FiD-ICL: A Fusion-in-Decoder Approach for Efficient In-Context Learning

Qinyuan Ye, Iz Beltagy, Matthew E. Peters, Xiang Ren, Hannaneh Hajishirzi qinyuany@usc.edu

USC University of Southern California · AI2

TL;DR

We adapt fusion-in-decoder models (originally designed for open-domain QA) to perform in-context learning.

📈 Performance

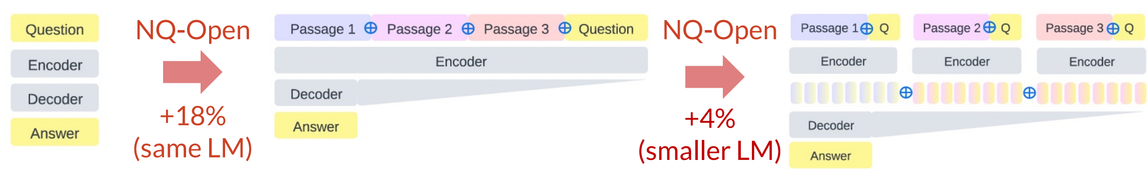
FiD-ICL ***outperforms*** Concat-ICL and Ensemble-ICL.
The ***gap*** between FiD-ICL and fine-tuning is ***<3%*** on P3 meta-test tasks.

✨ Efficiency

FiD-ICL is ***faster*** than Concat-ICL and Ensemble-ICL;
More efficient than fine-tuning when considering optimization costs.

Motivation: QA vs. ICL

Closed-book QA (Roberts et al., 2020) · **Retrieval-Augmented Generation** (Lewis et al., 2020) · **Fusion-in-Decoder** (Izacard et al., 2020)



Zero-shot Learning · **Few-shot In-Context Learning** · **This work: FiD-ICL**

Experiment Setting

Data

Public Pool of Prompts (P3)
(Sanh et al., 2022; Bach et al., 2022)

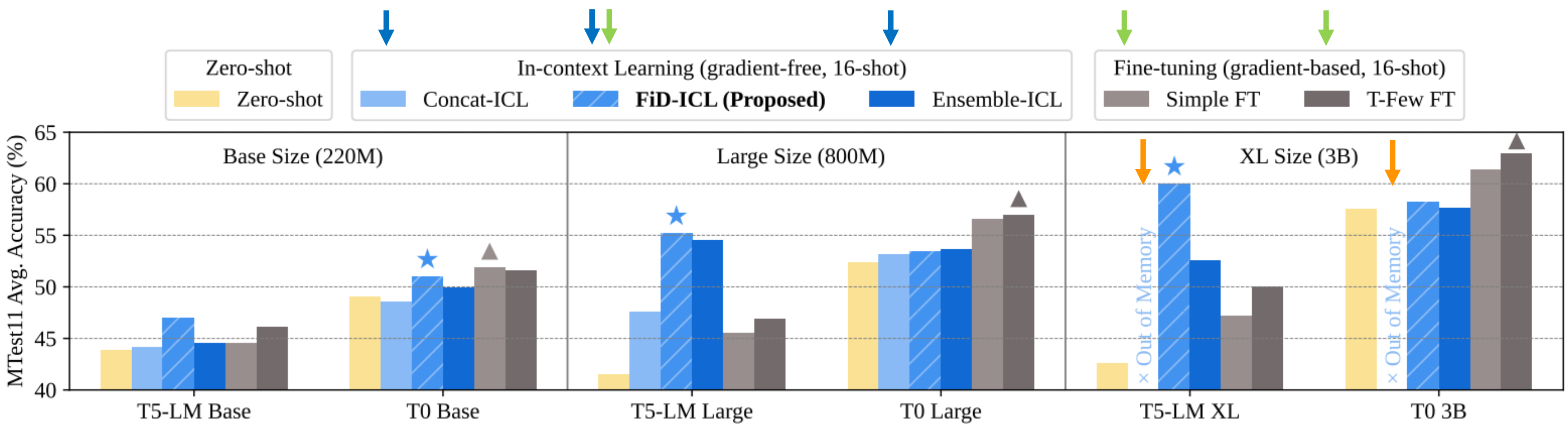
Training Procedure

Meta-train on *seen* tasks;
Meta-test on *unseen* tasks

Compared Methods

Method	Meta-Train T0	Meta-Train ICL	Meta-Test Fine-tune	Meta-Test # shots
Initialize from T5-LM				
Zero-shot	✗	✗	✗	0
Concat/FiD/Ensemble-ICL	✗	✓	✗	k
Simple/TFew Fine-tune	✗	✗	✓	k
Initialize from T0				
Zero-shot	✓	✗	✗	0
Concat/FiD/Ensemble-ICL	✓	✓	✗	k
Simple/TFew Fine-tune	✓	✗	✓	k

Performance: P3 Meta-Test (11 Held-out Tasks)



Key Observations

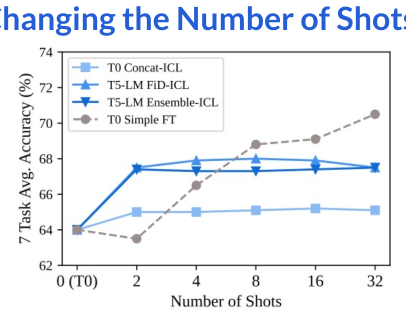
① FiD-ICL enables efficient meta-training; Concat-ICL would fail at the scale of 3B.
② FiD-ICL is comparable or outperforms the Concat-ICL and Ensemble-ICL.
③ The gap between FiD-ICL (★ gradient-free) and fine-tuning (▲ gradient-based) is <3%.

Efficiency: Comparing Runtime (Pre-inference + Inference)

	Zero-shot	Concat-ICL	FiD-ICL (Proposed)	Ensemble-ICL	Simple FT	T-Few FT
	Zero-shot	In-context Learning (gradient-free, 16-shot)			Fine-tuning (gradient-based, 16-shot)	
RTE	1x	17x	2x ✨	26x	26x	>26x
StoryCloze	1x	27x	3x ✨	29x	6x	>6x

Analysis

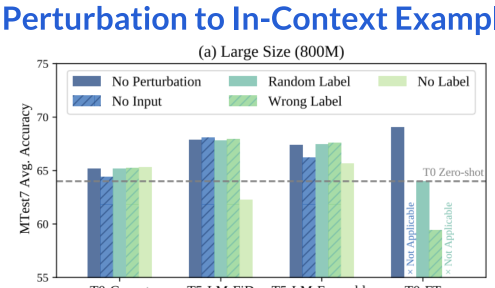
Changing the Number of Shots



Meta-test average performance ***does not*** grow with more shots.

The behavior is ***task-dependent***.

Perturbation to In-Context Examples



Performance is rather ***insensitive*** to perturbations to in-context examples.

FiD-ICL is still ***not*** learning ***effectively***.

Implications

Insights and methodologies from ***open-domain QA*** can be very useful to ICL!

FiD-ICL is also related to …
retrieval augmentation,
sparse attention,
hypernetworks.