It's Not Just Size That Matters: SLMs Are Also Few-Shot Learners

Project presentation

Lucas Fourest, Adéchola Kouande, Marius Roger

Article analysis

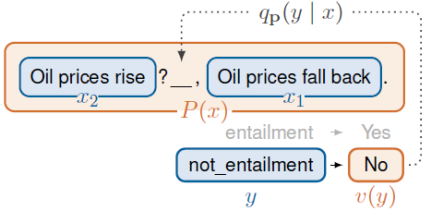
Context and goal of the article

Context :

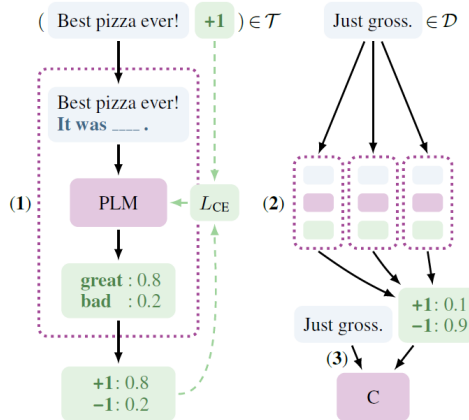
- Few-shot learning
- T. Schick and H. Schütze (we don't need 175 billion parameters model)
- Drawbacks :
  - Expressivity (model size) needed
  - Context length (GPT-3 : 2048, usual LMs : 128-512)
    at $\approx$ 20 to 50 tokens per few-shot sample, not scalable for "smaller" models

Goal : Attain the performance of GPT-3 with a much smaller model

Pattern-Exploit Training



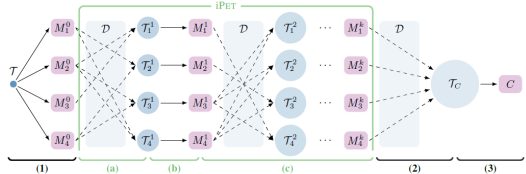
Tasks can be modeled as Cloze questions



PET for sentiment classification

Alternative techniques

Iterative PET : train k successive groups of PET models on increasingly big datasets through soft labelling



PET $(1-3)$ and iPET $(a-c)$

PET with multiple masks : useful to verbalize into more complex patterns
"x. It was _." $\longrightarrow$ ("good"/"bad") can become ("de·lic·ious"/"terri·ble")

- Compute $l(x) = max_{y \in Y_x} |v(y)|$, the maximum number of `MASK` tokens inserted
- Inference : $\forall y \in Y_x$, mask $|v(y)|$ tokens $\rightarrow$ AR predict $v(y) \rightarrow$ get score
- Training : mask $l(x)$ tokens, $\forall y \in Y_x \rightarrow$ predict $v(y)$ at once $\rightarrow$ get score

Their experiments

- Tasks from SuperGLUE :
  BoolQ (QA), CB / RTE (entailment), COPA (given premise p, which of $\{c_1, c_2\}$ is cause/effect), WiC (does w mean the same in s_1 and s_2), WSC (which noun n does pronoun p reference), MultiRC (QA with candidate answer), ReCoRD (Cloze)

- Model and training:
  `ALBERT-xxlarge-v2` + SeqClass head, trained on `FewGLUE`
  PET with MM is used for COPA, WSC, and ReCoRD, standard PET for others

- Protocol and metrics:
  Compare with different sizes of GPT-3 (and SotA for reference)
  Per-task metrics, mostly accuracy: average accuracy across tasks in the last column

Their results

	Model	Params (M)	BoolQ Acc.	CB Acc. / F1	COPA Acc.	RTE Acc.	WiC Acc.	WSC Acc.	MultiRC EM / F1a	ReCoRD Acc. / F1	Avg –
dev	GPT-3 Small	125	43.1	42.9 / 26.1	67.0	52.3	49.8	58.7	6.1 / 45.0	69.8 / 70.7	50.1
	GPT-3 Med	350	60.6	58.9 / 40.4	64.0	48.4	55.0	60.6	11.8 / 55.9	77.2 / 77.9	56.2
	GPT-3 Large	760	62.0	53.6 / 32.6	72.0	46.9	53.0	54.8	16.8 / 64.2	81.3 / 82.1	56.8
	GPT-3 XL	1,300	64.1	69.6 / 48.3	77.0	50.9	53.0	49.0	20.8 / 65.4	83.1 / 84.0	60.0
	GPT-3 2.7B	2,700	70.3	67.9 / 45.7	83.0	56.3	51.6	62.5	24.7 / 69.5	86.6 / 87.5	64.3
	GPT-3 6.7B	6,700	70.0	60.7 / 44.6	83.0	49.5	53.1	67.3	23.8 / 66.4	87.9 / 88.8	63.6
	GPT-3 13B	13,000	70.2	66.1 / 46.0	86.0	60.6	51.1	75.0	25.0 / 69.3	88.9 / 89.8	66.9
	GPT-3	175,000	77.5	82.1 / 57.2	92.0	72.9	**55.3**	75.0	32.5 / 74.8	**89.0 / 90.1**	73.2
	PET	223	79.4	85.1 / 59.4	**95.0**	69.8	52.4	**80.1**	**37.9 / 77.3**	86.0 / 86.5	74.1
	iPET	223	**80.6**	**92.9 / 92.4**	**95.0**	**74.0**	52.2	**80.1**	33.0 / 74.0	86.0 / 86.5	**76.8**
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
	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
	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**
	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*

PET or iPET are almost always better than GPT-3

No gradient updates on GPT-3, and some kind of knowledge distillation in PET

Still, using PET methods appears to be the more reasonable option in practical use

Our work

Experimental phase

Goal : Understand efficiency/usefulness of Cloze reformulation for few-shot learning

Dataset : Internet Movie Database (binary sentiment classification, review-sentiment)
25k training pairs, 25k test pairs, 50k unlabeled review texts (iPET ?)

Models : DistillBERT, BERT

Preliminary experiment : We compare the impact of multiple PVP (each denoted by a 2-digit *pv* code) against a vanilla fine-tuning strategy on identical sets of 32 pairs

Metric : Accuracy

Our patterns and verbalizer

$P_1 : r \rightarrow$ The review is: r Is it a positive review?
$P_2 : r \rightarrow$ r Did this user like the movie ?
$P_3 : r \rightarrow$ Read the following review: r Did this user enjoy its experience?
$P_4 : r \rightarrow$ The review is: " r ". Is it a positive review?
$P_5 : r \rightarrow$ " r ". Did this user like the movie ?
$P_6 : r \rightarrow$ Read the following review: " r ". Did this user enjoy its experience?

$$v : \begin{cases} 1 \rightarrow \textbf{"yes"} \\ 0 \rightarrow \textbf{"no"} \end{cases}$$

Applying a PVP :

(*What a great movie*, 1) $\rightarrow$ (The review is: What a great movie Is it a positive review?, **yes**)

Results

Preliminary results

Model	*classic*	*pvp=11*	*pvp=21*	*pvp=31*	*pvp=41*	*pvp=51*	*pvp=61*
DistilBERT	51.55	56.61	**56.82**	56.00	55.37	53.62	55.06
BERT	51.62	52.69	53.72	**54.65**	53.00	54.13	53.62

Comparisons of classical training and PET for few-shot learning with gradient updates
Bold is the best fine-tuning for a model (line)

What we noticed

- Accuracy still in 50-60% range : PET only slightly (but consistently) better
  Sentiment analysis task on "raw review" may be too complex for only 32 examples
- Patterns 1, 2, 3 perform better than "punctuated" counterparts
  Maybe punctuation adds unnecessary complexity
- DistilBERT better than BERT on all but one PVPs, while being similar in classic training (despite BERT being larger)
  Either BERT overfits (only 32 pairs) or DistilBERT benefits more from Cloze
- Bigger models (ALBERT xxlarge from the paper) couldn't fit in local GPU memory
  The problem set in the article is even more prominent on individual user hardware

How we intend to continue

- Experiment with bigger models (if we can manage it)
- Create new patterns and verbalizers and evaluate their performance
- Implement Iterative PET and their "knowledge distillation" technique
- Add possibility of using other datasets: SuperGLUE as in the article