JOINT TEXT AND LABEL GENERATION FOR SPOKEN LANGUAGE UNDERSTANDING

Yang Li[*]
Department of Computer Science
University of North Carolina at Chapel Hill
yangli95@cs.unc.edu

Ben Athiwaratkun & Cicero Nogueira dos Santos & Bing Xiang
Amazon Web Services
{benathi,cicnog,bxiang}@amazon.com

ABSTRACT

Generalization is a central problem in machine learning, especially when data is limited. Using prior information to enforce constraints is the principled way of encouraging generalization. In this work, we propose to leverage the prior information embedded in pretrained language models (LM) to improve generalization for intent classification and slot labeling tasks with limited training data. Specifically, we extract prior knowledge from pretrained LM in the form of synthetic data, which encode the prior implicitly. We fine-tune the LM to generate an *augmented language*, which contains not only text but also encodes both intent labels and slot labels. The generated synthetic data can be used to train a classifier later. Since the generated data may contain noise, we rephrase the learning from generated data as learning with noisy labels. We then utilize the mixout regularization for the classifier and prove its effectiveness to resist label noise in generated data. Empirically, our method demonstrates superior performance and outperforms the baseline by a large margin.

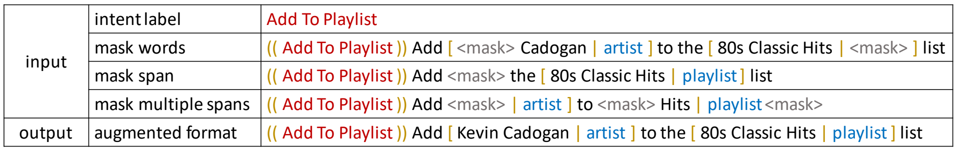
1 INTRODUCTION

Natural language processing has been profoundly impacted by leveraging pretrained large-scale language models. Many downstream tasks have achieved state-of-the-art performance by fine-tuning pretrained models (Devlin et al., 2018; Radford et al., 2019; Raffel et al., 2019). The success lies in the transfer ability of those models trained with large unlabeled corpus, which are typically thought of learning universal language representations (Howard & Ruder, 2018). Although the wide adoption, fine-tuning still requires a large dataset to achieve good performance. In some cases, however, it would be inconvenient, expensive, or even impossible to collect a large dataset. For instance, when adapting a personal voice assistant to a specific user, it is inconvenient to require the user to label a lot of utterances.

In this work, we focus on a setting, where we only have access to very limited data from each testing domains. Learning with limited data is challenging since the modern over-parametrized models can easily overfit the small training dataset while simple models usually suffer from insufficient representative power. According to the Bayesian Occam's razor theory (MacKay, 1992), exploiting prior information is a principled way of encouraging generalization when faced with limited data. In this work, we leverage the priors embedded in pretrained language models.

Following the Bayesian view of data augmentation (Zhang et al., 2016; Dao et al., 2019), we express synthetic data as an implicit form of prior that encodes data and task invariances. In our proposed approach, prior information is distilled by generating task-specific synthetic data from pretrained language models. In order to generate task-specific data, we fine-tune the language models over the small training dataset. The augmented datasets are then used to train the classifiers. The synthetic

[*] work done while interning at Amazon



<table>
<tr><td rowspan="4">input</td><td>intent label</td><td>Add To Playlist</td></tr>
<tr><td>mask words</td><td>((Add To Playlist)) Add [<mask> Cadogan | artist] to the [80s Classic Hits | <mask>] list</td></tr>
<tr><td>mask span</td><td>((Add To Playlist)) Add <mask> the [80s Classic Hits | playlist] list</td></tr>
<tr><td>mask multiple spans</td><td>((Add To Playlist)) Add <mask> | artist] to <mask> Hits | playlist <mask></td></tr>
<tr><td>output</td><td>augmented format</td><td>((Add To Playlist)) Add [Kevin Cadogan | artist] to the [80s Classic Hits | playlist] list</td></tr>
</table>

Figure 1: Input and output format of the conditional generator.

data embody the prior by teaching the classifier about possible tokens for each label. The generation process also bears similarity to knowledge distillation (Hinton et al., 2015). Here, we distill prior knowledge from pretrained language models.

We focus on the tasks of intent classification and slot labeling, which correspond to sentence and token level classification tasks, respectively. Therefore, we require the synthetic data to contain both intent and slot labels. To generate text (utterances) and labels simultaneously, we leverage an augmented language format (Athiwaratkun et al., 2020), where intent and slot label information is embedded in the generated sentences using an special format (see Fig. D.2). We fine-tune conditional language models to directly generate the augmented sentences. We test and compare multiple conditional generation strategies for synthesizing data with legitimate intent and slot labels.

Since we only have access to a very limited data to fine-tune the language model, it is inevitable that the generated data would contain noise. To resist label noise, we apply a recently proposed regularization method called mixout (Lee et al., 2019). Mixout is originally proposed to improve the generalization of large-scale language models. Here, we prove and empirically show that mixout regularized models are robust to label noise.

Our contributions are as follows: 1) We extract prior knowledge in the form of synthetic data to improve generalization of classification models 2) We utilize an augmented language format to simultaneously generate sentences and the corresponding intent and slot labels. 3) We propose two metrics that measure the correspondence between generated tokens and labels. 4) We reinterpret mixout as a regularization that resist label noise and prove its generalization bound under mild assumptions. 5) We significantly outperform BERT-based baselines for joint slot labeling and intent classification in limited data regime.

2 Methods

In this section, we formally describe the problem and introduce our proposed method, which employs three main steps as illustrated in Fig C.1: 1) finetuning of a pretrained encoder-decoder (enc-dec) language model (our *conditional generator*) using the small labeled set; 2) generation of synthetic data; 3) training of the *joint classifier* using the augmented training set. We detail our conditional generative model and describe the modifications we made to deal with noisy generations. Finally, we introduce the evaluation metrics to compare generations and present our training procedure.

Problem Formulation In this work, we are interested in slot labeling and intent classification. We focus on the limited data regime where a very small training dataset $\{(x_i, s_i, y_i)\}_{i=1}^N$ is given. x, s and y represent sentence, slot labels and intent label respectively. We will build a model $f(x;\theta) = p_\theta(s, y \mid x)$ to jointly classify slots and intents for novel sentences x: $s^*, y^* = \arg\max_{s,y} p_\theta(s, y \mid x)$. We augment the small training set with synthetic data $\{(x', s', y')\}_{j=1}^{N'}$ generated from pretrained language models and train a joint classifier to predict intent and slot labels simultaneously. We will describe the generative and discriminative components below.

Conditional Generator In this section, we introduce the generative component for drawing new training instances. First, we describe the augmented language format (Athiwaratkun et al., 2020) utilized to generate utterances (text) and labels simultaneously. As shown in the last row in Fig. 1, we use additional markers to indicate the token-spans and their associated labels. The augmented format can be converted from and to the traditional BIO format without loss of information (see Fig. D.2 in Appendix). With the training data converted into the augmented format, we can train a generative model to capture the joint distribution of the utterances and the labels, i.e., $p(x, s, y)$.

There are many options for modeling the joint distributions, such as VAE-based models (Kingma & Welling, 2013; Yang et al., 2017), autoregressive models (Radford et al., 2019) and GANs (Goodfellow et al., 2014; de Masson d'Autume et al., 2019). In this work, we employ a seq2seq model (Raffel et al., 2019) due to its verified ability for language modeling. It also gives us the flexibility to condition on additional information. That is, we essentially model the conditional distributions $p(x, s, y \mid c)$, where c represents the conditioning information described later. The conditional generation mechanism enables us to control the generation quality and diversity by varying the conditioning input c. We explore multiple different types of conditioning: 1) condition on the intent labels; 2) mask out several words from the augmented sentence at random; 3) mask out a random span from the augmented sentence; 4) mask out multiple spans. Please refer to Fig. 1 for an illustration.

Sampling and Filtering Sampling from the conditional generator is straight-forward, we just need to condition on the corresponding inputs. Intent labels are sampled from all possible labels and transformed to natural words separated by white spaces. Masked conditioning inputs are constructed by randomly sampling one augmented sentence from the given training set and replacing some tokens or spans with a mask token. The generated augmented sentences can then be decoded into BIO format and used to train the downstream models. However, generations could be noisy, some of them might not follow the exact format of the augmented language, thus cannot be decoded into the corresponding BIO format; some generations might contain invalid labels. We postprocess the generations by simply dropping those generations.

Joint Classifier After generating the synthetic data, we train a classifier to jointly predict intent and slot labels using both real and synthetic data. We again utilize a pretrained transformer model and add two classification heads above the language model backbone. The classification heads and the backbone are jointly trained with cross entropy loss.

Although we carefully tune the generator and filter the invalid samples to generate high quality data, it is inevitable to contain noise in the generated data. For example, some tokens might be labeled incorrectly in the synthetic data. The label noise can be detrimental since modern over-parametrized models can easily overfit to the noisy labels (Zhang et al., 2016). We rephrase learning with generated data as learning with noisy labels thus connect it with a well-studied literature (Song et al., 2020). As a simple modification, we apply a recently proposed regularization, mixout (Lee et al., 2019). Mixout is originally proposed to improve the generalization of fine-tuning large-scale language models. It has been shown effective particularly when training data is limited. Here, we provide another perspective by showing that mixout regularization is naturally robust to label noise. Please refer to Sec. A for more details.

Training Procedure As our model consists of two separate components, conditional generator and joint classifier, we would like to tune the generator so that the classifier trained with generated data can achieve the best performance on validation set. However, it would be difficult to determine the hyperparameters for generator based on validation set performance, since it requires to train a classifier till convergence. The likelihood of the validation set cannot be used as a metric, since some hyperparameters, like the probability of replacing a token, can affect the underlying conditional distribution $p(x, s, y \mid c)$, and comparing different conditional likelihood is meaningless. Instead, we propose two new metrics that measure the correspondence between generated tokens and labels. Our proposed metrics leverage the augmented language format to measure if the token-label joint distribution matches with the real one. We empirically verify the correlation between those metrics and the downstream task performance. Hence, we can search hyperparameters for the generator until it achieves the best metric scores without having to train the classifier. Please see Sec. B for details.

3 Experiments

In this section, we evaluate our framework on several public benchmark datasets, including SNIPS (Coucke et al., 2018), ATIS (Hemphill et al., 1990) and NLUED (Liu et al., 2019). Please see appendix for preprocessing details. Baseline models include **JBERT** which directly fine-tune the BERT model with the small training set to jointly predict the intent and the slots. The classification heads are two fully connected layers. We also compare to a factorized augmentation strategy similar to (Yoo et al., 2019) and denote it as **G(factor)+JBERT**, where the augmentation is performed by first sampling a sentence from a fine-tuned T5 model condition on the intent and then obtaining the slot labels from a fine-tuned BERT classifier with mixout regularization, i.e., $p(x, s, y) = p(y)p(x \mid y)p(s \mid x)$. Other classic data augmentation methods, such as EDA (Wei & Zou, 2019), are not suitable for out setting since the augmentation may change the slot types. Our proposed methods are denoted as

Table 1: Slot labeling and intent classification performance with four different sampling ratios.

(a) SNIPS

	Slot Labeling (F1)				Intent Classificaton (Acc.)			
	0.25%	0.5%	1%	2%	0.25%	0.5%	1%	2%
JBERT	45.96±5.16	62.22±1.64	78.36±1.86	88.94±0.92	87.11±8.09	90.96±0.47	96.68±0.40	98.21±0.52
G(factor)+JBERT Yoo et al. (2019)	48.34±1.54	64.73±2.15	79.10±1.36	88.87±1.21	88.91±3.95	94.66±0.92	96.36±0.82	98.02±0.57
G(intent)+JBERT [ours]	52.26±4.55	74.21±3.62	**85.33±0.87**	89.75±0.45	91.68±2.46	95.21±2.29	97.46±0.97	98.43±0.65
G(words)+JBERT [ours]	59.12±4.27	73.30±2.62	84.26±0.64	89.86±0.99	91.04±1.59	**96.50±0.46**	97.71±0.79	98.46±0.37
G(span)+JBERT [ours]	61.24±1.36	**75.39±2.76**	84.99±1.30	89.27±0.30	**93.21±3.75**	95.36±3.00	**97.79±0.46**	**98.46±0.23**
G(multi_spans)+JBERT [ours]	**63.00±4.48**	74.51±1.67	85.03±0.72	**90.07±0.91**	91.68±0.27	95.25±0.97	97.57±0.39	98.43±0.30

(b) ATIS

	Slot Labeling (F1)				Intent Classificaton (Acc.)			
	0.25%	0.5%	1%	2%	0.25%	0.5%	1%	2%
JBERT	52.20±2.83	63.25±4.12	71.53±2.52	79.67±0.84	79.96±1.64	83.76±1.18	87.07±2.21	90.12±0.64
G(factor)+JBERT Yoo et al. (2019)	53.68±3.74	64.39±3.71	72.68±2.93	79.45±0.90	77.99±9.74	**86.53±3.31**	**91.09±1.09**	**93.59±1.65**
G(intent)+JBERT [ours]	55.52±2.39	66.48±2.49	74.39±1.23	79.71±1.67	70.46±6.21	83.51±5.19	87.29±3.41	90.99±0.88
G(words)+JBERT [ours]	**61.31±1.66**	**69.49±2.98**	**74.97±0.73**	**81.79±1.90**	**80.43±3.86**	84.71±1.55	88.52±2.45	92.86±1.09
G(span)+JBERT [ours]	58.08±1.45	66.51±2.38	72.68±3.08	80.30±1.72	76.26±3.12	80.29±5.39	84.71±1.92	86.51±2.60
G(multi_spans)+JBERT [ours]	58.21±3.38	68.44±2.41	73.63±2.23	81.30±1.04	73.26±14.1	83.62±2.45	88.27±1.03	90.30±3.23

(c) NLUED

	Slot Labeling (F1)				Intent Classificaton (Acc.)			
	0.5%	1%	2%	4%	0.5%	1%	2%	4%
JBERT	28.58±2.17	39.21±1.37	53.45±1.66	60.93±0.84	31.76±1.54	53.42±4.09	72.35±0.52	79.44±0.60
G(factor)+JBERT Yoo et al. (2019)	30.34±1.85	39.35±0.95	52.07±2.51	59.62±1.10	**59.87±0.93**	**66.89±1.24**	74.65±1.64	**80.97±0.75**
G(intent)+JBERT [ours]	32.39±2.96	40.90±1.90	50.71±1.56	58.26±1.03	45.24±5.40	55.25±3.25	72.21±0.82	78.69±1.00
G(words)+JBERT [ours]	39.53±1.15	42.94±2.49	54.12±1.54	60.07±1.35	44.63±4.47	58.99±1.79	71.75±1.78	79.69±1.02
G(span)+JBERT [ours]	**42.00±0.95**	**47.32±1.61**	**56.11±1.65**	**60.57±1.08**	51.97±3.85	64.22±0.96	**74.86±1.28**	79.76±0.43
G(multi_spans)+JBERT [ours]	38.94±1.97	43.48±0.93	52.36±1.35	59.35±1.46	55.11±2.12	60.27±2.11	72.42±2.75	79.86±0.99

G(intent)+JBERT, **G(words)+JBERT**, **G(span)+JBERT** and **G(multi_spans)+JBERT** to represent models with generator conditioning on intent labels and different masking schemes respectively. We fine-tune the `T5-large` as the generator and `BERT-large-cased` as the classifier for the main results. Ablation studies are performed in Sec. F.3. To deal with the small training set size and noisy labels in generated data, mixout regularization is applied to the classifier in all models. We generate 500 synthetic data per intent class for SNIPS dataset and 50 for ATIS and NLUED.

Table 1 shows the slot labeling and intent classification performance on SNIPS, ATIS and NLUED. We conduct experiments with four different sampling ratios to subsample the training set. Intent classification performance is evaluated by accuracy and the slot labeling performance is evaluated by F1. Mean and standard deviation are reported from four independent runs. The results are consistent across different datasets. For slot labeling, the synthetic data can significantly improve the performance especially when training data is extremely low. For example, with only 0.25% of the training data, we improve the slot labeling performance on SNIPS from 45.96 to 63.00, which is a 37.08% relative improvement. Although all generation strategies are useful, conditioning on masked sentences are more effective than others with low training data. We believe that is because the masked sentences can provide more diverse context to synthesize the augmented sentences. It also provides a template so that the generator can leverage the token-label correspondences to generate faithful outputs. We can also see that using augmented language format gives better results than using the factorized generation scheme, which we believe is because of the difficulty of fine-tuning the classification head from scratch with only limited data. Given more real training data, the difference between different generation strategies gets smaller, which is as expected. For intent classification, the trends are similar, but the improvement is relatively small since the intent classification is a relatively simpler task. Using factorized generation is actually competitive, sometimes even better, for intent classification, because the synthetic sentences are generated directly conditioned on the intent labels. Fig F.3 (Appendix) shows some generated sentences from our generator conditioned on augmented sentences masked with multiple spans. The generation is diverse and fluent. Thanks to the powerful pretrained models, our generator is capable of generating phrases beyond the training data. Please see Sec. F.2 for additional studies about our proposed metrics that evaluate the generation quality.

4 Conclusion

In this work, we present a framework to extract prior information from pretrained language models to improve spoken language understanding. We express the prior knowledge in the form of synthetic data and propose to use an augmented language format to generate both sentences and intent and slot labels simultaneously. The generated data as well as the small real dataset are used to fine-tune a classifier to predict intent and slot labels jointly. We also utilize the mixout regularization for classifiers to resist label noise in generated data. On three public benchmark datasets, we achieve superior performance over baselines. For future directions, we will apply the augmented language format for other tasks, such as named entity recognition. We will also explore the classical few-shot setting, where we have other related tasks to accumulate common knowledge.

REFERENCES

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pp. 8141–8150, 2019.

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. Augmented natural language for generative sequence labeling. *arXiv preprint arXiv:2009.13272*, 2020.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, 2018.

Tri Dao, Albert Gu, Alexander J Ratner, Virginia Smith, Christopher De Sa, and Christopher Ré. A kernel theory of modern data augmentation. *Proceedings of machine learning research*, 97:1528, 2019.

Cyprien de Masson d'Autume, Shakir Mohamed, Mihaela Rosca, and Jack Rae. Training language gans from scratch. In *Advances in Neural Information Processing Systems*, pp. 4300–4311, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.

Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. *arXiv preprint arXiv:1712.09482*, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pp. 8527–8537, 2018.

Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pp. 6626–6637, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Wei Hu, Zhiyuan Li, and Dingli Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. *arXiv preprint arXiv:1905.11368*, 2019.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3609–3619, 2019.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. *arXiv preprint arXiv:1909.11299*, 2019.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5447–5456, 2018.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1910–1918, 2017.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. *arXiv preprint arXiv:1903.05566*, 2019.

Sidi Lu, Yaoming Zhu, Weinan Zhang, Jun Wang, and Yong Yu. Neural text generation: Past, present and beyond. *arXiv preprint arXiv:1803.07133*, 2018.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. *arXiv preprint arXiv:2006.13554*, 2020.

David JC MacKay. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992.

Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". In *Advances in Neural Information Processing Systems*, pp. 960–970, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, pp. 5228–5237, 2018.

Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. On accurate evaluation of gans for language generation. *arXiv preprint arXiv:1806.04936*, 2018.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Advances in Neural Information Processing Systems*, pp. 5596–5605, 2017.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 322–330, 2019.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Congying Xia, Chenwei Zhang, Hoang Nguyen, Jiawei Zhang, and Philip Yu. Cg-bert: Conditional text generation with bert for generalized few-shot intent detection. *arXiv preprint arXiv:2004.01881*, 2020.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. *arXiv preprint arXiv:1702.08139*, 2017.

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. Data augmentation for spoken language understanding via joint variational generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 7402–7409, 2019.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? *arXiv preprint arXiv:1901.04215*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pp. 8778–8788, 2018.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. Adversarially regularized autoencoders. In *International Conference on Machine Learning*, pp. 5902–5911, 2018.