Dataset Distillation for Large Language Models 

Anirudh, Xindi, Sabhya

PRINCETON UNIVERSITY

Outline

Motivation

Related Work

Approach

Results & Analysis

Future Work

Motivation | What is dataset distillation?

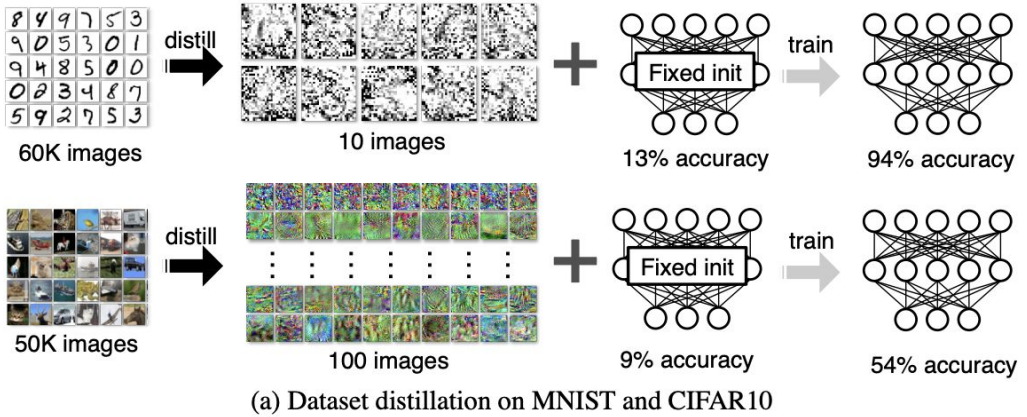
Dataset distillation aims to learn a small dataset which preserves the **critical information** from the original dataset.

Instead of training on the original dataset, we can use this condensed dataset to train a model, and have the model achieve **similar performance**.

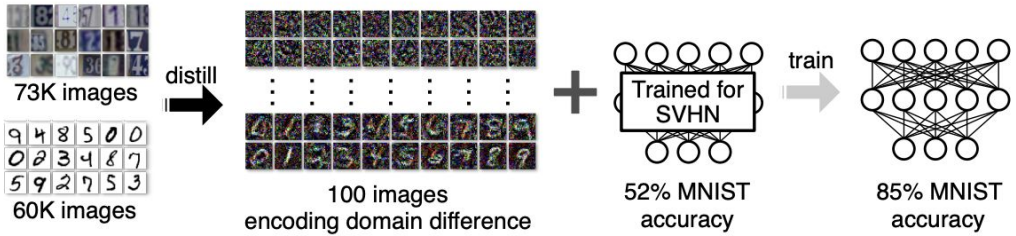
Motivation | Why is it important?

- Fundamentally a compression problem - **pragmatic compression**.
- Finding equivalent bits that can represent datasets + store information in a **continuous** environment.
- Example: after learning on two tasks, dataset distillation enables **merging knowledge**.
- Potential Applications: **continual learning, knowledge base update, adaptation tasks, etc,.**

Related Work | Vision Domain



(a) Dataset distillation on MNIST and CIFAR10



(b) Dataset distillation can quickly fine-tune pre-trained networks on new datasets

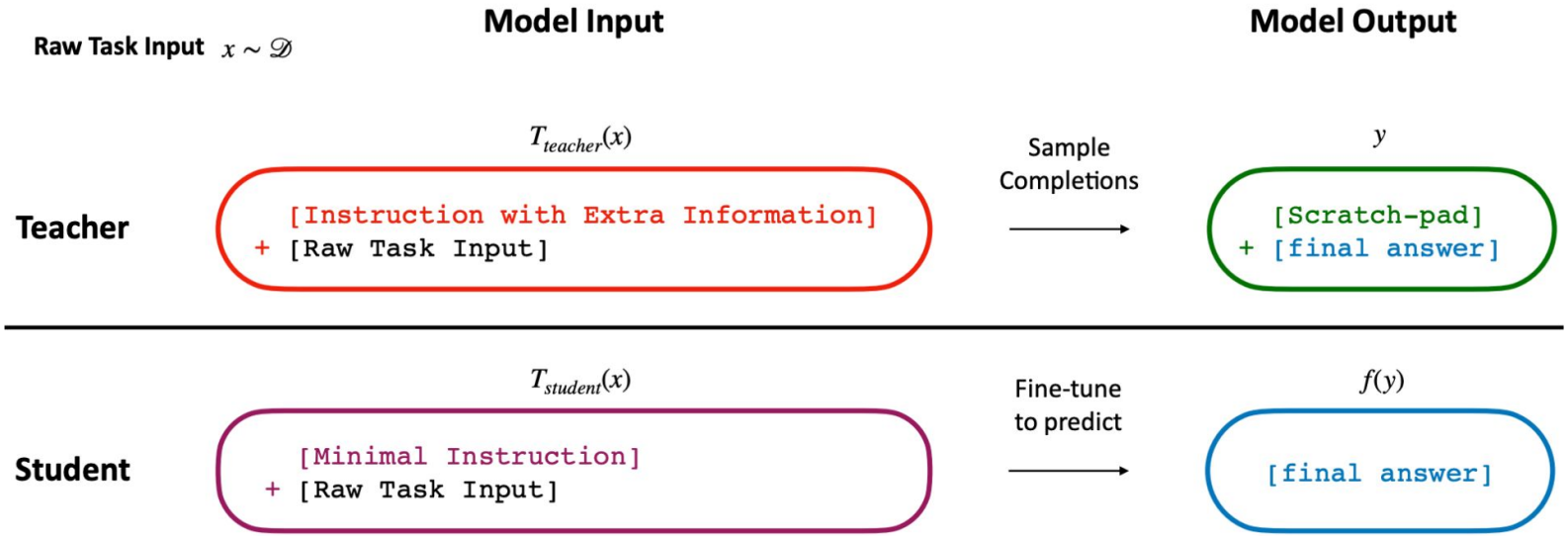
Wang, Tongzhou, et al. "Dataset distillation." *arXiv preprint arXiv:1811.10959* (2018).

Related Work | Language Domain

Distilled Sentence	Label Class 0	1	2	3	4	5
allan milk banned yellow planted successfully introduced bombay 1936 grass mines iron delhi 1942 male heir throne oath clouds 7th occur millennium smoking flows truth powder judiciary pact slim profit	**2.72**	-0.48	-0.07	-0.62	-0.53	-0.27
whom engineer grandfather joan officer entered victoria 1940s taxi romania motorcycle italian businessman photographer powerful driving u brilliant affect princess 1940s enemies conflicts southwestern retired cola appearances super dow consumption	-0.05	**3.21**	-1.15	-0.79	-0.71	-0.64
necessarily factors pronounced pronounced define bow destroying belonged balls 1923 storms buildings 1925 victorian sank dragged reputation sailed nn occurs darkness blockade residence traveled banner chef ruth rick lion psychology	-0.67	-0.66	**3.28**	-0.27	-0.77	0.47
accommodate accommodate peak 2.5 adults thin teenagers hike aged nurse policeman admit aged median philippines define baghdad libya ambassador admit baseman burma inning bills trillion donor fined visited stationed clean	-0.98	-0.14	-1.12	**5.57**	-0.85	-1.85

Sucholutsky, Ilia, and Matthias Schonlau. "Soft-label dataset distillation and text dataset distillation." *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.

Related Work | Language Domain

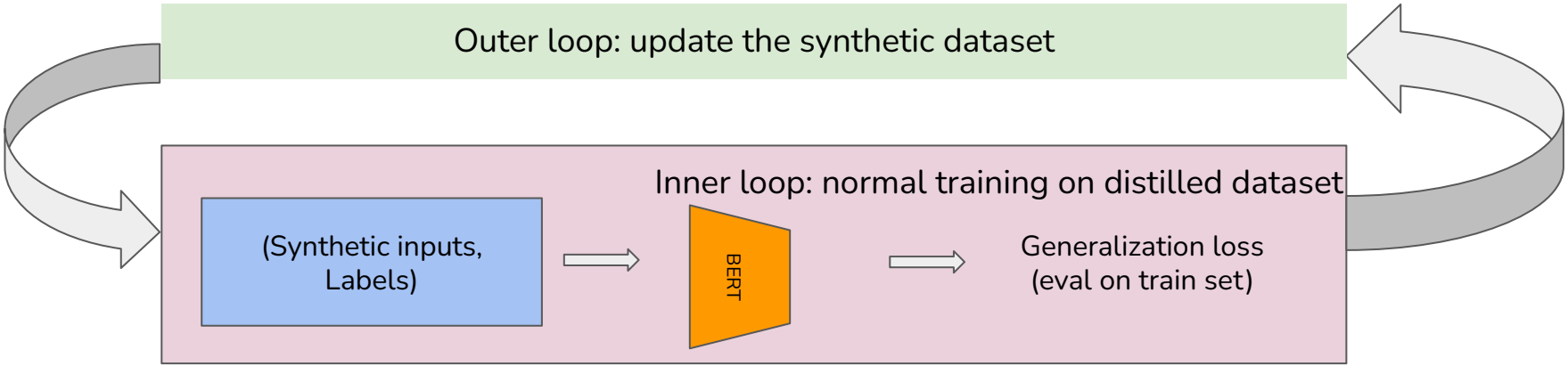


Snell, Charlie, Dan Klein, and Ruiqi Zhong. "Learning by distilling context." *arXiv preprint arXiv:2209.15189* (2022).

Approach | Create few synthetic inputs per output class

- Focus on classification tasks for now.

Approach | Trainable synthetic inputs



Approach | Few synthetic inputs

- Few synthetic inputs
  - |C| classes
  - k synthetic inputs per class
  - Each input contains 512 tokens

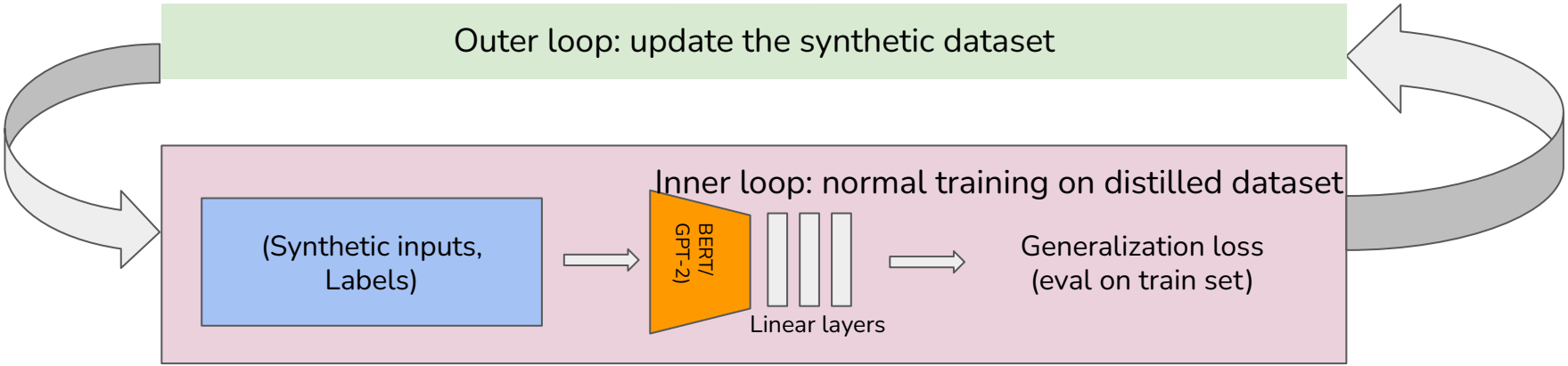
$$\mathbf{v}_i^c \in \mathbb{R}^{512\times 768} \qquad \begin{matrix} c \in \{1,\ldots,|C|\} \\ i \in \{1,\ldots,k\} \end{matrix}$$

$$\mathbf{v}_i^c = [w_1, w_2, \ldots, w_{512}]_i^c$$

- Number of parameters:
  - $|C| \cdot k \cdot 512 \cdot 768$

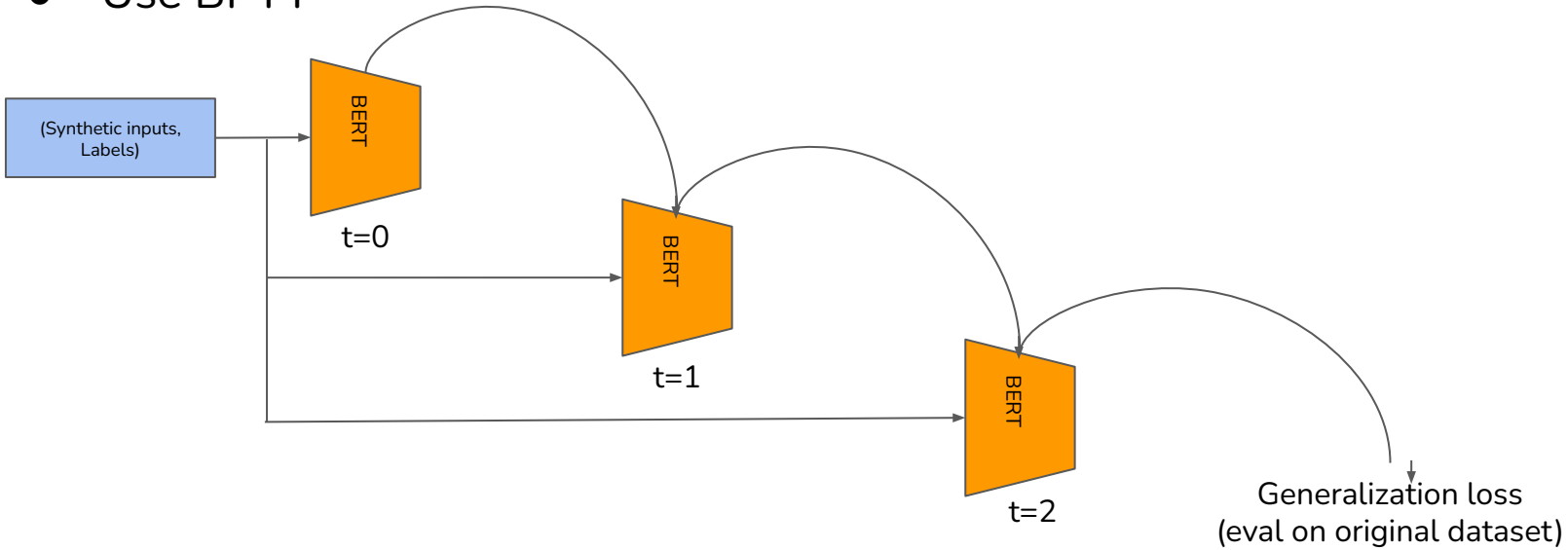
Approach

$$\frac{\partial \mathcal{L}}{\partial \phi}$$



Approach | Back-propagation through time for $\frac{\partial \mathcal{L}}{\partial \phi}$

- Handling gradient computation through multiple inner-loop iterations is non-trivial.
- Use BPTT



Approach | Convert back to text?

- Outerloop trains synthetic input embeddings.
- Convert back to tokens using nearest neighbour search over token embeddings.
- Finally, use for downstream applications $\{(x_i'^c, y^c)\}$
  - Train (much) quicker using distilled dataset.
  - Combine with other distilled datasets to create super-dataset?
  - Continual learning / Knowledge-base updates?
  - Theoretically interesting!

Baselines | Settings

- We compare our method with the following baselines:
  - **Fine-Tuned BERT Base**
  - **Fine-Tuned DistilBERT**
  - **Fine-Tuned RoBERTa**
- We test our method on the following datasets:
  - **AG News, SST2, IMDB, Rotten Tomatoes**
  - We choose classification tasks because of their primitive nature

Baselines | Accuracy Scores

Model / Dataset	AG News	IMDB	SST-2	Rotten Tom.
DistilBERT	92.97	87.32	80.74	78.71
BERT Base	94.46	93.86	91.86	84.71
RoBERTa	94.98	95.11	93.35	88.46
BERT Base	pending	pending	pending	pending
DistilBERT	pending	pending	pending	pending
RoBERTa	pending	pending	pending	pending

Fine-tuned on entire dataset
Fine-tuned on distilled dataset

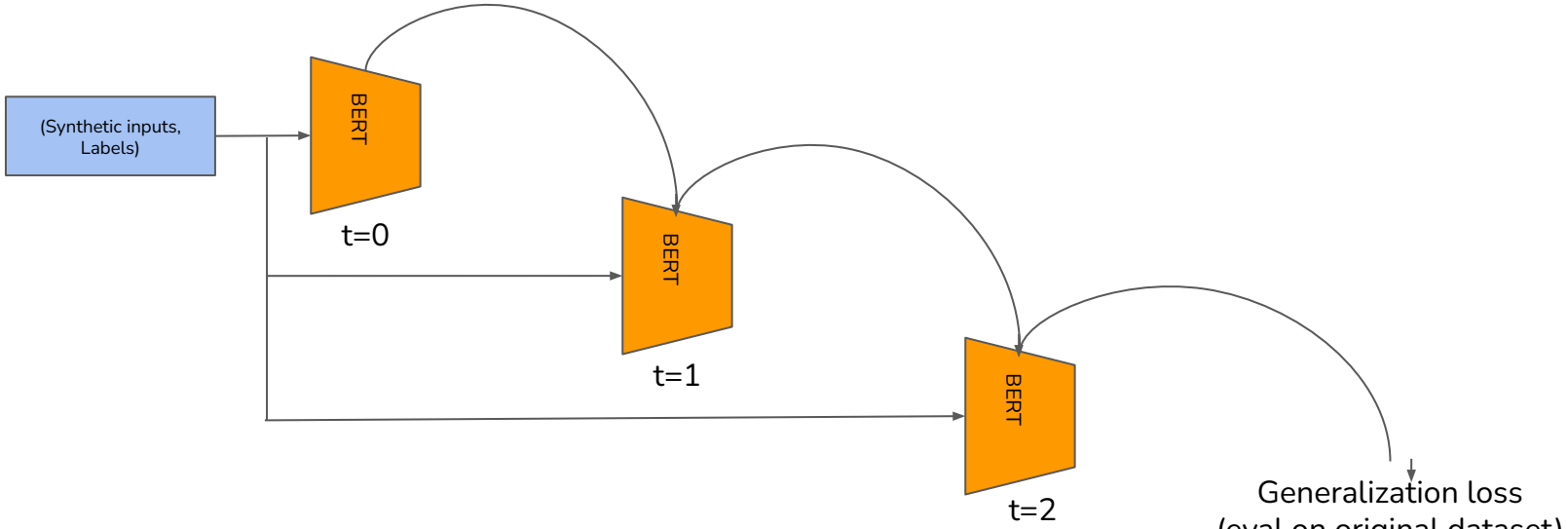
Prompting Experiment | Few-shot Capabilities?

- Do distilled examples make for good prompts?
- Compare few-shot performance of using distilled examples within prompts for GPT family models

Analysis Experiments | Going Further

- We hope to run the following analysis after we have our results
  - How does performance scale with number of distilled examples / class
  - Compare compute usage of fine-tuning distilled examples vs. full dataset
  - Qualitative analysis of distilled examples vs regular examples
- Future Work
  - Extend methodology to more complex tasks
  - Extend distillation method to use optimizers other than SGD

Thank You


(Synthetic inputs, Labels)
BERT
t=0
BERT
t=1
BERT
t=2
Generalization loss
(eval on original dataset)