WHAT IF I DON'T HAVE IN-DOMAIN UNLABELED DATA FOR SEMI-SUPERVISED LEARNING? WELL, GENERATE SOME!

Xuanli He, Islam H. Nassar & Gholamreza Haffari
Faculty of Information Technology
Monash University
Melbourne, Australia
{xuanli.he1,islam.nassar,gholamreza.haffari}@monash.edu

Jamie Kiros & Mohammad Norouzi
Google Research
Toronto, Canada
{kiros,mnorouzi}@google.com

ABSTRACT

Semi-Supervised Learning (SSL) has seen success in many application domains, but this success often hinges on the availability of in-domain unlabeled data. We present Generative Self-Training (GeST), a simple refinement of SSL algorithms, in particular self-training, which alleviates the need for in-domain unlabeled data. The key idea is to train an unconditional domain-specific generative model, and use it to generate synthetic unlabeled data for SSL. To train strong domain-specific generative models, one fine-tunes generic generative models (trained on open-domain data) on specific domains. GeST enables combining the benefits of large language models and large self-supervised representations; when GPT-2-large is fine-tuned on the inputs of each GLUE task separately and used as the generative model of GeST to self-train RoBERTa-large, we achieve an average improvement of 1.3% over fine-tuned RoBERTa-large, yielding state-of-the-art performance of 90.1% on GLUE dev sets. We also show that knowledge distillation using generated unlabeled data can help bridge the gap between 12- and 6-layer transformers on GLUE tasks.

1 INTRODUCTION

Unlabeled data is abundant in the real world, but domain-specific unlabeled data within the scope of a given machine learning problem is challenging to find. For instance, one cannot easily find in-domain unlabeled data conforming to the input distribution of a specific Natural Language Processing (NLP) task from the GLUE benchmark (Wang et al., 2019). Some NLP tasks require an input comprising a sentence pair with a particular relationship between them or a question-paragraph pair. If domain-specific unlabeled data were available, one could adopt self-training (Yarowsky, 1995) to automatically annotate unlabeled data with pseudo labels to help improve accuracy and robustness of machine learning models. This paper aims to make self-training more universally applicable by leveraging *generated* unlabeled data within self-training.

The dependence of self-training on in-domain unlabeled data has made it hardly applicable to realistic problems without in-domain unlabeled data. To address this challenge, Du et al. (2020) have used nearest neighbor retrieval to harvest in-domain unlabeled data from a large corpus of open-domain text, leading to a successful application of self-training to certain NLP tasks. While retrieval can indeed help find in-domain data for problems with simple inputs, it is not practical for problems with complex input schemes, *e.g.,* sentence pairs with certain relations and tabular data. Accordingly, to our knowledge, no prior work has successfully applied self-training to tasks from the GLUE benchmark that often involve mulit-sentence inputs.

We present Generative Self-Training (GeST), a simple refinement of self-training that alleviates the need for in-domain unlabeled data. The key idea of GeST is to train an unconditional domain-specific generative model, and use it to generate lots of synthetic unlabeled data, useful for self-training. Thus, the difference between self-training and GeST is that self-training uses existing in-domain unlabeled data, annotated with synthetic labels, whereas GeST uses both synthetic unlabeled data and synthetic labels. Building on recent advances in text generation (Radford et al., 2019), we train strong domain-specific generative model for GeST, by fine-tuning an existing generative model that has been pretrained on open-domain data on specific domains.

Our main contributions are summarized as:

- We propose GeST: a novel wrapper around SSL and KD that advocates the use of unconditional generative models to synthesize in-domain unlabeled data for SSL and KD.
- We demonstrate the efficacy of GeST on GLUE benchmark tasks.

2 Generative Self-Training (GeST)

Given a labeled dataset $L = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ and an unlabeled dataset $U = \{\boldsymbol{x}_j\}_{j=1}^{M}$, we summarize the general family of SSL algorithms known as self-training as:

1. First, an initial model denoted f_1 is trained using supervised learning on the labeled dataset L.
2. Then, at iteration t, one adopts f_t as the teacher model to annotate the unlabeled dataset U using *pseudo labels*, denoted as $S_t = \{(x, f_t(x)) \mid x \in U\}$.
3. A student model f_{t+1} is trained to optimize a classification loss on the combination of L and S_t:
$$\ell_{t+1} = \mathbb{E}_{(\boldsymbol{x},y)\sim(L \cup S_t)} H(y, f_{t+1}(\boldsymbol{x})) \,, \tag{1}$$
where $H(q, p) = q^\top \log p$ is the softmax cross entropy loss, and y is assumed to be a one-hot vector (original labels) or a vector of class probabilities (pseudo labels).
4. Self-training iterations are repeated T times or until performance plateaus.

However, the unlabeled in-domain dataset U is usually not available.Thus, given a labeled dataset $L = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, we first train an unconditional domain-specific generative model $g(\boldsymbol{x})$ on $L_x = \{\boldsymbol{x}_i\}_{i=1}^{N}$, and then use it to synthesize unlabeled data U[1]. Such synthetic unlabeled data is used to enable the adoption of self-training even without in-domain unlabeled data. We call this general framework Generative Self-Training (GeST) because it uses generated data within self-training.

3 Experiments

We assess the effectiveness of GeST on GLUE benchmark (Wang et al., 2019) (see Appendix B for benchmark details). To generate domain-specific synthetic data, we fine-tune GPT-2-large on the training set of each downstream task, excluding labels. For tasks with multiple input sentences, we concatenate input sentences into a long sequences and separate sentences by special [SEP] tokens. We generate new domain-specific data by using top-k random sampling similar to Radford et al. (2019). We do not feed any prompt to the LM, but a special [BOS] token to initiate the generation chain. A generation episode is terminated when a special [EOS] token is produced. We generate diverse sentences by varying the random seed. After collecting enough synthetic data, we only retain unique sentences. For tasks with α input sentences, we discard generated samples that violate this constraint (approximately 10% of samples were rejected). Our final synthetic unlabeled dataset U includes $40\times$ as many examples as the original dataset for each task.

GeST. We fine-tune pretrained RoBERTa models provided by fairseq (Ott et al., 2019) on each task. Fine-tuned RoBERTa serves as the first teacher model for self-training. Each student model is initialized with the original pretrained RoBERTa. We combine the labeled dataset L and the synthetic dataset U with a ratio of 1:1, by oversampling labeled data.

Table 1 shows that GeST provides an average improvement of +1.2% over RoBERTa-base. We see consistent improvements with more GeST iterations, but performance saturates after three iterations. Finally, we apply 3 iterations of GeST to RoBERTa-large and compare with state-of-the-art techniques in Table 2. We observe that RoBERTa-large + GeST outperforms strong recent techniques in terms of average performance on the GLUE tasks.

[1] The detailed algorithm can be found in Appendix C

Model	SST-2	QQP	QNLI	RTE	MNLI	MRPC	CoLA	STS-B	Avg
RoBERTa base	94.8	91.5	92.6	78.8	87.7	90.1	63.6	90.8	86.2
+ GeST (iter 1)	95.3	91.8	93.1	81.4	87.9	91.7	65.1	91.4	87.2
+ GeST (iter 2)	95.3	91.7	93.2	82.4	88.0	92.2	65.2	91.5	**87.4**
+ GeST (iter 3)	95.3	91.7	93.2	82.0	87.9	92.2	65.5	91.7	**87.4**

Table 1: RoBERTa base and GeST results with few iterations on GLUE dev sets. Reported results are the average of 5 independent runs.

Model	SST-2	QQP	QNLI	RTE	MNLI	MRPC	CoLA	STS-B	Avg
BERT	93.2	91.3	92.3	70.4	86.6	88.0	60.6	90.0	84.1
RoBERTa	96.4	92.2	93.9	86.6	90.2	90.9	68.0	92.4	88.8
XLNET	**97.0**	92.3	94.9	85.9	90.8	90.8	69.0	92.5	89.2
ELECTRA	96.9	**92.4**	95.0	88.0	90.9	90.8	69.1	92.6	89.5
DeBERTa	96.8	92.3	**95.3**	88.3	**91.1**	91.9	70.5	**92.8**	89.9
RoBERTa + GeST	96.9	92.1	94.7	**90.1**	90.7	**93.0**	**70.8**	92.2	**90.1**

Table 2: RoBERTa large and GeST results (average of 5 runs) on GLUE dev sets in comparison with strong recent baselines: BERT large (Devlin et al., 2019), RoBERTa large (Liu et al., 2019b), XLNET large (Yang et al., 2019), ELECTRA large (Clark et al., 2020), DeBERTa large (He et al., 2020),

In what follows, we conduct an in-depth ablation of different components of GeST. Unless stated otherwise, we use a RoBERTa-base model with a combination of the original training data and $40\times$ synthetic data for each experiment.

Synthetic dataset size. Deep neural networks typically benefit from large training datasets (Koehn & Knowles, 2017). Because we use a generative model to synthesize data, we can use as much synthetic data as practically possible given our computational budget. To investigate the impact of synthetic dataset size on GeST, we vary the synthetic dataset size from $1\times$ to $40\times$ of the labeled dataset. We also study the use of synthetic data only, without mixing it with the original labeled dataset. Table 3 shows that for both GeST and synthetic data only settings, larger synthetic datasets translate to better performance. On the other hand, the use of synthetic data only, without mixing in the labeled dataset, does not consistently outperform the RoBERTa baseline.

Setup	SST-2	RTE	MRPC	CoLA
RoBERTa base	94.8	78.8	90.1	63.6
Synthetic-only $1\times$	94.9	73.1	88.7	56.1
Synthetic-only $5\times$	94.9	76.5	90.0	59.1
Synthetic-only $10\times$	95.0	77.6	91.1	59.2
Synthetic-only $40\times$	95.1	80.3	90.7	59.9
GeST $1\times$	95.3	79.1	90.0	63.6
GeST $5\times$	95.3	80.5	91.0	64.9
GeST $10\times$	95.2	80.5	91.3	65.0
GeST $40\times$	95.3	81.4	91.7	65.1

Table 3: The impact of synthetic dataset size on GLUE dev set results. Synthetic dataset size is $k\times$ of the original dataset. GeST leverages both synthetic unlabeled data and labeled data.

Soft v.s. hard pseudo label. We investigate the use of soft and hard pseudo labels within the GeST framework. The results in Table 4 suggest that GeST using soft pseudo labels is more effective than hard labels on the GLUE benchmark. This finding is compatible with the intuition that soft labels enable measuring the functional similarity of neural networks better (Hinton et al., 2015).

Pseudo label	SST-2	RTE	MRPC	CoLA
hard	95.0	80.7	90.8	63.0
soft	95.3	81.4	91.7	65.1

Table 4: GeST with soft *v.s.* hard pseudo labels on GLUE dev sets.

Class-conditional synthetic data generation. Previous work (Kumar et al., 2020) suggests that it is challenging to utilize synthetic data from class-conditional generative models to boost the accuracy of text classifiers. We also study this phenomenon, by fine-tuning GPT-2 in a class-conditional manner. Table 5 shows that not only class-conditional LMs underperform unconditional LMs (GeST), but also they are much worse than the baseline.

Source of synthetic data	SST-2	RTE	MRPC	CoLA
No synthetic data (baseline)	94.8	78.8	90.1	63.6
Class-conditional LM	92.9	74.4	86.0	58.4
Unconditional LM (GeST)	95.3	81.4	91.7	65.1

Table 5: Synthetic data from class-conditional LMs underperforms GeST and original RoBERTa base on GLUE dev sets.

GPT-2 model size. Radford et al. (2019) present a few variants of the GPT-2 model including *GPT-2*, *GPT-2-medium*, and *GPT-2-large*. Larger GPT-2 models yield better perplexity scores and higher generation quality. We utilize these models within the GeST framework to study the impact of the generative model's quality on downstream task's performance. Table 6 shows that SST-2 and RTE datasets are not sensitive to the capacity of the GPT-2 model, but higher quality synthetic text improves the results on MRPC and CoLA datasets.

GPT-2	SST-2	RTE	MRPC	CoLA
small	95.5	81.3	90.9	63.9
medium	95.3	81.3	91.3	63.7
large	95.3	81.4	91.7	65.1

Table 6: GeST with various GPT-2 model sizes on GLUE dev sets.

Knowledge distillation. The goal of knowledge distillation (KD) (Buciluǎ et al., 2006; Hinton et al., 2015) is to distill the knowledge of a powerful teacher model into a compact student model with as little loss in performance as possible. This can help with model compression (Jiao et al., 2019; Sun et al., 2019) and multi-task learning (Liu et al., 2019a; Clark et al., 2019).

Model	Data	SST-2	QQP	QNLI	RTE	MNLI	MRPC	CoLA	STS-B	Avg
BERT base	Orig.	93.2	89.7	91.6	67.1	84.6	87.9	58.3	88.1	82.6
DistilBERT	Orig.	91.1	88.7	88.4	60.3	82.4	87.7	52.8	86.8	79.8
BERT-PKD	Orig.	91.3	88.4	88.4	66.5	81.3	85.7	45.5	86.2	79.2
BERT-Thes.	Orig.	91.5	89.6	89.5	68.2	82.3	**89.0**	51.1	**88.7**	81.2
DistilBERT	GeST	**92.1**	**89.7**	**90.6**	**70.4**	**83.6**	88.6	**56.6**	88.1	**82.5**

Table 7: Knowledge Distillation results on GLUE dev sets with different models. All models use 6-layer transformer, except BERT base. Orig. indicates the original training data, and BERT-Thes. is BERT-Theseus (Xu et al., 2020).

We use the HuggingFace implementation (Wolf et al., 2020) for KD experiments and adopt a standard experimental setup consistent with previous work (Sun et al., 2019; Xu et al., 2020). A fine-tuned BERT base model (12-layer transformer) (Devlin et al., 2019) represents the teacher and a DistilBERT model (6-layer transformer) (Sanh et al., 2019) is used as the student. Similar to GeST, we train the student model on U and L, where U is annotated by a fixed teacher. Table 7 shows that GeST dramatically surpasses all existing KD baselines, including DistilBERT (Sanh et al., 2019), BERT-PKD (Sun et al., 2019) and BERT-Theseus (Xu et al., 2020). All of the baselines use the same student architecture. This marks a new state-of-the-art for KD on GLUE benchmark.

4 Conclusion

We present Generative Self-Training (GeST): a framework for self-training with generated unlabeled data. We demonstrate that GeST leverages advances in deep generative models to help supervised learning and can have implications for learning from limited labeled data. Particularly, the proposed approach works surprisingly well on dev sets of GLUE benchmark and helps improve knowledge distillation. We hope that GeST will stimulate new research on the evaluation and development of deep generative models.

REFERENCES

A. Agrawala. Learning with a probabilistic teacher. *IEEE Transactions on Information Theory*, 16 (4):373–379, 1970. doi: 10.1109/TIT.1970.1054472.

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

Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv:1810.10863*, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

TB Brown, B Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, et al. Language models are few-shot learners. arxiv 2020. *arXiv:2005.14165*, 4, 2020.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT, 2009.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D Manning, and Quoc Le. Bam! born-again multi-task networks for natural language understanding. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5931–5937, 2019.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *International Conference on Learning Representations*, 2020.

David B Cooper and John H Freeman. On the asymptotic improvement in the out-come of supervised learning provided by additional nonsupervised learning. *IEEE Transactions on Computers*, 1970.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *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. 4171–4186, 2019.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv:1910.08854*, 2019.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. Self-training improves pre-training for natural language understanding. *arXiv:2010.02194*, 2020.

S Fralick. Learning to recognize patterns without a teacher. *IEEE Transactions on Information Theory*, 1967.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv:2012.15723*, 2020.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv:2006.03654*, 2020.

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. *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:1503.02531*, 2015.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv:1909.10351*, 2019.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39, 2017.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv:2003.02245*, 2020.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv:1904.09482*, 2019a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*, 2019b.

Geoffrey J McLachlan and S Ganesalingam. Updating a discriminant function on the basis of unclassified data. *Communications in Statistics-Simulation and Computation*, 1982.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, 2019.

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

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in Neural Information Processing Systems*, pp. 12268–12279, 2019.

Ellen Riloff. Automatically generating extraction patterns from untagged text. *Proceedings of the national conference on artificial intelligence*, pp. 1044–1049, 1996.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2234–2242, 2016.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

H Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 1965.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4314–4323, 2019.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *International Conference on Learning Representations*, 2019.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Kdgan: Knowledge distillation with generative adversarial networks. *NeurIPS*, 2018.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, October 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2020.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. Bert-of-theseus: Compressing bert by progressive module replacing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7859–7869, 2020.

I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv:1905.00546*, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763, 2019.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. *33rd annual meeting of the association for computational linguistics*, pp. 189–196, 1995.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in Neural Information Processing Systems*, 32:9054–9065, 2019.

Xiaofeng Zhang, Zhangyang Wang, Dong Liu, and Qing Ling. Dada: Deep adversarial data augmentation for extremely low data regime classification. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2807–2811, 2019.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems*, 33, 2020.

A Related Work

Semi-supervised learning (SSL) has received considerable attention over the last few decades (Cooper & Freeman, 1970; McLachlan & Ganesalingam, 1982; Riloff, 1996; Chapelle et al., 2009; Van Engelen & Hoos, 2020). One of the oldest family of SSL algorithms is known as *self-training*, *a.k.a.* self-learning or self-labeling (Scudder, 1965; Fralick, 1967; Agrawala, 1970; Yarowsky, 1995). The main intuition of self-training is to encourage knowledge transfer between a *teacher* and a *student* model in such a way that the student can outperform the teacher. Specifically, one leverages the teacher's knowledge to annotate unlabeled data with so-called *pseudo labels*, and the student learns from a mixture of pseudo- and human-labeled data. Self-training has seen a surge of recent interest across vision and NLP applications (Yalniz et al., 2019; Xie et al., 2020; Zoph et al., 2020; Du et al., 2020).

Knowledge Distillation (KD) (Buciluǎ et al., 2006; Hinton et al., 2015) uses a procedure similar to self-training to distill knowledge of an expressive teacher model into a smaller student model. Previous work uses unlabeled data (Buciluǎ et al., 2006) and adversarial training (Wang et al., 2018) to improve KD. We demonstrate that synthetic data generated by unconditional generative models can improve KD on NLP, outperforming strong baselines (*e.g.,* Xu et al. (2020)).

Advanced generative models are able to generate realistic images and text (Karras et al., 2017; Brock et al., 2019; Karras et al., 2019; Radford et al., 2019; Brown et al., 2020). The quality of synthetic samples has improved to the extent that deep fake detection has become an important research topic itself (Zellers et al., 2019; Dolhansky et al., 2019). Recent work has aimed to utilize class-conditional generative models to help improve supervised learning (Antoniou et al., 2017; Bowles et al., 2018; Zhang et al., 2019; Kumar et al., 2020; Gao et al., 2020). However, Ravuri & Vinyals (2019) have shown that images generated by state-of-the-art class-conditional generative models fall short of improving ImageNet classification accuracy, despite strong sample quality scores (Salimans et al., 2016; Heusel et al., 2017). Similarly, Kumar et al. (2020) find that it is difficult for sentences generated by label-conditioned GPT-2 (Radford et al., 2019) to retain the semantics or pragmatics of a specified category, which leads to poor performance on downstream tasks.

B Datasets

Dataset	task	domain	#train	#dev	#test	#classes
SST-2	sentiment analysis	movie reviews	67k	872	1.8k	2
QQP	paraphrase	social QA questions	364k	40k	391k	2
QNLI	QA/natural language inference	Wikipedia	105k	5k	5.4k	2
RTE	natural language inference	news, Wikipedia	2.5k	277	3k	2
MNLI	natural language inference	misc.	393k	20k	20k	3
MRPC	paraphrase	news	3.7k	408	1.7k	2
CoLA	acceptability	misc.	8.5k	1043	1k	2
STS-B	sentence similarity	misc.	5.8k	15k	1.4k	–

Table 8: Summary of the tasks used for evaluation of GeST. STS-B is a regression task, so #classes is not applicable.

C GeST Algorithm

The algorithms of the generic self-training and GeST are summarized in algorithm 1 and algorithm 2.

D Training Details

We use the fairseq codebase Ott et al. (2019) for self-training experiments. Training details are summarized in Table 9. We use the HuggingFace codebase (Wolf et al., 2020) for KD experiments. All models are trained for 5 epochs with a learning rate of 2e-5 and a batch size of 32.

Algorithm 1: $\mathrm{SelfTraining}(L, U, f_0, T)$

Input: Labeled dataset $L = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$
Unlabeled dataset $U = \{\boldsymbol{x}_j\}_{j=1}^{M}$
Initial parameters of a classifier f_0
Output: A better classifier f_{T+1} after T self-training steps
1: train a base model f_1 by fine-tuning f_0 on L
2: **for** $t = 1$ to T do:
3: apply f_t to unlabeled instances of U to obtain $S_t = \{(x, f_t(x)) \mid x \in U\}$
4: train a new model f_{t+1} by either fine-tuning f_0 on $L \cup S_t$
5: **return** f_{T+1}

Algorithm 2: $\mathrm{GeST}(L, g_0, f_0, k, T)$

Input: Labeled dataset $L = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$
Initial parameters of a generative model g_0
Initial parameters of a classifier f_0
Output: A better classifier f_{T+1} after T GeST steps
1: train a generative model g by fine-tuning g_0 on L_x where $L_x = \{\boldsymbol{x} \mid (\boldsymbol{x}, y) \in L\}$
2: generate $U = \{\widetilde{\boldsymbol{x}}_j\}_{j=1}^{kN}$ by drawing kN random samples *i.i.d.* from $g(\boldsymbol{x})$, *i.e.,* $\widetilde{\boldsymbol{x}}_j \sim g(\boldsymbol{x})$ for $j = 1$ to kN.
3: **return** $\mathrm{SelfTraining}(L, U, f_0, T)$

	SST-2	**QQP**	**QNLI**	**RTE**	**MNLI**	**MRPC**	**CoLA**	**STS-B**
lr	1e-5	1e-5	1e-5	2e-5	1e-5	1e-5	1e-5	2e-5
#sent.	32	32	32	16	32	16	16	16
warmup steps	1256	28318	1986	122	7432	137	320	214
validate steps	2093	11307	3310	203	12386	203	535	360
#epoch	10	10	10	10	10	10	10	10

Table 9: Training details for GLUE tasks.

E GENERATED UNLABELED EXAMPLES ANNOTATED WITH PSEUDO LABELS

When did the third Digimon series begin? [SEP] Unlike the two seasons before it and most of the seasons that followed, Digimon Tamers takes a darker and more realistic approach to its story featuring Digimon who do not reincarnate after their deaths and more complex character development in the original Japanese. (**not entailment**)
KNN: 1: What is the name of the third season? [SEP] In addition to the first two seasons, the third season is the season that introduced new characters such as Captain Malice, a supervillain who became the antagonist in season two; and the villains known as the Heartbreakers, who introduced a group of crime fighters. (**not entailment**) 2: When did the "Walking Dead" series end? [SEP] In 2013, AMC announced that it would develop a "superhero series", which would follow the storylines and characters from the "Walking Dead" series in order to bring the popular AMC original series to a new and younger audience. (**not entailment**) 3: What is the main objective of the first season of the X-Files? [SEP] The first season was notable in that the characters were introduced and developed within the space of a single season, as was the format of the show itself. (**not entailment**)
What did Arsenal consider the yellow and blue colors to be after losing a FA Cup final wearing red and white? [SEP] Arsenal then competed in three consecutive FA Cup finals between 1978 and 1980 wearing their "lucky" yellow and blue strip, which remained the club's away strip until the release of a green and navy away kit in 1982–83. (**entailment**)
KNN: 1: Who was the most important player for Arsenal Football Club in the 1950s? [SEP] Wenger continued to use Arsenal's famous red shirts and red kits throughout the 1950s and 1960s, and the red strip became the club's most recognised and recognizable symbol. (**not entailment**) 2: When were the first two teams to play for the trophy in the Premier League? [SEP] The trophy was awarded to Manchester United in 1990-91 and was named after Sir Bobby Charlton, the club's manager until 1990, and later Sir Stanley Matthews, the club's most successful manager. (**not entailment**) 3: What were the last four players to wear the yellow in the final? [SEP] With Arsenal having won all four major trophies in the period, they became the only club to have won five in a row. (**not entailment**)

Table 10: **QNLI**: Two labeled examples, along with 3 nearest neighbors (based on RoBERTa representations) from our synthetic dataset. We include **labels** for original examples and **pseudo-labels** for synthetic examples in parenthesis.

are more deeply thought through than in most ‘ right-thinking ’ films (**positive**)
KNN: 1: is far more sophisticated , insightful and thought-provoking than his previous films . (**positive**) 2: is more sophisticated than its more obvious and less-than-dazzling counterparts (**positive**) 3: is about as well-thought as the idea of a bad hair day , (**negative**)
contains no wit , only labored gags (**negative**)
KNN: 1: lacks insight , and lacks empathy (**negative**) 2: has little humor or intelligence (**negative**) 3: lacks all wit and humanity (**negative**)

Table 11: **SST-2**: Two labeled examples, along with 3 nearest neighbors (based on RoBERTa representations) from our synthetic dataset. We include **labels** for original examples and **pseudo-labels** for synthetic examples in parenthesis.

Like the United States, U.N. officials are also dismayed that Aristide killed a conference called by Prime Minister Robert Malval in Port-au-Prince in hopes of bringing all the feuding parties together. [SEP] Aristide had Prime Minister Robert Malval murdered in Port-au-Prince. (**not entailment**)
KNN: 1: The government has been criticized for failing to prevent the mass protests that led to the ouster of President Nicolas Sarkozy earlier this month, which led to his second election defeat since assuming office two years ago. [SEP] Prime Minister Jean-Marc Ayrault is a former president of France. (**not entailment**) 2: The French president, Jacques Chirac, has been urged by both the Vatican and the U.N. Security Council to step up efforts to prevent the return of former dictator Nicolas Sarkozy. [SEP] Nicolas Sarkozy left France. (**not entailment**) 3: The French newspaper Le Monde says the French President Nicolas Sarkozy was advised by U.S. President George W. Bush about a possible trip to Iraq on Thursday. [SEP] Nicolas Sarkozy is a member of the United States. (**not entailment**)
Only a week after it had no comment on upping the storage capacity of its Hotmail e-mail service, Microsoft early Thursday announced it was boosting the allowance to 250MB to follow similar moves by rivals such as Google, Yahoo, and Lycos. [SEP] Microsoft's Hotmail has raised its storage capacity to 250MB. (**entailment**)
KNN: 1: The company, known as Microsoft Office, said it plans to sell all of the copies of its popular Office suite at a loss in the wake of the launch of Microsoft Windows 7, saying it will also make $25 million in advertising costs, a move likely to hurt its long-standing position among consumers and business leaders. [SEP] Microsoft Office is a popular office suite. (**entailment**) 2: The company's shares shot up more than 35% after the company said it has sold all of its remaining inventory of the new Kindle e-readers at $70 each. The shares rose to $65.20 on Wednesday, their highest since March 6, 2011. "The Kindle is our best selling product," said Jeff Bezos, founder and CEO of Amazon.com. [SEP] Amazon.com is based in Seattle. (**not entailment**) 3: In response to concerns expressed by some investors, Microsoft last week said it would reduce the amount of shares that will be available to the public by 10 percent in the first quarter, with a further reduction to 3 percent in the second quarter. The stock price has plunged from $24 to $17, and Microsoft is currently offering $17 to $19 a share to its most senior employees. Some investors had criticized Microsoft's response to concerns about the price of its stock and about the perception that the company is in trouble. [SEP] Microsoft is struggling to sell its stock. (**not entailment**)

Table 12: **RTE**: Two labeled examples, along with 3 nearest neighbors (based on RoBERTa representations) from our synthetic dataset. We include **labels** for original examples and **pseudo-labels** for synthetic examples in parenthesis.

How is the life of a math student? Could you describe your own experiences? [SEP] Which level of prepration is enough for the exam jlpt5? (**not duplicated**)
KNN: 1: What are the best courses for a mechanical engineering student? [SEP] What is the best course to do after completing a B.Tech in mechanical engineering? (**not duplicated**) 2: How much marks are needed to get through the GATE with electronics? [SEP] What is the average score of the Gate EE exam? What are the cut-offs? (**not duplicated**) 3: What is the best time table for students to prepare for IAS? [SEP] How can one study for IAS in a best time? (**not duplicated**)
How does an IQ test work and what is determined from an IQ test? [SEP] How does IQ test works? (**duplicated**)
KNN: 1: What is the average IQ of the U.S. population? [SEP] How does an IQ test work? (**not duplicated**) 2: Is the Iq test an effective way to measure intelligence? [SEP] How do IQ tests work? (**duplicated**) 3: How is an IQ test on a scale from 1 to 100 scored? [SEP] How do you get your IQ tested? (**not duplicated**)

Table 13: **QQP**: Two labeled examples, along with 3 nearest neighbors (based on RoBERTa representations) from our synthetic dataset. We include **labels** for original examples and **pseudo-labels** for synthetic examples in parenthesis.

A BMI of 25 or above is considered overweight ; 30 or above is considered obese . [SEP] A BMI between 18.5 and 24.9 is considered normal , over 25 is considered overweight and 30 or greater is defined as obese . (**paraphrase**)
KNN: 1: The report said that the average woman in her twenties who takes oral contraceptives daily can expect a loss of around 40 per cent of her bone density between the ages of 20 and 45 . [SEP] The study said the average woman in her twenties who used the pill every day , or every day for up to five years , can expect a loss of about 40 per cent of her bone density between the ages of 20 and 45 . (**paraphrase**) 2: The report found that 17 percent of U.S. adults between ages 18 and 64 have a body mass index at or above the " normal " 20 . [SEP] For people of that age , 17.1 percent of adults have a body mass index at or above the " normal " 20 , while 12.6 percent have a body mass index of 30 or above . (**not paraphrase**) 3: The survey shows the proportion of women between 20 and 44 who were obese was 6.3 percent , up from 5.7 percent in 2001 . [SEP] The proportion of women between 20 and 44 who were obese increased to 6.3 percent from 5.7 percent in 2001 . (**paraphrase**)
Shares of Genentech , a much larger company with several products on the market , rose more than 2 percent . [SEP] Shares of Xoma fell 16 percent in early trade , while shares of Genentech , a much larger company with several products on the market , were up 2 percent .(**not paraphrase**)
KNN: 1: Shares in Aventura fell as much as 5 percent , while shares in Medi-Cal climbed 2.5 percent . [SEP] Shares in Aventura were up 2.5 percent , while shares in Medi-Cal rose 2.5 percent . (**paraphrase**) 2: Shares of Amgen rose $ 2.29 , or 2.2 percent , to $ 41.10 in after-hours trading . [SEP] Shares of Amgen , a division of Sanofi-Aventis , rose $ 1.62 , or 1.6 percent , to $ 41.06 in after-hours trading .(**paraphrase**) 3: Shares of General Electric Co . GE.N rose more than 6 percent on the New York Stock Exchange , while shares of PepsiCo Inc . PEP.N rose 4.7 percent . [SPE] General Electric 's shares jumped almost 6 percent on the New York Stock Exchange , while PepsiCo 's climbed 4.7 percent . (**paraphrase**)

Table 14: **MRPC**: Two labeled examples, along with 3 nearest neighbors (based on RoBERTa representations) from our synthetic dataset. We include **labels** for original examples and **pseudo-labels** for synthetic examples in parenthesis.

One of our number will carry out your instructions minutely. [SEP] A member of my team will execute your orders with immense precision. (**entailment**)
KNN: 1: We are at your disposal to help you with your investigation and provide a full range of pro bono services. [SEP] We are the only ones who can help you with your investigation. (**neutral**) 2: I will speak with the chief officer of the contractor, who will be informed about the results of this effort. [SEP] The contractor is being informed about the results of the effort. (**entailment**) 3: We have an office here to assist you. [SEP] An office is where we will assist you, said the manager. (**neutral**)
Conceptually cream skimming has two basic dimensions - product and geography. [SEP] Product and geography are what make cream skimming work. (**neutral**)
KNN: 1: There are two main types of analysis and they are the case study and the case report. [SEP] The case study is the most popular method used to analyze a subject. (**neutral**) 2: A third approach to capturing and using this type of experience is to engage the program management and finance systems of the organization. [SEP] There are two strategies to capturing and using experience. (**contradiction**) 3: The first is to see the basic elements of a business model in action. [SEP] Basic elements of business models are the most important for the success of any company. (**neutral**)
I don't mean to be glib about your concerns, but if I were you, I might be more concerned about the near-term rate implications of this $1. [SEP] I am concerned more about your issues than the near-term rate implications. (**contradiction**)
KNN: 1: I'm not here to tell you of my own experiences, but they are important to others who might have similar concerns. [SEP] If you were to have similar concerns, I'd like to encourage you to tell them to me. (**neutral**) 2: I don't mean to sound judgmental, but as a person, I think that's an issue you're probably pretty much on your own if you think about it. [SEP] You're probably right if you think about it. (**neutral**) 3: But I don't mean to take your word for it. [SEP] I know you are correct, but I want to make sure it's clear that I do not agree. (**contradiction**)

Table 15: **MNLI**: Two labeled examples, along with 3 nearest neighbors (based on RoBERTa representations) from our synthetic dataset. We include **labels** for original examples and **pseudo-labels** for synthetic examples in parenthesis.