GistScore: Learning Better Representations for In-Context Example Selection with Gist Bottlenecks

Shivanshu Gupta shivag5@uci.edu

Clemens Rosenbaum cgbr@cs.umass.edu

Ethan R. Elenberg ethan@permanence.ai

https://arxiv.org/abs/2311.09606

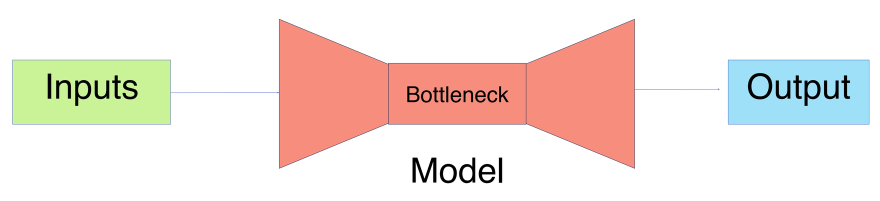
https://github.com/Shivanshu-Gupta/gist-icl

Motivation

- In-Context Learning (ICL) uses LLMs for new tasks by conditioning them on prompts comprising a few task examples. However, ICL performance is critically sensitive to the choice of those examples.
- Prior example selection approaches are either not very effective, require expensive task and/or LLM-specific training[3,4,5], or are too slow[1].
- Thus, the standard approach remains to use general-purpose retrievers like BM25 or Cosine Similarity with SentenceBERT.
- We propose a novel approach for selecting informative ICL examples that is not only superior in performance, but is also fast and can be used without any training!

Intuition

Idea: Train a model to perform a task with a *bottleneck* between the inputs and output

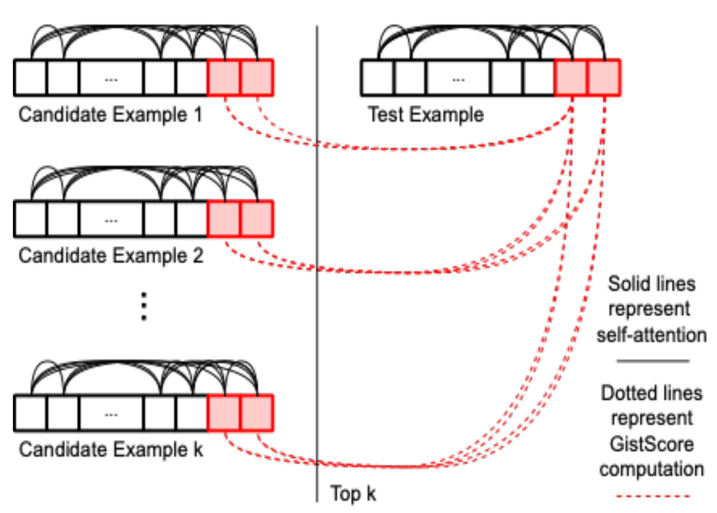


- Encodes the task-specific salient aspects[1] of the inputs.
- Then use encoding to retrieve informative ICL examples with similar salient aspects.

Example Selection

1. Compute the gist embeddings for candidates z and test input x_{test}.
2. Score each candidate using **GistScore** to rank and pick top-k.

$$\mathrm{GS}(\mathbf{x}, \mathbf{z}) = \underbrace{\frac{1}{l}\sum_{i=1}^{l}}_{\text{Sum over test input's gist tokens}} \underbrace{\max_{j=1,\ldots l}}_{\text{Maximum similarity across candidate gist tokens}} \frac{\mathbf{x}_i^T \mathbf{z}_j}{\|\mathbf{x}_i\| \|\mathbf{z}_j\|}$$



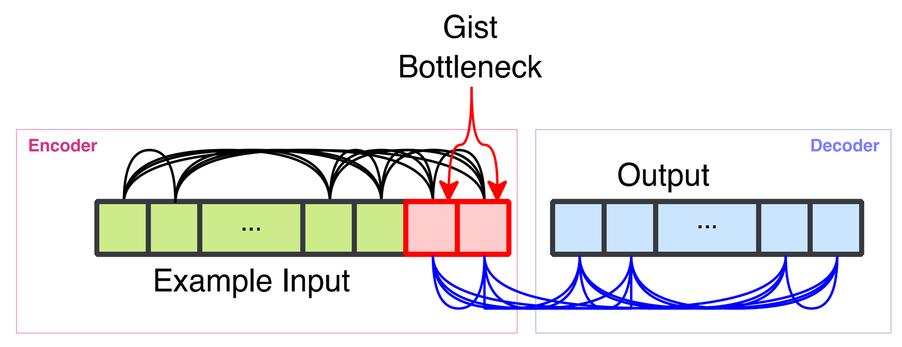
GistScore can also be extended to a set-level metric[1] Set-GS, selecting examples together as a set rather than ranking independently.

References

[1] Gupta, S., Gardner, M., and Singh, S. Coverage-based example selection for in-context learning. EMNLP Findings 2023.
[2] Mu, J., Li, X. L., and Goodman, N. Learning to compress prompts with gist tokens. NeurIPS 2023.
[3] Rubin, O., Herzig, J., and Berant, J. Learning to retrieve prompts for in-context learning. NAACL 2022.
[4] Ye, J., Wu, Z., Feng, J., Yu, T., and Kong, L. Compositional exemplars for in-context learning. ICML 2023.
[5] Wang, L., Yang, N., and Wei, F. Learning to retrieve in-context examples for large language models. EACL 2024.

Example Gisting

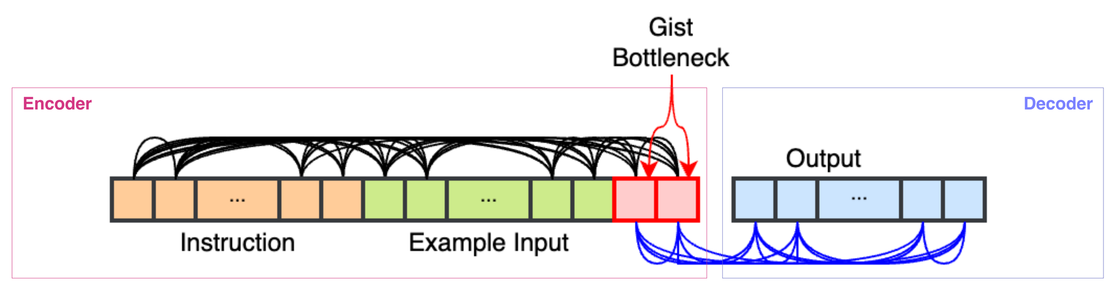
Supervised finetuning of an encoder-decoder model with an attention-masking bottleneck such that the decoder can only attend to the input x through gist tokens $G = g_1 g_2 \ldots g_l$.



Encoder learns to encode salient information of the inputs in activations above G. Use the top-layer embeddings as the encoding for retrieval.

Multi-task Training

Problem: Finetuning improves performance but sacrifices the training-free ICL pipeline. Design an easy-to-use encoder that works with new datasets and tasks out-of-the-box.



Solution: Multi-task training with task instructions where Instruction t and input x are attended to through G. Instructions allow the encoder to infer the task and extract task-specific salient information from the input. Use top-layer embeddings for gist tokens are used during retrieval as before.

Experimental Setup

Selectors (l: number of gist tokens)

- GS[F, l]: `flan-t5-base` finetuned on each dataset.
- GS[M, l]: `flan-t5-large` multi-task trained on a subset (~5M) of zero-shot prompts from FLAN-2022, then applied to downstream datasets.

Evaluation

- 21 datasets spanning 9 diverse task categories + 8 diverse LLMs.
- Compositional generalization and multilingual settings.
- Held-out datasets, domains, and tasks for the multi-task model.

Conclusion

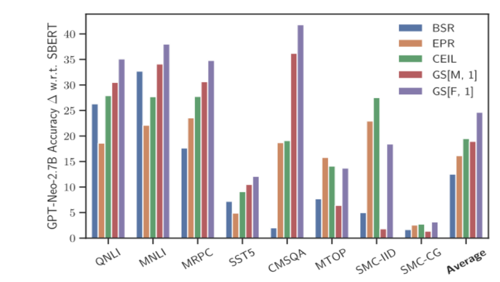
- Example Gisting extracts task-specific salient information useful for selecting informative in-context examples.
- Finetuned GistScore yields state-of-the-art ICL performance.
- Multi-task pretrained GistScore generalizes out-of-the-box to new datasets and presents the best trade-off of performance, ease-of-use, and selection speed.

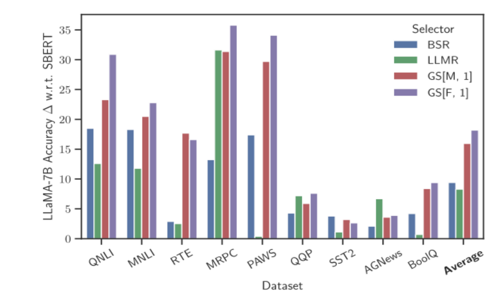
Results

- Finetuned GistScore GS[F, 1] beats all prior approaches, outperforming SBERT by over 21 points!

Selector	Neo	L7B	L13B	Mis.	Zeph.	Bab.	Dav.
RAND	38.0	46.3	48.9	56.4	58.8	39.9	52.4
BM25	46.2	53.6	57.3	64.0	65.1	45.4	57.4
SBERT	46.5	53.7	57.7	64.6	65.5	47.3	58.1
BSR	57.1	60.8	64.6	70.9	70.1	57.3	65.4
GS[M, 1]	63.5	65.8	68.1	73.6	71.7	63.1	68.4
GS[F, 1]	**68.1**	**70.1**	**71.8**	**76.5**	**74.9**	**67.3**	**71.0**

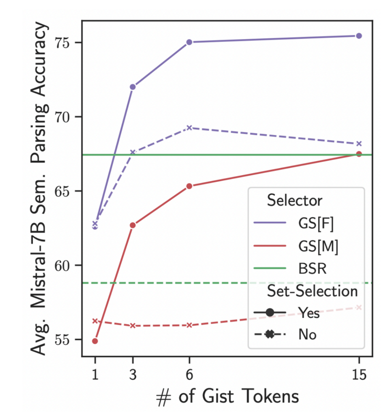
- Multi-task trained GistScore GS[M, 1] is best among training-free methods
- Competitive with trained baselines (EPR[3], CEIL[4], LLM-R[5]) without additional finetuning.



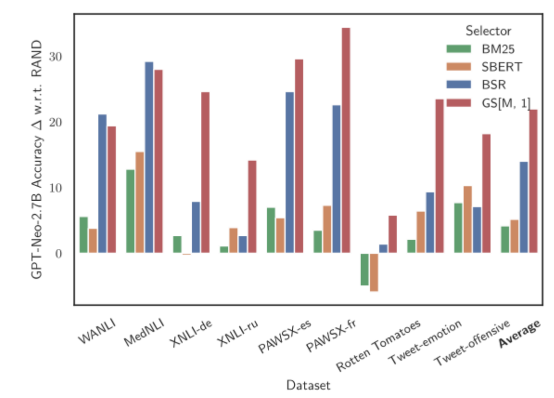


Semantic Parsing (Held-out)

- Unlike other tasks, significantly benefits from using >1 gist token and set-selection.
- Set-GS[M, 15] matches Set-BSR[1] despite never training on this task.
- Set-GS[F, 15] performs best.

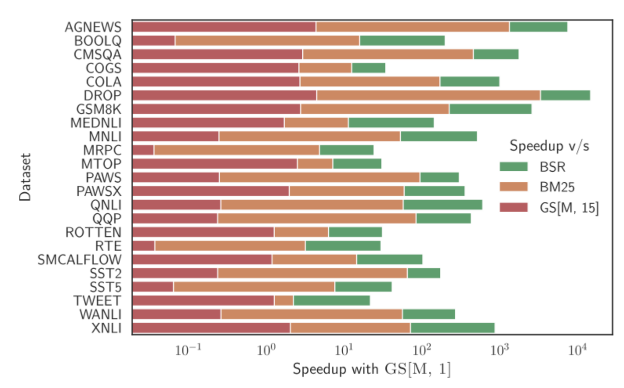


- GS[M] beats other training-free selection methods on **held-out** datasets.

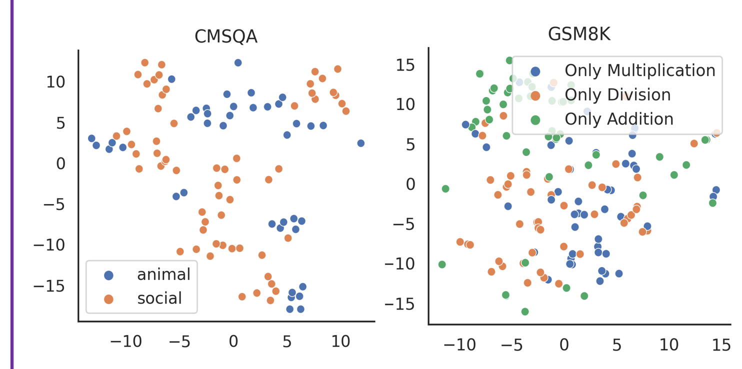


Analysis

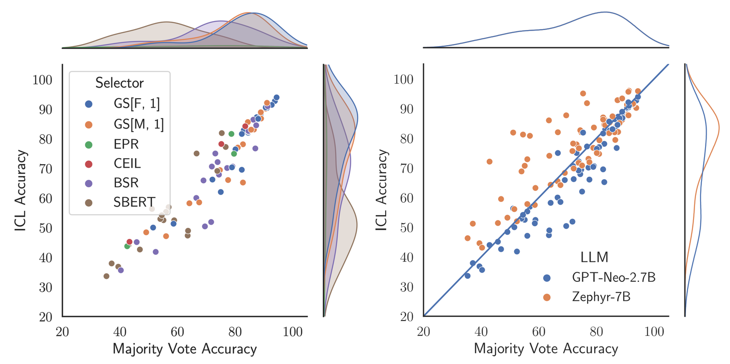
- Selection with single-token GistScore is up to 10000x faster than BSR[1] and even faster than BM25!



- Gist embeddings encode tasks-specific salient information
- CMSQA: Relevant question concepts
- GSM8k: Required arithmetic operations



- **(Left)** ICL on classification tasks is tied to the retriever accuracy itself. This seems true for all selection approaches.
- **(Right)** However, stronger LLMs seem less reliant on accurate retrieval, especially with weak retrievers.



- GistScore-based selection can improve ICL performance beyond that of the underlying gist model GM[F] itself.

	Method	SST5	QNLI	CSQA	SMC CG	SMC IID	COGS CG	COGS IID	GSM	DROP
	GM[F]	53.7	85.6	64.6	0.0	64.7	45.7	99.0	0.0	32.5
Zephyr	RAND	52.3	73.4	72.5	0.0	5.9	15.4	17.7	37.9	37.0
Zephyr	SBERT	51.2	72.1	71.6	13.4	50.8	39.7	55.4	35.9	46.3
Zephyr	GS[F, 1]	56.1	85.2	73.0	16.1	66.8	68.5	78.0	39.0	53.6