STraTA: Self-Training with Task Augmentation for Better Few-shot Learning

Tu Vu, Minh-Thang Luong , Quoc V. Le , Grady Simon , Mohit Iyyer
EMNLP 2021

Motivation

Large generative model (e.g. GPT3) still lags behind well-finetuned model

This paper proposes STraTa to leverage unlabeled data:

- Self-Training
- Task Augmentation

Framework

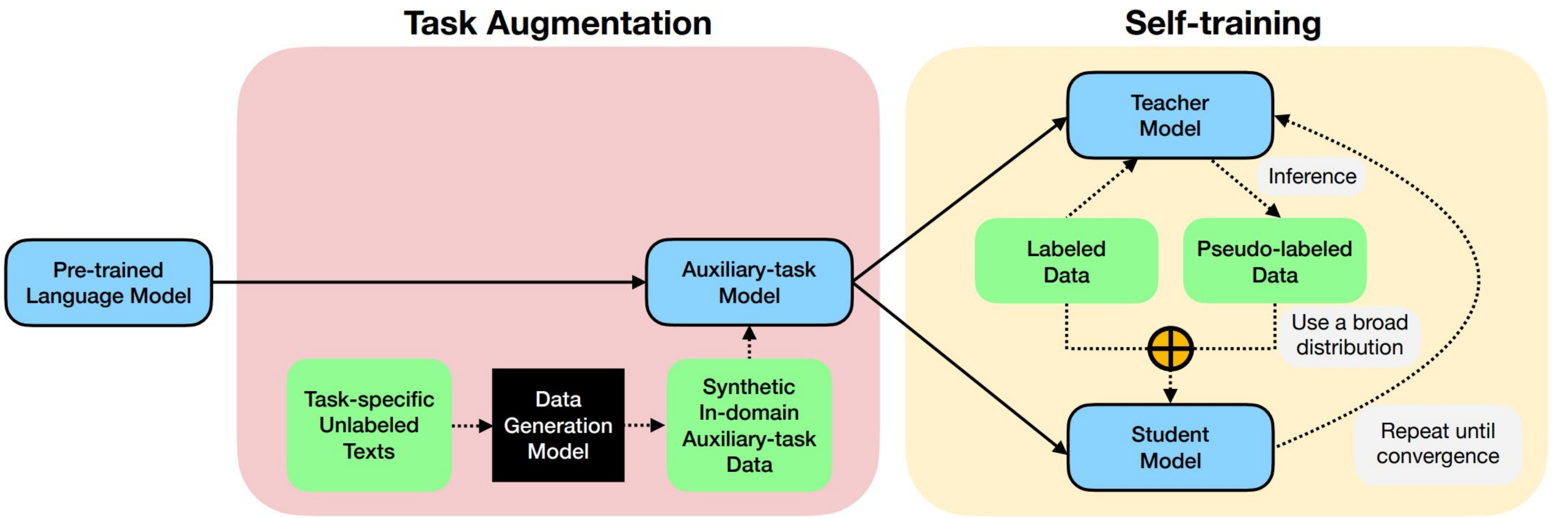

Figure 2: An illustration of our **S**elf-**Tra**ining with **T**ask **A**ugmentation (**STraTA**) approach. In task augmentation, we train an NLI data generation model and use it to synthesize a large amount of *in-domain* NLI training data for each given target task, which is then used for auxiliary (intermediate) fine-tuning. Our self-training algorithm iteratively learns a better model using a concatenation of labeled and pseudo-labeled examples. At each iteration, we always start with the auxiliary-task model produced by task augmentation and train on a broad distribution of pseudo-labeled data.

Task Augmentation

- Goal:
  - **Finetune LMs on auxiliary task before the target task**
- Which task?
  - Domain mismatch: MNLI and SQuAD
- In this paper:
  - Finetune on NLI data

Task Augmentation (2)

- Synthetic NLI data generation
    - Generate from MNLI
        - (sentA, sentB)-> label
    - To NLI
        - (label, sentA) -> sentB
    - Finetune T5 model
- Benefit
    - Training label is free
    - Be able to overgeneration in-domain NLI training data
- Overgeneration
    - Generate 100 output samples per input (top-k=40)
- Filtering
    - Use BERT model finetuned on MNLI as NLI classifier
    - Filter synthetic data:
        - If produce the same label -> kept
        - Otherwise -> Remove

Self-training

- Goal
  - Improve the model using pseudo-labeled data
- **A strong base model**
  - Strong base model can produce quite good pseudo labeled data
  - Otherwise, errors can be propagated through incorrect data
- Steps
  - Starts with the same strong base model
  - Finetune all parameters using labeled and pseudo-labeled data
  -

Self-training (2)

- **Self-training on a broad distribution of pseudo-labeled data**
- Confident data ~ narrow distribution
  - Overconfident teacher
  - Poorly calibrated
  - Harmful to the self-training
- Less confident data ~ broad distribution
  - Using the whole set of pseudo-labeled data
  - At each iteration, the pseudo-labels are regenerated as the teacher improves gradually

Experiment

- LMFT: Target task language model finetune
  - Finetune LM on in-domain unlabeled text using MLM signal
  - Finetune LM on the target task
- ITFT_MNLI: intermediate task finetuning on MNLI
  - Finetune LM on MNLI
  - Finetune LM on target task
- TA: task augmentation
- ST: self-training
-

Model	SNLI	QQP	QNLI	SST-2	SciTail	SST-5	STS-B	SICK-E	SICK-R	CR	MRPC	RTE
						Full						
$\text{BERT}_{\text{LARGE}}$	91.1	88.4	91.9	92.4	95.3	$53.7_{0.9}$	$89.6_{0.2}$	$87.9_{0.6}$	$84.4_{0.4}$	$91.7_{0.6}$	$89.0_{0.8}$	$68.6_{7.2}$
+ LMFT	91.0	88.1	90.4	93.5	95.3	$54.0_{0.4}$	$89.5_{0.2}$	$87.7_{0.5}$	$84.0_{0.5}$	$91.6_{0.8}$	$89.5_{1.0}$	$66.5_{7.3}$
+ $\text{ITFT}_{\text{MNLI}}$	91.1	88.2	91.6	93.5	96.5	$54.0_{0.8}$	$90.3_{0.3}$	$89.9_{0.2}$	$86.3_{0.3}$	$92.0_{0.6}$	$89.7_{0.9}$	$82.3_{1.4}$
+ TA	**91.9**	**88.5**	**92.5**	**94.7**	**96.9**	$\mathbf{55.7}_{0.8}$	$\mathbf{90.9}_{0.2}$	$\mathbf{90.7}_{0.3}$	$\mathbf{87.0}_{0.3}$	$\mathbf{93.3}_{0.6}$	$\mathbf{90.8}_{0.7}$	$\mathbf{83.8}_{1.1}$
						Limited *(1024 total training examples)*						
$\text{BERT}_{\text{LARGE}}$	$77.4_{0.6}$	$74.1_{1.0}$	$81.7_{0.9}$	$89.8_{0.6}$	$90.9_{0.7}$	$49.1_{1.3}$	$88.2_{0.4}$	$84.8_{0.7}$	$80.2_{0.4}$	$91.2_{0.6}$	$85.7_{1.7}$	$66.8_{2.7}$
+ LMFT	$75.8_{1.5}$	$71.6_{0.5}$	$80.5_{2.0}$	$88.9_{0.8}$	$87.7_{2.3}$	$49.2_{3.1}$	$88.4_{0.4}$	$83.2_{0.6}$	$78.5_{0.6}$	$90.9_{0.7}$	$84.9_{1.1}$	$65.2_{3.4}$
+ $\text{ITFT}_{\text{MNLI}}$	$85.2_{0.4}$	$74.0_{0.5}$	$83.5_{0.5}$	$90.0_{0.8}$	$92.1_{1.1}$	$49.4_{1.2}$	$87.8_{0.8}$	$88.8_{0.5}$	$83.2_{0.7}$	$91.3_{0.7}$	$86.4_{0.9}$	$81.1_{1.3}$
+ TA	$\mathbf{87.3}_{0.3}$	$\mathbf{75.7}_{0.5}$	$\mathbf{85.0}_{0.5}$	$\mathbf{91.7}_{0.7}$	$\mathbf{92.3}_{1.1}$	$\mathbf{51.4}_{1.0}$	$\mathbf{89.0}_{0.6}$	$\mathbf{89.4}_{0.4}$	$\mathbf{84.3}_{0.4}$	$\mathbf{92.6}_{0.6}$	$\mathbf{88.0}_{0.8}$	$\mathbf{82.9}_{1.8}$
						Few-Shot *(8 training examples per class)*						
$\text{BERT}_{\text{BASE}}$	$43.7_{2.2}$	$55.9_{6.5}$	$59.0_{10.9}$	$59.1_{8.4}$	$67.1_{6.6}$	$30.5_{2.0}$	$73.6_{4.5}$	$61.3_{4.1}$	$59.7_{2.7}$	$65.2_{8.2}$	$72.4_{10.2}$	$51.4_{2.5}$
+ LMFT	$45.2_{3.9}$	$57.2_{6.2}$	$57.6_{9.1}$	$64.9_{8.7}$	$64.0_{8.0}$	$33.4_{1.9}$	$75.4_{4.4}$	$59.3_{4.0}$	$58.3_{2.0}$	$72.4_{6.0}$	$73.9_{8.6}$	$50.9_{3.9}$
+ $\text{ITFT}_{\text{MNLI}}$	$75.2_{5.7}$	$63.7_{7.0}$	$62.8_{5.1}$	$76.8_{7.2}$	$75.8_{5.6}$	$35.0_{2.6}$	$80.2_{1.1}$	$80.4_{1.9}$	$73.5_{2.7}$	$79.2_{3.6}$	$74.3_{8.0}$	$62.2_{13.5}$
+ TA	$83.3_{0.8}$	$68.7_{1.5}$	$70.1_{3.4}$	$80.3_{6.6}$	$78.5_{3.2}$	$37.4_{3.0}$	$80.7_{1.5}$	$81.1_{2.4}$	$75.9_{1.8}$	$86.5_{2.2}$	$74.5_{6.5}$	$67.6_{7.1}$
+ ST	$65.0_{5.8}$	$69.9_{5.9}$	$71.6_{11.3}$	$62.7_{10.4}$	$68.6_{8.3}$	$33.9_{3.5}$	$80.5_{2.2}$	$68.1_{4.5}$	$64.0_{2.4}$	$78.2_{6.3}$	$80.5_{1.8}$	$50.7_{3.1}$
+ $\text{ITFT}_{\text{MNLI}}$ + ST	$83.2_{0.3}$	$70.7_{5.9}$	$81.5_{1.2}$	$88.0_{2.1}$	$83.7_{4.4}$	$39.5_{2.0}$	$84.2_{0.8}$	$81.8_{2.6}$	$75.8_{2.2}$	$85.6_{2.3}$	$80.6_{1.2}$	$62.5_{12.0}$
+ STraTA	$\mathbf{85.7}_{0.2}$	$\mathbf{74.5}_{0.4}$	$\mathbf{82.1}_{0.5}$	$\mathbf{90.1}_{0.8}$	$\mathbf{86.3}_{3.5}$	$\mathbf{41.3}_{1.5}$	$\mathbf{84.7}_{0.5}$	$\mathbf{84.9}_{1.2}$	$\mathbf{77.6}_{1.6}$	$\mathbf{90.5}_{0.8}$	$\mathbf{81.0}_{0.8}$	$\mathbf{70.6}_{2.4}$
$\text{BERT}_{\text{LARGE}}$	$43.1_{4.4}$	$58.5_{4.7}$	$64.4_{6.1}$	$66.1_{8.7}$	$68.8_{9.5}$	$35.2_{1.3}$	$74.6_{3.8}$	$66.5_{4.5}$	$66.6_{3.3}$	$72.0_{6.0}$	$79.9_{2.0}$	$53.1_{3.3}$
+ LMFT	$39.6_{2.6}$	$52.7_{4.7}$	$52.2_{1.6}$	$66.3_{9.3}$	$66.4_{10.6}$	$36.8_{2.9}$	$75.4_{9.4}$	$58.8_{6.9}$	$51.6_{7.0}$	$75.6_{5.9}$	$80.5_{2.4}$	$52.8_{4.8}$
+ $\text{ITFT}_{\text{MNLI}}$	$79.9_{3.1}$	$62.6_{9.0}$	$64.5_{4.4}$	$80.7_{5.0}$	$72.3_{11.2}$	$36.4_{2.1}$	$75.5_{4.0}$	$77.8_{3.8}$	$73.5_{2.8}$	$82.6_{3.0}$	$72.8_{7.9}$	$69.7_{14.6}$
+ TA	$84.8_{0.7}$	$64.6_{6.3}$	$71.5_{4.0}$	$85.5_{1.4}$	$79.0_{4.5}$	$38.5_{3.0}$	$78.9_{2.4}$	$81.2_{3.9}$	$77.5_{1.4}$	$88.6_{1.3}$	$78.2_{6.6}$	$77.0_{6.3}$
+ ST	$69.3_{9.2}$	$74.3_{1.2}$	$85.4_{1.7}$	$81.9_{12.2}$	$79.9_{4.8}$	$42.0_{1.5}$	$82.8_{2.3}$	$77.3_{3.1}$	$73.1_{2.3}$	$88.1_{1.3}$	$81.2_{0.5}$	$53.9_{4.3}$
+ $\text{ITFT}_{\text{MNLI}}$ + ST	$85.4_{0.3}$	$74.8_{0.7}$	$86.1_{1.1}$	$89.7_{0.7}$	$86.2_{4.2}$	$42.2_{2.0}$	$84.1_{1.7}$	$84.3_{2.0}$	$78.4_{1.3}$	$89.3_{1.0}$	$81.4_{1.2}$	$72.7_{5.4}$
+ STraTA	$\mathbf{87.3}_{0.3}$	$\mathbf{75.1}_{0.2}$	$\mathbf{86.4}_{0.8}$	$\mathbf{91.7}_{0.7}$	$\mathbf{87.3}_{2.9}$	$\mathbf{43.0}_{2.3}$	$\mathbf{84.5}_{1.6}$	$\mathbf{86.3}_{1.8}$	$\mathbf{79.0}_{1.0}$	$\mathbf{90.0}_{0.6}$	$\mathbf{81.5}_{0.7}$	$\mathbf{77.1}_{5.4}$
						Prompt-based (LM-BFF; Gao et al., 2021) and entailment-based (EFL; Wang et al., 2021) fine-tuning approaches						
$\text{RoBERTa}_{\text{LARGE}}$	$38.4_{1.3}$	$58.8_{9.9}$	$52.7_{1.8}$	$60.5_{3.1}$	–	–	$24.5_{8.4}$	–	–	$61.9_{5.1}$	$76.1_{3.9}$	$55.0_{1.3}$
+ LM-BFF	$52.0_{1.7}$	$\mathbf{68.2}_{1.2}$	$61.8_{3.2}$	$79.9_{6.0}$	–	–	$66.0_{3.2}$	–	–	$88.6_{2.3}$	$\mathbf{78.5}_{2.3}$	$63.3_{2.1}$
+ EFL	$\mathbf{81.0}_{1.1}$	$67.3_{2.6}$	$\mathbf{68.0}_{3.4}$	$\mathbf{90.8}_{1.0}$	–	–	$\mathbf{71.0}_{1.3}$	–	–	$\mathbf{92.3}_{0.4}$	$76.2_{1.3}$	$\mathbf{85.8}_{0.9}$