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

Tu Vu

November, 2021
STraTA: Self-Training with Task Augmentation for Better Few-shot Learning

Tu Vu¹,²
Thang Luong¹
Quoc Le¹
Grady Simon¹
Mohit Iyyer²
Agenda

- Motivation
- STraTA: Self-training with Task Augmentation
- Results and Discussion
- Conclusion
The current dominant learning paradigm

Pre-training  Fine-tuning

target task

labeled data

Credit to Jay Alammar for creating the BERT image
Exploiting task-specific unlabeled data

Pre-training  Task augmentation  Self-training

auxiliary task

unlabeled data  labeled data

synthetic training data  pseudo-labeled data

Credit to Jay Alammar for creating the BERT image
STraTA substantially improves sample efficiency

![Graph](image-url)

- **SST-2**
  - BERT\textsubscript{BASE}
  - BERT\textsubscript{BASE} + STraTA
  - BERT\textsubscript{BASE} w/ 67K examples

- **SciTail**
  - BERT\textsubscript{BASE}
  - BERT\textsubscript{BASE} + STraTA
  - BERT\textsubscript{BASE} w/ 27K examples

Accuracy vs. # labeled examples per class
What is self-training?

Teacher Model

Labeled Data

Inference

Pseudo-labeled Data

Repeat until convergence

Student Model

what pseudo-labeled examples to use?
Self-training on a broad distribution of pseudo-labeled data

Iterative confidence filtering

Our self-training algorithm

Labeling accuracy

# self-training iterations

-dev
-test
-predict
-self_train
Our self-training algorithm

Teacher Model

Labeled Data

Pseudo-labeled Data

Student Model

Inference

Use a broad distribution

Repeat until convergence
Our self-training algorithm (cont.)

what model to use?

what model to use?
Task Augmentation

Pre-trained Language Model

Task-specific Unlabeled Texts

Data Generation Model

Synthetic In-domain Auxiliary-task Data

Auxiliary-task Model
Task augmentation

1. Train an NLI data generator by fine-tuning a pre-trained generative model on the MNLI dataset in a text-to-text format

**entailment:** I have met a woman whom I am attracted to → I am attracted to a woman I met

2. Use the model to simulate a large amount of NLI data using target-task unlabeled text

**contradiction:** his acting was really awful

→ he gave an incredible performance

3. Create synthetic in-domain NLI training examples

**[his acting was really awful, he gave an incredible performance]**

→ contradiction
### unlabeled text from SST-2

**Input:** a cast of a-list brit actors

- **Entailment** → **Output:** the cast includes British actors.
- **Neutral** → **Output:** the cast includes a-list brit actors such as Nicole Kidman and Charlie Chaplin.
- **Contradiction** → **Output:** the cast was all new and unknown.

**Input:** both heartbreaking and heartwarming ... just a simple fable done in an artless sytle, but it ’s tremendously moving

- **Entailment** → **Output:** the story is moving and touching.
- **Neutral** → **Output:** the artless style used to make the fable was a favorite of many.
- **Contradiction** → **Output:** the fable is artful and sophisticated but it is also heartbreaking.

### unlabeled text from SciTail

**Input:** Surface tension is responsible for the fact that small insects can walk on water.

- **Entailment** → **Output:** small insects can walk on water due to surface tension.
- **Neutral** → **Output:** it’s difficult for large insects to walk on water.
- **Contradiction** → **Output:** small insects cannot walk on water because of surface tension.

**Input:** Tamiflu inhibits spread of virus.

- **Entailment** → **Output:** the virus is less spread when tamiflu is used.
- **Neutral** → **Output:** tamiflu is a good way to treat hiv.
- **Contradiction** → **Output:** tamiflu promotes viral spread.
STraTA: Self-training with Task Augmentation

**Task Augmentation**
- Pre-trained Language Model
- Task-specific Unlabeled Texts
- Data Generation Model
- Synthetic In-domain Auxiliary-task Data
- Auxiliary-task Model

**Self-training**
- Teacher Model
- Labeled Data
- Pseudo-labeled Data
- Inference
- Use a broad distribution
- Student Model
- Repeat until convergence
### Experimental setup: datasets

| Task                        | Train | Task type                      | Domain                   |
|-----------------------------|-------|--------------------------------|--------------------------|
| *text classification/regression* |       |                                |                          |
| SNLI (Bowman et al., 2015) | 570K  | NLI                            | misc.                    |
| MNLI (Williams et al., 2018) | 393K  | NLI                            | misc.                    |
| QQP (Iyer et al., 2017)    | 364K  | paraphrase identification      | social QA                |
| QNLI (Wang et al., 2019b)  | 105K  | QA-NLI                         | Wikipedia                 |
| SST-2 (Socher et al., 2013) | 67K   | sentiment analysis             | movie reviews            |
| SciTail (Khot et al., 2018) | 27K   | NLI                            | science QA               |
| SST-5 (Socher et al., 2013) | 8.5K  | sentiment analysis             | movie reviews            |
| STS-B (Cer et al., 2017)   | 7K    | semantic similarity            | misc.                    |
| SICK-E (Marelli et al., 2014) | 4.5K  | NLI                            | misc.                    |
| SICK-R (Marelli et al., 2014) | 4.5K  | semantic similarity            | misc.                    |
| CR (Hu and Liu, 2004)      | 4K    | sentiment analysis             | product reviews          |
| MRPC (Dolan and Brockett, 2005) | 3.7K  | paraphrase identification      | news                     |
| RTE (Dagan et al., 2005, et seq.) | 2.5K  | NLI                            | news, Wikipedia          |

Datasets used in our experiments and their characteristics, sorted by training dataset size.
Experimental setup: baselines

**LMFT & ITFT**
- **LMFT**: target-task language model fine-tuning ([Howard and Ruder, 2018](#)); ([Gururangan et al., 2020](#))
- **ITFT**: intermediate-task fine-tuning with MNLI ([Phang et al., 2019](#))

**Prompt/entailment-based fine-tuning**
- **LM-BFF**: prompt-based fine-tuning ([Gao et al., 2021](#))
- **EFL**: entailment-based fine-tuning ([Wang et al., 2021](#))

**Du et al. (2021)**
- **SentAugST**: Retrieval-based augmentation (SentAug) + self-training (ST)
Main results

STraTA significantly improves results across 12 NLP benchmark datasets (numbers in the subscript indicate the standard deviation across 10 random seeds).

| Model            | SNLI | QQP | QNLI | SST-2 | SciTail | SST-5 | STS-B |
|------------------|------|-----|------|-------|---------|-------|-------|
| **FULL (1024 total training examples)** |      |     |      |       |         |       |       |
| BERT\_LARGE      | 91.1 | 88.4| 91.9 | 92.4  | 95.3    | 53.7,9| 89.6,2|
| + LMFT           | 91.0 | 88.1| 90.4 | 93.5  | 95.3    | 54.0,4| 89.5,2|
| + ITFT\_MNLI     | 91.1 | 88.2| 91.6 | 93.5  | 96.5    | 54.0,8| 90.3,3|
| + TA             | **91.9** | **88.5** | **92.5** | **94.7** | **96.9** | **55.7,8** | **90.9,2** |

| **LIMITED (1024 total training examples)** |      |     |      |       |         |       |       |
| BERT\_LARGE      | 77.4,6| 74.1,0| 81.7,9| 89.8,6| 90.9,7 | 49.1,3| 88.2,4|
| + LMFT           | 75.8,5| 71.6,0| 80.5,2| 88.9,8| 87.7,3 | 49.2,1| 88.4,0|
| + ITFT\_MNLI     | 85.2,4| 74.0,0| 83.5,5| 90.0,8| 92.1,1 | 49.4,2| 87.8,8|
| + TA             | **87.3,3** | **75.7,5** | **85.0,5** | **91.7,0** | **92.3,1** | **51.4,0** | **89.0,6** |

| **FEW-SHOT (8 training examples per class)** |      |     |      |       |         |       |       |
| BERT\_LARGE      | 43.1,4| 58.5,4| 64.4,1| 66.1,8| 68.8,5 | 35.2,3| 74.6,38|
| + LMFT           | 39.6,6| 52.7,4| 52.2,16| 66.3,9,3| 66.4,10,6| 36.8,2,9| 75.4,9,4|
| + ITFT\_MNLI     | 79.9,3,1| 62.6,9,0| 64.5,4| 80.7,5,0| 72.3,12,2| 36.4,2,1| 75.5,4,0|
| + 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|
| + 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|
| + STraTA         | **87.3,0,3** | **75.1,0,2** | **86.4,0,8** | **91.7,0,7** | **87.3,2,9** | **43.0,2,3** | **84.5,1,6** |

*Prompt-based (LM-BFF; Gao et al., 2021) and entailment-based (EFL; Wang et al., 2021) methods*
Main results (cont.)

| Model            | SST-2 | SST-5 | CR   |
|------------------|-------|-------|------|
| **Ours (8 examples per class)** |       |       |      |
| BERT<sub>BASE</sub> | 69.8<sub>6.5</sub> | 32.8<sub>2.0</sub> | 73.1<sub>0.5</sub> |
| + TA             | 85.5<sub>0.6</sub> | 41.0<sub>0.8</sub> | 88.7<sub>0.2</sub> |
| + ST             | 74.9<sub>9.0</sub> | 38.3<sub>0.8</sub> | 85.6<sub>1.8</sub> |
| + STraTA         | 90.8<sub>0.6</sub> | 43.1<sub>1.1</sub> | 91.4<sub>0.2</sub> |
| BERT<sub>LARGE</sub> | 75.6<sub>3.3</sub> | 36.6<sub>0.4</sub> | 79.3<sub>0.7</sub> |
| + TA             | 87.3<sub>0.3</sub> | 41.7<sub>1.1</sub> | 90.0<sub>0.4</sub> |
| + ST             | 90.6<sub>0.3</sub> | 43.8<sub>0.4</sub> | 89.0<sub>1.1</sub> |
| + STraTA         | 92.4<sub>0.1</sub> | 45.5<sub>0.7</sub> | 90.6<sub>0.0</sub> |
| **Du et al. (2021) (20 examples per class)** |       |       |      |
| RoBERTa<sub>LARGE</sub> | 83.6<sub>2.7</sub> | 42.3<sub>1.6</sub> | 88.9<sub>1.7</sub> |
| + SentAugST      | 86.7<sub>2.3</sub> | 44.4<sub>1.0</sub> | 89.7<sub>2.0</sub> |

Compared to Du et al. (2021), our approach leads to better downstream performance, despite using a weaker base model (BERT vs. RoBERTa) and with less labeled examples.
STraTA improves a randomly-initialized base model while being competitive on SciTail. Additionally, vanilla BERT typically add a small set of unlabeled examples et al. algorithms (pseudo-labeled data: Self-training on a broad distribution of and SciTail, respectively. BERT the vanilla BERT as BERT model (RAND applied to a randomly initialized Transformer base model to exhibit improvements: when approach does not require a powerful pre-trained model: STraTA improves a randomly-initialized base dataset of 27K labeled examples at performance of standard fine-tuning with the whole dataset of 67K labeled examples. On the harder task of SciTail, STraTA already nearly saturated its performance, achieving results competitive with standard fine-tuning even when starting with a randomly-initialized model, but pre-training helps considerably.

| Model          | SST-2 | SciTail |
|----------------|-------|---------|
| RAND_{BASE}    | 50.0_{1.6} | 50.7_{2.4} |
| + STraTA       | 78.6_{0.9}  | 64.4_{3.1}  |
| BERT_{BASE}    | 59.1_{8.4}  | 67.1_{6.6}  |
| + STraTA       | 90.1_{0.8}  | 86.3_{3.5}  |
| BERT_{LARGE}   | 66.1_{8.7}  | 68.8_{9.5}  |
| + STraTA       | 91.7_{0.7}  | 87.3_{2.9}  |

Our approach yields improvements even when starting with a randomly-initialized model, but pre-training helps considerably.
Does self-training work with out-of-domain/distribution unlabeled data?

| Model            | SciTail | CR  | MRPC | RTE  |
|------------------|---------|-----|------|------|
| \( \text{BERT}_{\text{BASE}} \) | 67.1\,_{6.6} | 65.2\,_{8.2} | 72.4\,_{10.2} | 51.4\,_{2.5} |
| \( \text{BERT}_{\text{BASE}} + \text{TA} \) | 78.5\,_{3.2} | 86.5\,_{2.2} | 74.5\,_{6.5} | 67.6\,_{7.1} |
| \( + \text{ST}_{\text{IN}} \) | 86.3\,_{3.5} | 90.5\,_{0.8} | 81.0\,_{0.8} | 70.6\,_{2.4} |
| \( + \text{ST}_{\text{OUT}} \) | 81.4\,_{3.7} | 88.3\,_{1.9} | 80.3\,_{1.9} | 71.2\,_{3.2} |
| \( + \text{ST}_{\text{IN} + \text{OUT}} \) | 82.6\,_{2.6} | 88.3\,_{1.5} | 80.2\,_{1.1} | 69.9\,_{4.0} |

Self-training with out-of-domain unlabeled examples also results in improvements, but using in-domain data works significantly better.
Towards realistic evaluation in few-shot learning

| Model               | SST-2       | SciTail     |
|---------------------|-------------|-------------|
| BERT<sub>BASE</sub> | 58.8<sub>8.4</sub> (↓ 0.3) | 61.5<sub>5.4</sub> (↓ 5.6) |
| + LMFT              | 64.0<sub>8.1</sub> (↓ 0.9) | 59.3<sub>5.6</sub> (↓ 4.7) |
| + ITFT<sub>MNLI</sub> | 76.5<sub>7.2</sub> (↓ 0.3) | 76.2<sub>5.4</sub> (↑ 0.4) |
| + TA                | 79.8<sub>6.3</sub> (↓ 0.5) | 77.8<sub>3.3</sub> (↓ 0.7) |
| + STraTA            | 86.6<sub>2.6</sub> (↓ 3.5) | 80.6<sub>3.0</sub> (↓ 5.7) |

In a realistic evaluation without a development set, our STraTA approach still leads to significant improvements on top of BERT<sub>BASE</sub>. In parentheses, we show the absolute increase (↑) or decrease (↓) in performance compared to the same method used with a development set.
Conclusion

STraTA

✧ two *complementary* and *independently effective* methods to leverage task-specific unlabeled data for improved downstream performance

• *task augmentation*: synthesizes a large amount of in-domain data for auxiliary-task fine-tuning from target-task unlabeled texts

• *self-training*: trains on a broad distribution of pseudo-labeled data

✧ substantially improves sample efficiency across 12 NLP benchmark datasets
Thank you!

Code will be available at
https://github.com/google-research/google-research/tree/master/STraTA