Few-Shot Learning in NLP

A Survey

Minh Duc Bui, mbui@mail.uni-mannheim.de

June 11, 2021

Outline

Few-Shot Learning and Why its hard

What is Few-Shot Learning?

Supporting Set	
(A) capital_of	(1) *London* is the capital of *the U.K*. (2) *Washington* is the capital of *the U.S.A*.
(B) member_of	(1) *Newton* served as the president of *the Royal Society*. (2) *Leibniz* was a member of *the Prussian Academy of Sciences*.

Test Instance	
(A) or (B)	*Euler* was elected a foreign member of *the Royal Swedish Academy of Sciences*.

Figure 1: 2-Shot Relation Classification.

What is Few-Shot Learning?

Few-Shot Learning approaches **use prior knowledge** to generalize to new tasks containing only a **few samples** with supervised information.

Task T	**Experience E**		**Performance P**
	Supervised Information	**Prior Knowledge**	
Relation Classification	Few examples of Relations	Pre-learned language semantics	Acc.

Transfer Learning: Sample Efficiency

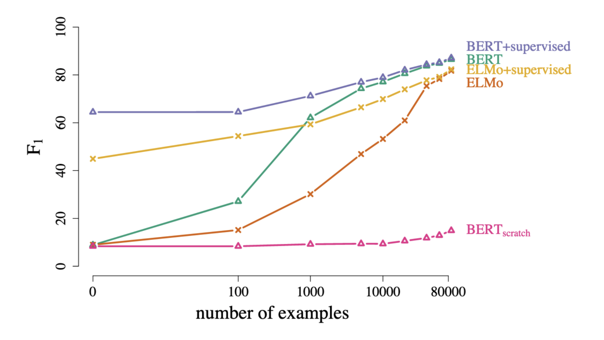


Figure 2: F1 scores on SQuAD as a function of the number of training examples (log scale). BERT+supervised denote BERT that is pretrained on other datasets and tasks.[1]

[1] *Yogatama, D. et al. Learning and Evaluating General Linguistic Intelligence. 2019.*

Approaches

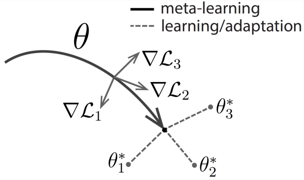
Optimization-Based Approaches

Goal:

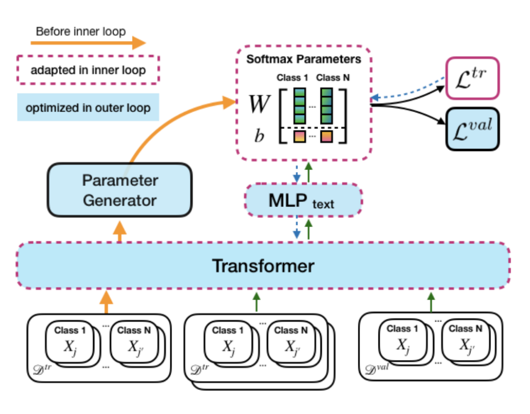
- **Learn a good set of parameter initialization** by using many tasks and treating **each task as a training example**

Training:

- Fine-tuning the model on a training set $\mathcal{D}_i^{tr}$ of a selected training task, which **only consists of K examples**
- Use the task loss $\mathcal{L}_i$ on $\mathcal{D}_i^{test}$ to update our original not fine-tuned model parameters by computing the gradient



Optimization-Based: LEOPARD[1]



Trained on 7 tasks and evaluated on 17 tasks in few-shot scenario

[1] *Bansal, T., Jha, R., and McCallum, A. "Learning to Few-Shot Learn Across Diverse Natural Language Classification Tasks". 2019.*

Optimization-Based: Discussion

- LEOPARD is first meta-learning approach that could generalize to test tasks, significantly different than training tasks (NLP)

	N	k	$\text{BERT}_{\text{base}}$	$\text{MT-BERT}_{\text{softmax}}$	MT-BERT	Proto-BERT	LEOPARD
Overall Average		4	38.13	40.13	40.10	36.29	**45.99**
		8	36.99	45.89	44.25	39.15	**50.86**
		16	48.55	49.93	49.07	39.85	**55.50**

Figure 3: Few-shot generalization performance across tasks not seen during training.

Meta-Learning: Discussion

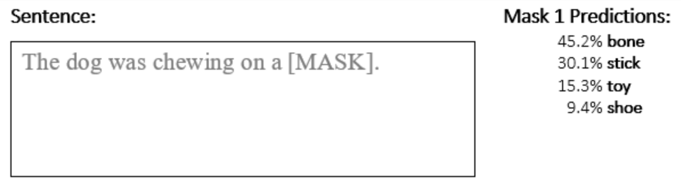
Supporting Set	
(A) capital_of	
(B) member_of	

Test Instance	
(A) or (B)	*Euler* was elected a foreign member of *the Royal Swedish Academy of Sciences*.

Figure 4: 2-Shot Relation Classification. Can you do zero-shot learning?

How to use pre-trained language models?

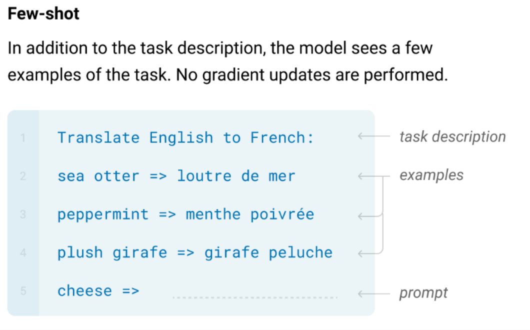
- Transformers are simply pre-trained on a language modeling objective in a semi-supervised fashion



- **Reformulate Tasks as language modeling problems!**

Reformulate Tasks: Few-Shot with GTP-3[1]

- Model is given a task description and K examples of context and completion, which they call model *priming*
- To make predictions, one final context is given, but the model has to fill in the completion



[1] *Brown, T. B. et al. Language Models are Few-Shot Learners. 2020.*

Reformulate Tasks: GTP-3 Result

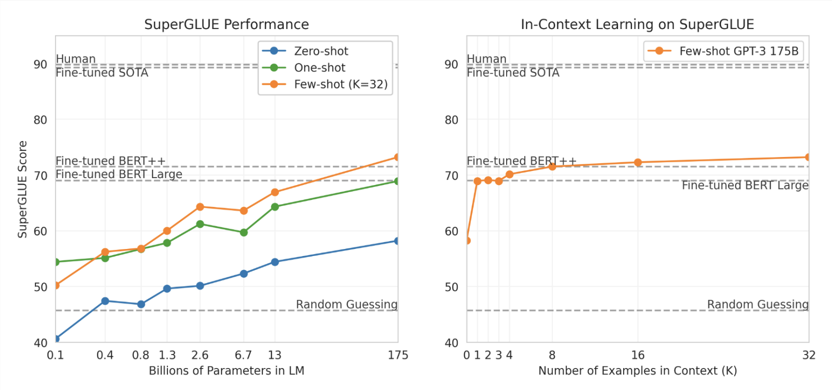
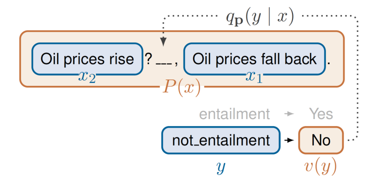


Figure 5: Performance on SuperGLUE increases with model size and number of examples in context.[1]

[1] *Brown, T. B. et al. Language Models are Few-Shot Learners. 2020.*

Reformulate Tasks: Other Approaches

- **iPET**[1]
  - Reformulates tasks as cloze questions and uses ALBERT with regular gradient-based finetuning
  - Uses knowledge distillation, which in turn need unlabeled data



	Model	Params (M)	BoolQ Acc.	CB Acc. / F1	COPA Acc.	RTE Acc.	WiC Acc.	WSC Acc.	MultiRC EM / F1a	ReCoRD Acc. / F1	Avg –
test	GPT-3	175,000	76.4	75.6 / 52.0	**92.0**	69.0	49.4	80.1	30.5 / 75.4	**90.2 / 91.1**	71.8
test	PET	223	79.1	87.2 / 60.2	90.8	67.2	**50.7**	**88.4**	**36.4 / 76.6**	85.4 / 85.9	74.0
test	iPET	223	**81.2**	**88.8 / 79.9**	90.8	**70.8**	49.3	**88.4**	31.7 / 74.1	85.4 / 85.9	**75.4**
test	SotA	11,000	*91.2*	*93.9 / 96.8*	*94.8*	*92.5*	*76.9*	*93.8*	*88.1 / 63.3*	*94.1 / 93.4*	*89.3*

Figure 6: Results on SuperGLUE for GPT-3 primed with 32 randomly selected examples and for iPET after training on 32 random examples.

[1] *Schick, T. and Schütze, H. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. 2021.*

Few Shot Learning in Cross Lingual Setting

PRE-TRAINING ON SIMILAR TASKS

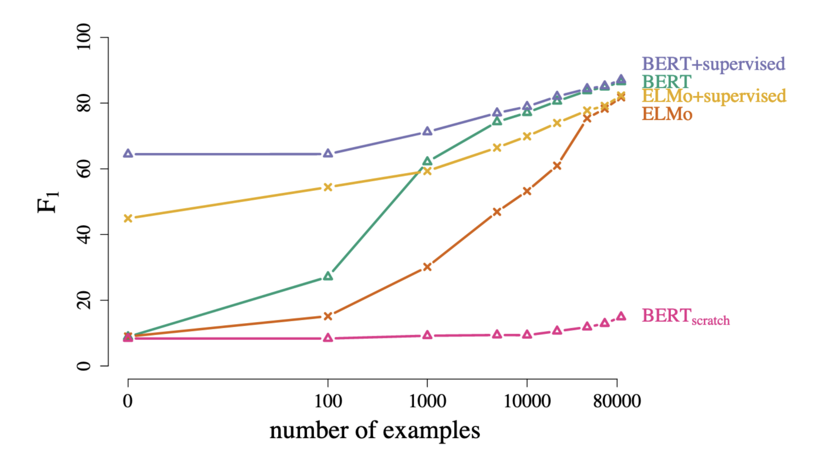


Figure 7: F1 scores on SQuAD as a function of the number of training

Cross-Lingual Setting

- Transfer the knowledge about the same task **from a high resource language to low resource language**

- Process:
  1. Couple a multilingual Transformer with task-specific classifier
  2. **Fine-tune model using task-specific supervised training data** from one **high resource language** (*source-adaption*)
     - If stop here: Zero-Shot Transfer
  3. **Continue fine-tuning** on K task-specific examples in the (low resource) target language (*target-adaption*).

Cross-Lingual Results: Zero Shot

Fine-tuning \ Eval	EN	DE	ES	IT
EN	**96.82**	89.40	85.91	91.60
DE	83.99	**93.99**	86.32	88.39
ES	81.64	88.87	**96.71**	93.71
IT	86.79	87.82	91.28	**98.11**

Figure 8: Zero Shot POS accuracy.[1]

Task	Model	EN	ZH Δ	TR Δ	RU Δ	AR Δ	HI Δ	EU Δ	FI Δ	HE Δ	IT Δ	JA Δ	KO Δ	SV Δ
DEP	B	92.3	-40.9	-41.2	-23.5	-47.9	-49.6	-42.0	-26.7	-29.7	-10.6	**-55.4**	-53.4	-12.5
POS	B	95.5	-33.6	-26.6	-9.5	-32.8	-33.9	-28.3	-14.6	-21.4	-6.0	**-47.3**	-37.3	-6.2
NER	B	92.3	-31.5	-6.5	-9.2	-29.2	-12.8	-8.5	-0.9	-9.2	-0.8	**-51.1**	-12.9	-1.9

Figure 9: Zero-shot cross-lingual transfer performance with mBERT (B).[2]

[1] *Pires, T., Schlinger, E., and Garrette, D. How multilingual is Multilingual BERT?. 2019.*

[2] *Lauscher, A. et al. From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers. 2020.*

Cross-Lingual Results: Few Shot

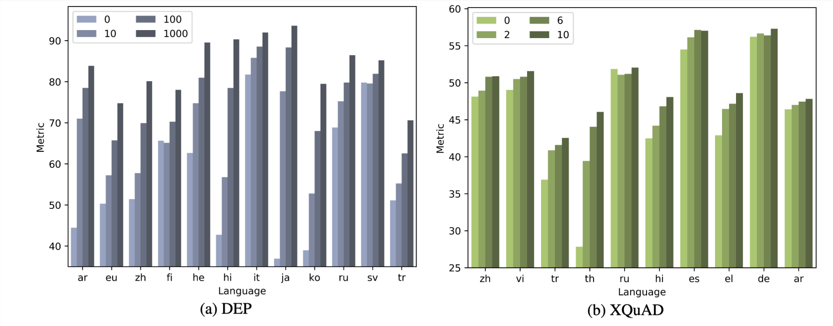


Figure 10: Results of the few-shot experiments with varying numbers of target-language examples k.[1]

[1] *Lauscher, A. et al. From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers. 2020.*

Summary & Discussion

Big Transformers, High Variance

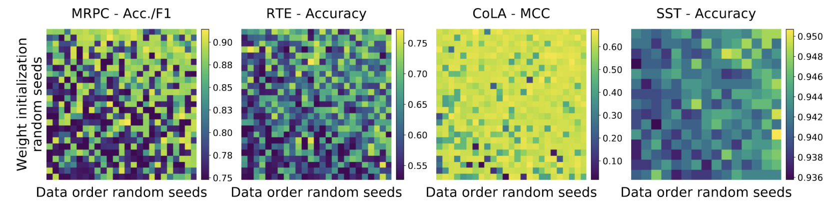


Figure 11: Visualization of validation performance, where each colored cell represents the performance of a training run[1].

[1] *Dodge, J. et al. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. 2020.*

Summary & Discussion

- Discussed Optimization-based Meta-Learning
  - + Efficient use of few examples through optimal parameter initialization
  - - Creating training tasks that enable finding a good initialization set to solve the target task is difficult.
  - - Overfitting on task distribution

- Discussed reformulating to Language Modelling problems
  - + Achieve impressive results with small number of examples
  - - Favor tasks, that can naturally be reformulated as ”fill-in-the-blank”
  - - Finding the right prompt is an art, needs a big enough validation set.[1]
  - - Restricted input size

[1] *Gao, T., Fisch, A., and Chen, D. Making Pre-trained Language Models Better Few-shot Learners. 2020.*

Summary & Discussion

- Discussed Few-Shot Cross-Lingual Transfer
  + Finetuning on few examples, can significantly improve performances on distant languages - exactly where zero-shot CLT fail.
  - Inefficient way of finetuning.
  - Need labeled data in the low resource target language, which is typically hard to acquire.

- Few-Shot in General
  + Enables to train without needing many training examples
  + Advances to ultimate goal of NLP: General-purpose language understanding
  - Since most method rely on Transformers, they suffer high variance.
  - Essential to use the same set or average between multiple equal sets when comparing few-shot approaches, which is still lacking

Appendix

Pre-training on Language Modelling objective

Figure 12: Pretraining on a language modelling objective.

Prior Knowledge: General Purpose Language Understanding

- Solving NLP tasks requires the model to learn about syntax, semantics, as well as certain facts about the world

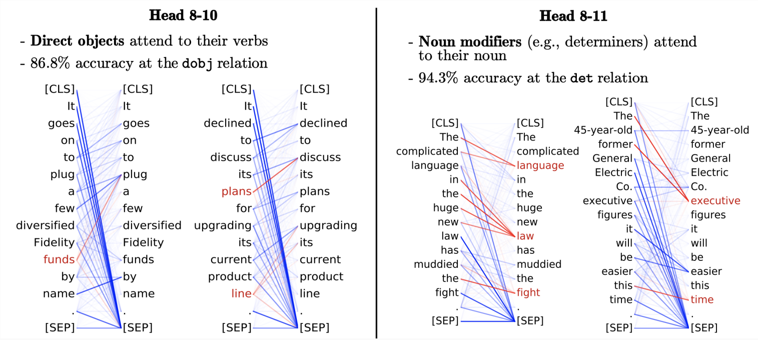


Figure 13: What Does BERT Look At? An Analysis of BERT's Attention [Clark et al. 2019]

Meta-Learning: Discussion

- Challenge: Create training tasks enabling meta-learning algorithms to find a good initialization
  - Requires labeled data from many different tasks and additionally

- Major obstacle for meta-learning approaches is to solve diverse NLP few-shot tasks with one model
  - Suffers from overfitting to the training task-distribution (meta-overfitting)[1]
  - Does not use any information of the underlying task

[1] *Bansal, T., Jha, R., Munkhdalai, T., et al. "Self-Supervised Meta-Learning for Few-Shot Natural Language Classification Tasks". 2020.*

Reformulate Tasks: Discussion

- Achieve **impressive results** with only a small amount of examples
- Approaches **favor tasks, that can naturally be reformulated as ”fill-in-the-blank” problems**, leaving room for future work
- LMs have **restricted input size**
  - Tasks that have long input sequences can not be properly solved
  - Future work: Use LM that allow long input sequences[1]

[1] *Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The Long-Document Transformer. 2020.*

Big Transformers, High Variance

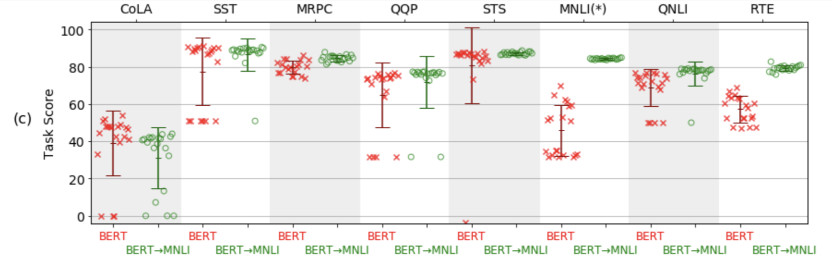


Figure 14: Distribution of task scores across 20 random restarts for BERT, and BERT with intermediary fine-tuningon MNLI. Fine-tuned on nomore than 1k examples for each task.[1]

[1] *Phang, J., Févry, T., and Bowman, S. R. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. 2019.*