantly benefits many ML-based solutions for data management tasks.

3 DA FOR DATA MANAGEMENT

In this section, we will focus on two main problems, data preparation and data integration, and describe how existing ML-based solutions benefit from the data augmentation process.

3.1 DA for Data Preparation

Data preparation plays an essential role for many data analytic applications. Two significant tasks for data preparation include information extraction and data cleaning. Recently, ML-based solutions, in particular those that are enhanced by data augmentation, have become one of the main streams for solving these tasks.

Information extraction. Information extraction focuses on extracting structured information from unstructured or semi-structured data sources, and it is a popular research topic for both data management (DB) and natural language processing (NLP). Information extraction includes several core tasks, such as named entity recognition, relation extraction, and coreference extraction. Many recent solutions to these core information extraction tasks rely on training examples and machine learning approaches. Thus, such solutions can greatly benefit from data augmentation to diversify and enrich the training dataset, which will further alleviate the cost for collecting high quality labeled data and improve the model accuracy.

Named entity recognition task is often formulated as a sequence tagging problem. Mathew et al. [11] and Snippext [44] adapt several basic data augmentation operators, commonly used for sequence classification tasks, to perform the sequence tagging task; DAGA [13] uses conditional generation to produce synthetic training examples. Relation extraction focuses on assigning a relation label for two entities in a given context. To augment examples in the training set, Xu et al. [63] use dependency path between entities to classify the relation and they augment the paths via entity directions; Lin and Miller et al. [37] leverage external ontology knowledge to augment training examples.

Data Cleaning. Error detection is an essential step for data cleaning. Given a cell in a database, the goal of error detection is to determine whether its values are correct or not. Thus, it is natural to use machine learning for error detection by classifying whether the given cell is either clean or dirty.

HoloDetect [26] uses a data augmentation-based approach for detecting erroneous data. In essence, HoloDetect enriches and balances the labels of a small training data through the learned data transformations and data augmentation policies.

3.2 DA for Data Integration

The research of data integration has been expanded in several core directions, such as schema matching/mapping and entity matching. Many of these core tasks have benefited significantly from the recent advances in machine learning (ML) [15, 17] and human annotated datasets.

Schema Matching. Schema Matching focuses on finding the correspondence among schema elements in two semantically correlated schema. To use machine learning for schema matching, the problem can also be formulated as a classification problem [14]: for each schema element from the source schema, the task is to assign labels that correspond to schema element from the target schema.

Augmenting training examples has been applied for schema matching solutions for years. Madhavan et al. [39] use mapping between schemas in the same corpus to augment existing training examples; Dong et al. [16] adapt a similar augmentation method to enrich training examples for a ML model that predicts the similarity between two schema elements; ADnEv [54] augments training data of similarity matrices used for improving schema matching results.

Entity Matching. Entity Matching (also known as record linkage and entity resolution) is the problem of identifying records that refer to the same real-world entity, and it is an important task for data integration. A fundamental step for entity matching is to classify whether entity pairs as either matching or non-matching.

To enrich training examples for entity matching, Ditto [34] applies 5 distinct basic data augmentation operators in three different levels to transform existing examples into new (synthetic) ones; Thirumuruganathan et al. [56] use data augmentation to create new training examples from the unlabeled dataset, by assigning both positive and negative labels to a transformed data point in the unlabeled set, to enforce a strong co-regularization on the classifier.

4 ADVANCED DATA AUGMENTATION

In this section, we will cover some advanced DA techniques emerging from NLP/CV tasks and discuss their usage in data management tasks. These techniques heavily rely on recent ML techniques like representation learning, neural sequence generation, and Auto-ML.

Interpolation-based DA. MixUp [64], a recent data augmentation method for image classification, produces virtual training examples by combining two randomly sampled training examples into their linear interpolations. Variants of MixUp have also achieved significant improvements on sequence classification and tagging tasks. We will first introduce methods that adapted the MixUp technique to sequential data by performing interpolations between two sequences in their embedding space [6, 22]. Then we present MixDA [44], which interpolate original training examples' encoded representations with augmented sentences by simple augmentation operators mentioned in Section 3. After that, we discuss how to apply MixDA to data integration tasks, using Ditto [34] for entity matching as an example.

Generation-based DA. Leveraging the recent advancements in generative pre-trained language modeling [32, 50], this category of DA methods attempts to overcome the challenge of lacking diversity as in simple DA operators. We will review the background knowledge about neural text generation and introduce recent DA techniques that it inspires. With the goal of reducing the label corruptions and to further diversify the augmented examples, these techniques filter out low-quality generations using the target model [1] or apply conditional generation on the given labels [31]. We also discuss a recent DA method InvDA [43] trained on the task-specific corpus in a self-supervised manner, which learns how to augment existing examples by "inverting" the effect of multiple simple DA operators and has been shown effective for entity matching and data cleaning. There is another line of generation-based DA methods using Generative Adversarial Networks (GANs) [21] in CV. For relational data, researchers have used GANs to synthesize tables [19, 48], which can also be used for DA.

Learned DA policy. This category of DA methods aims at automatically finding best DA policies (combination of DA operators), by solving an additional learning task. We first introduce different optimization goals for the DA-learning task [9, 10, 27, 33, 35, 38, 47] and the different searching techniques to solve the DA-learning task, including Bayesian optimization [36], reinforcement learning [9, 27, 47, 52], and meta-learning [23, 33, 35, 38]. Among these approaches, meta-learning-based searching techniques show better efficiency since they enable the use of gradient descent by differentiating the search space. Finally, we present a meta-learning-based framework Rotom [43] that adapts the most popular optimization objective (minimizing the validation loss) and the to select and combine augmented examples.

5 DA WITH OTHER LEARNING-PARADIGMS

We finally discuss several opportunities and open challenges in combining data augmentation with other learning paradigms other than supervised learning for data preparation and integration.

Semi-supervised and active learning. In addition to labeled examples, data augmentation can also be applied to unlabeled data in a semi-supervised manner to exploit the large number of unlabeled examples [2, 43, 62] for consistency regularization. Active learning, which selects the most informative unlabeled examples for human to assign labels and update the model, has also been used in data integration tasks [29, 42]. Both the initial model training and the iterative labeling process of active learning can benefit from data augmentation to further reduce the label requirement [20], but it is non-trivial to make the DA process and the fine-tuning of deep learning models interactive enough to support user inputs.

Weak-supervision. Data augmentation is sometimes referred to as a special form of weak supervision, which in general uses noisy sources such as crowd-sourcing and user-defined heuristics to provide supervision signals from unlabeled examples. Data programming [51, 57] enables developers to provide data programs (labeling functions) that labels a subset of the unlabeled examples. In the same manner, Snorkel [52] takes as input the user-defined DA operators (transformation functions) and learns to apply them in sequence, which can be a good accompany of the DA methods discussed in this tutorial. One challenge remains in data programming is the difficulty of generating functions by enumerating heuristics rules, which may be potentially addressed by data transformation techniques [24, 25, 28] that have been extensively studied in the DB community.

Pre-training for relational data. It has been shown that pre-trained language models can be used to construct distributed representations of relational data entries and provide significant performance gain [34]. However, LMs did not characterize the structure information and factual knowledge in relational data. Very recently people have started investigating structure-aware representation learning for relational data in different data integration tasks [4, 12, 55], and it is promising but also challenging to have pre-trained models for different domains and tasks. We expect pre-trained models for relational data to provide effective DA for data integration tasks, like LMs for text data augmentation [27, 31, 61]. Given the huge success of pre-trained LMs in NLP community, publicly available pre-trained models for relational data would boost future research for data integration and table understanding.

6 BIO SKETCHES

Yuliang Li is a senior research scientist at Megagon Labs where he leads the efforts of building data integration (entity matching) and extraction systems with low label requirements. He received his PhD from UC San Diego in 2018.

Xiaolan Wang is a research scientist at Megagon Labs. At Megagon Labs, she is leading the ExtremeReading project that automatically summarizes text-based customer reviews. She received her PhD from University of Massachusetts Amherst in 2019.

Zhengjie Miao is a PhD candidate in Computer Science at Duke University. He is broadly interested in building techniques to reduce human effort in data analytics.

Wang-Chiew Tan is a research scientist at Facebook AI. Prior to that, she was at Megagon Labs and was a Professor of Computer Science at University of California, Santa Cruz. She also spent two years at IBM Research - Almaden. Her research interests include data integration and exchange, data provenance, and natural language processing.

REFERENCES

[1] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do Not Have Enough Data? Deep Learning to the Rescue!. In *AAAI*. 7383–7390.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*. 5049–5059.

[3] Ursin Brunner and Kurt Stockinger. 2020. Entity matching with transformer architectures-a step forward in data integration. In *EDBT*.

[4] Riccardo Cappuzzo, Paolo Papotti, and Saravanan Thirumuruganathan. 2020. Creating embeddings of heterogeneous relational datasets for data integration tasks. In *SIGMOD*. 1335–1349.

[5] Surajit Chaudhuri, Bee-Chung Chen, Venkatesh Ganti, and Raghav Kaushik. 2007. Example-driven design of efficient record matching queries.. In *VLDB*, Vol. 7. 327–338.

[6] Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. In *ACL*. 2147–2157.

[7] Xu Chu, Ihab F. Ilyas, and Paolo Papotti. 2013. Holistic data cleaning: Putting violations into context. In *ICDE*, Christian S. Jensen, Christopher M. Jermaine, and Xiaofang Zhou (Eds.). IEEE Computer Society, 458–469.

[8] Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang, and Yin Ye. 2015. KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing. In *SIGMOD*, Timos K. Sellis, Susan B. Davidson, and Zachary G. Ives (Eds.). ACM, 1247–1261.

[9] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *CVPR*. 113–123.

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR workshops*. 702–703.

[11] Xiang Dai and Heike Adel. 2020. An Analysis of Simple Data Augmentation for Named Entity Recognition. In *COLING*. 3861–3867.

[12] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2020. TURL: table understanding through representation learning. *PVLDB* 14, 3 (2020), 307–319.

[13] Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data Augmentation with a Generation Approach forLow-resource Tagging Tasks. In *EMNLP*. 6045–6057.

[14] AnHai Doan, Pedro Domingos, and Alon Y Halevy. 2001. Reconciling schemas of disparate data sources: A machine-learning approach. In *SIGMOD*. 509–520.

[15] AnHai Doan, Alon Halevy, and Zachary Ives. 2012. *Principles of data integration*. Elsevier.

[16] Xin Dong, Jayant Madhavan, and Alon Halevy. 2004. Mining structures for semantics. *ACM SIGKDD Explorations Newsletter* 6, 2 (2004), 53–60.

[17] Xin Luna Dong and Theodoros Rekatsinas. 2018. Data integration and machine learning: A natural synergy. In *SIGMOD*. 1645–1650.

[18] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440* (2017).

[19] Ju Fan, Tongyu Liu, Guoliang Li, Junyou Chen, Yuwei Shen, and Xiaoyong Du. 2020. Relational Data Synthesis using Generative Adversarial Networks: A Design Space Exploration. *PVLDB* 13, 11 (2020), 1962–1975.

[20] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. 2020. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*. Springer, 510–526.

[21] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NeurIPS*.

[22] Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941* (2019).

[23] Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. 2020. Faster AutoAugment: Learning Augmentation Strategies Using Backpropagation. In *ECCV*, Vol. 12370. Springer, 1–16.

[24] Yeye He, Kris Ganjam, Kukjin Lee, Yue Wang, Vivek Narasayya, Surajit Chaudhuri, Xu Chu, and Yudian Zheng. 2018. Transform-data-by-example (tde) extensible data transformation in excel. In *SIGMOD*. 1785–1788.

[25] Jeffrey Heer, Joseph M. Hellerstein, and Sean Kandel. 2015. Predictive Interaction for Data Transformation. In *CIDR*.

[26] Alireza Heidari, Joshua McGrath, Ihab F Ilyas, and Theodoros Rekatsinas. 2019. Holodetect: Few-shot learning for error detection. In *SIGMOD*. 829–846.

[27] Zhiting Hu, Bowen Tan, Russ Salakhutdinov, Tom Mitchell, and Eric Xing. 2019. Learning data manipulation for augmentation and weighting. In *NeurIPS*. 15764–15775.

[28] Zhongjun Jin, Michael R Anderson, Michael Cafarella, and HV Jagadish. 2017. Foofah: Transforming data by example. In *SIGMOD*. 683–698.

[29] Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource Deep Entity Resolution with Transfer and Active Learning. In *ACL*. 5851–5861.

[30] Sosuke Kobayashi. 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *NAACL-HLT*. 452–457.

[31] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data Augmentation using Pre-trained Transformer Models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*. 18–26.

[32] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[33] Yonggang Li, Guosheng Hu, Yongtao Wang, Timothy Hospedales, Neil M Robertson, and Yongxing Yang. 2020. DADA: Differentiable Automatic Data Augmentation. *arXiv preprint arXiv:2003.03780* (2020).

[34] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep entity matching with pre-trained language models. *PVLDB* 14, 1 (2020), 50–60.

[35] Hanwen Liang, Shifeng Zhang, Jiacheng Sun, Xingqiu He, Weiran Huang, Kechen Zhuang, and Zhenguo Li. 2019. Darts+: Improved differentiable architecture search with early stopping. *arXiv preprint arXiv:1909.06035* (2019).

[36] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. 2019. Fast autoaugment. In *NeurIPS*. 6665–6675.

[37] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2016. Improving temporal relation extraction with training instance augmentation. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. 108–113.

[38] Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2018. DARTS: Differentiable Architecture Search. In *ICLR*.

[39] Jayant Madhavan, Philip A Bernstein, AnHai Doan, and Alon Halevy. 2005. Corpus-based schema matching. In *ICDE*. IEEE, 57–68.

[40] Mohammad Mahdavi and Ziawasch Abedjan. 2020. Baran: Effective Error Correction via a Unified Context Representation and Transfer Learning. *PVLDB* 13, 11 (2020).

[41] Mohammad Mahdavi, Ziawasch Abedjan, Raul Castro Fernandez, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and Nan Tang. 2019. Raha: A configuration-free error detection system. In *SIGMOD*. 865–882.

[42] Venkata Vamsikrishna Meduri, Lucian Popa, Prithviraj Sen, and Mohamed Sarwat. 2020. A Comprehensive Benchmark Framework for Active Learning Methods in Entity Matching. In *SIGMOD*. 1133–1147.

[43] Zhengjie Miao, Yuliang Li, and Xiaolan Wang. 2021. Rotom: A Meta-Learned Data Augmentation Framework for Entity Matching, Data Cleaning, Text Classification, and Beyond. In *SIGMOD*. 1303–1316.

[44] Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *Proceedings of The Web Conference 2020*. 617–628.

[45] Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. 2018. Deep learning for entity matching: A design space exploration. In *SIGMOD*. 19–34.

[46] Hao Nie, Xianpei Han, Ben He, Le Sun, Bo Chen, Wei Zhang, Suhui Wu, and Hao Kong. 2019. Deep sequence-to-sequence entity matching for heterogeneous entity resolution. In *CIKM*. 629–638.

[47] Tong Niu and Mohit Bansal. 2019. Automatically Learning Data Augmentation Policies for Dialogue Tasks. In *EMNLP-IJCNLP*. 1317–1323.

[48] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. Data synthesis based on generative adversarial networks. *PVLDB* 11, 10 (2018), 1071–1083.

[49] Luis Perez and Jason Wang. 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621* (2017).

[50] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[51] Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. *PVLDB* 11, 3 (2017), 269–282.

[52] Alexander J Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. 2017. Learning to compose domain-specific transformations for data augmentation. In *NeurIPS*. 3236–3246.

[53] Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*. Vancouver, CA, 1–10.

[54] Roee Shraga, Avigdor Gal, and Haggai Roitman. 2020. ADnEV: cross-domain schema matching using deep similarity matrix adjustment and evaluation. *PVLDB* 13, 9 (2020), 1401–1415.

[55] Nan Tang, Ju Fan, Fangyi Li, Jianhong Tu, Xiaoyong Du, Guoliang Li, Samuel Madden, and Mourad Ouzzani. 2021. RPT: Relational Pre-trained Transformer Is Almost All You Need towards Democratizing Data Preparation. *PVLDB* 14, 8 (2021), 1254–1261.

[56] Saravanan Thirumuruganathan, Shameem A Puthiya Parambath, Mourad Ouzzani, Nan Tang, and Shafiq Joty. 2018. Reuse and adaptation for entity resolution through transfer learning. *arXiv preprint arXiv:1809.11084* (2018).

[57] Paroma Varma and Christopher Ré. 2018. Snuba: Automating Weak Supervision to Label Training Data. *PVLDB* 12, 3 (2018), 223–236.

[58] Jason W. Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *EMNLP-IJCNLP*. Association for Computational Linguistics, 6381–6387.

[59] Steven Euijong Whang and Jae-Gil Lee. 2020. Data collection and quality challenges for deep learning. *PVLDB* 13, 12 (2020), 3429–3432.

[60] William E Winkler. 1999. The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*. Citeseer.

[61] Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*. Springer, 84–95.

[62] Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised Data Augmentation for Consistency Training. In *NeurIPS*.

[63] Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. 2016. Improved relation classification by deep recurrent neural networks with data augmentation. In *COLING*. 1461–1470.

[64] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

[65] Chen Zhao and Yeye He. 2019. Auto-EM: End-to-end Fuzzy Entity-Matching using Pre-trained Deep Models and Transfer Learning. In *The World Wide Web Conference*. 2413–2424.