Efficient Multi-Task Auxiliary Learning: Selecting Auxiliary Data by Feature Similarity

National Taiwan University
國立臺灣大學

http://github.com/MiuLab/FastMTL

Po-Nien Kung, Yi-Cheng Chen, Sheng-Siang Yin
Tse-Hsuan Yang, Yun-Nung (Vivian) Chen

Outline

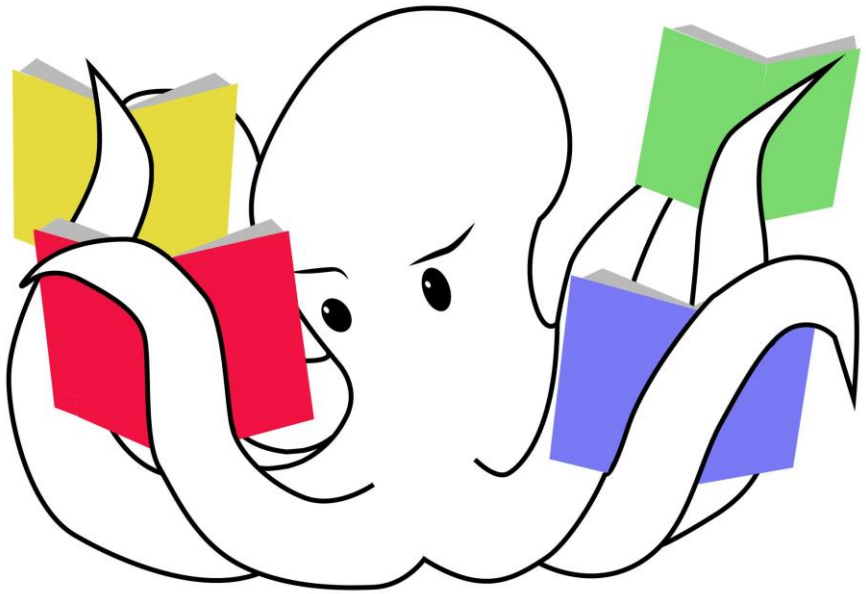
- Background
- Two-Stage Multi-Task Auxiliary Learning
- Efficient Data Selection
- Experiments
- Conclusion

Outline

- Background
  - What is **Multi-Task Auxiliary Learning**
- Two-Stage Multi-Task Auxiliary Learning
- Efficient Data Selection
- Experiments
- Conclusion

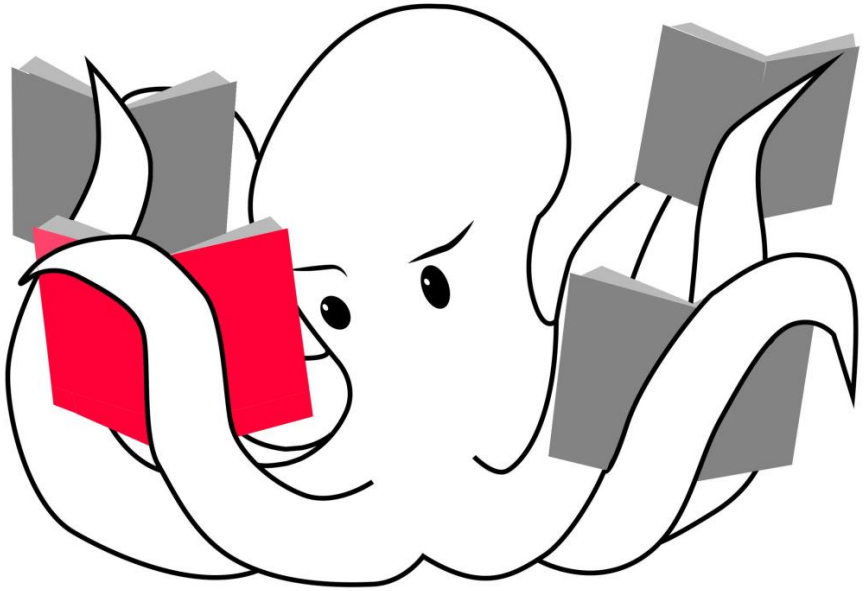
Background: Multi-Task Auxiliary Learning

- Multi-task learning



All tasks are important!

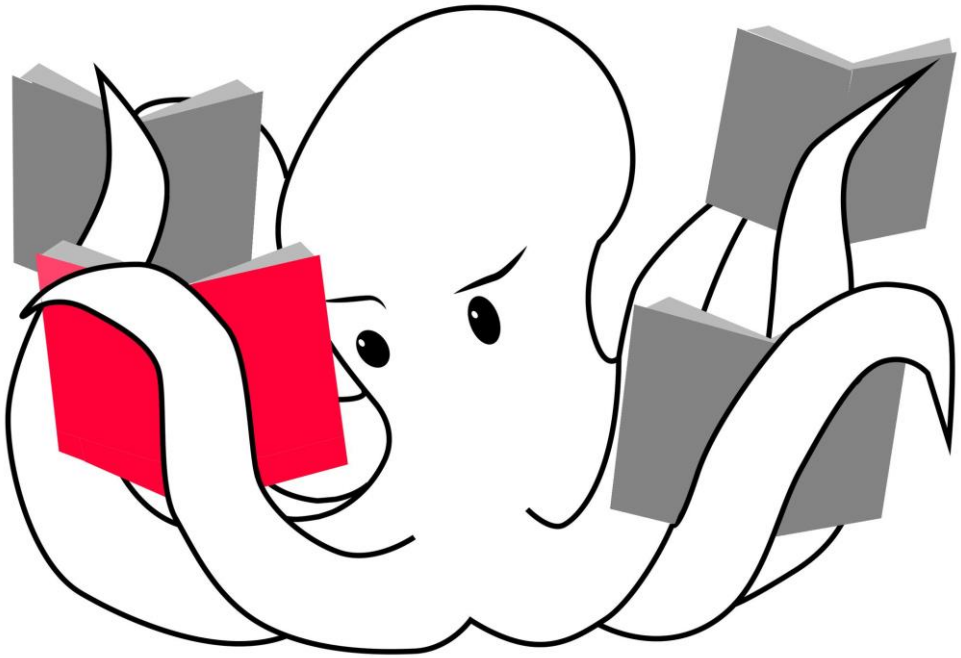
- Multi-task auxiliary learning



One primary task and multiple auxiliary tasks.

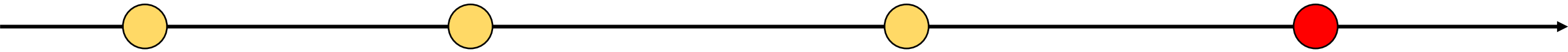
Background: Multi-Task Auxiliary Learning

- To achieve **better performance** on the primary task
- More useful when the size of the primary task is **small**



Background: Multi-Task Learning Chronology

More tasks, more **power**…



MTDNN	*T5*	*Muppet*	There will be **more…**
MSR	Google	Facebook	new NLP tasks
8 tasks	**18** tasks	**50** tasks	&
960k data	GLUE, Super GLUE, WMT…	**4.8mil** data	multi-modal tasks

Background: Auxiliary Data Size

More tasks, more **power**… Also, more **computing**!

Compare **single-task finetuning** and **multi-task auxiliary learning** when the primary task is **RTE:**

In *MTDNN* setting…

RTE (2.4k) vs **Auxiliary(960k)** data

400x Computing cost!

In *Muppet* setting…

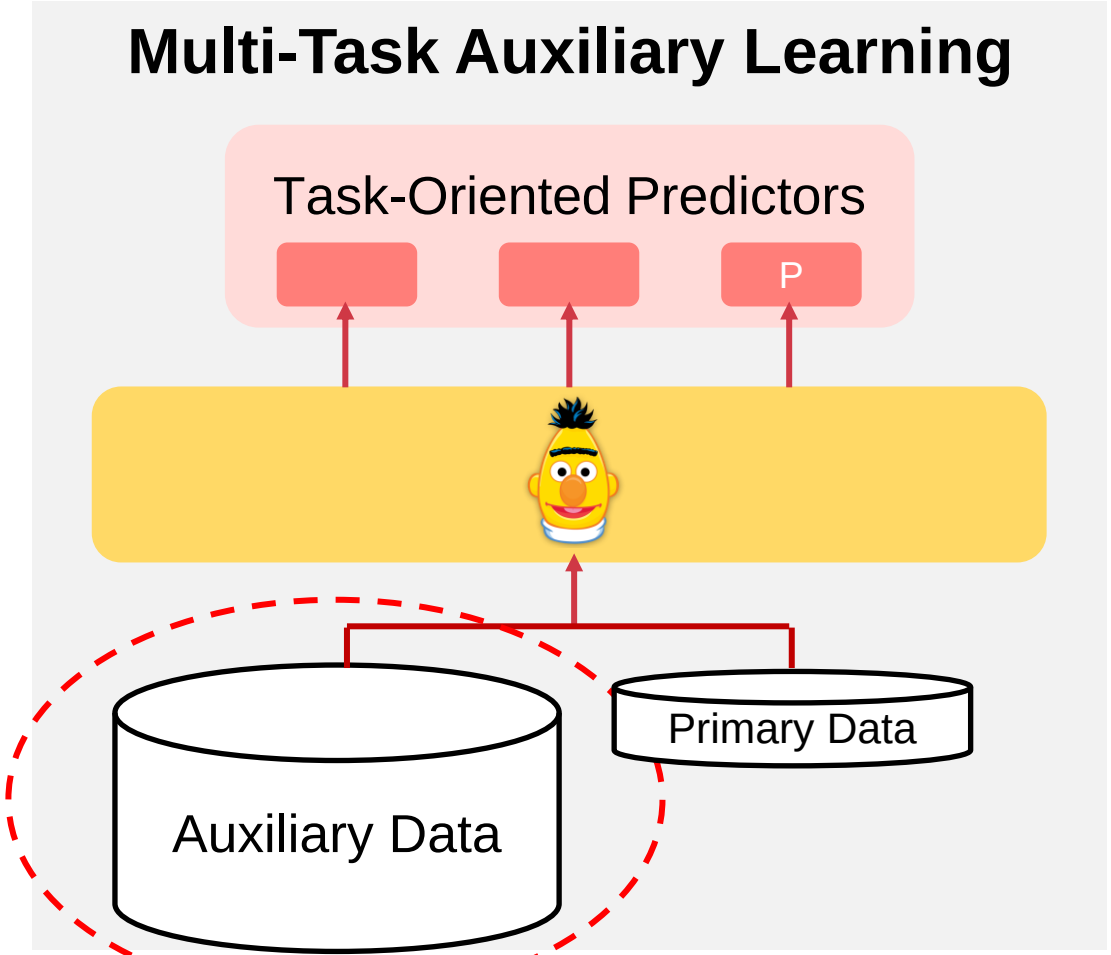
RTE (2.4k) vs **Auxiliary (4.8mil)** data

2000x Computing cost!

Outline

- Background
  - What is **Multi-Task Auxiliary Learning**
  - Why we need an efficient **Multi-Task Auxiliary Learning** method
- Two-Stage Multi-Task Auxiliary Learning
- Efficient Data Selection
- Experiments
- Conclusion

Inefficient Multi-Task Auxiliary Learning



Less Auxiliary Data is Possible?

The first question: Are all auxiliary data **beneficial**?

- Single-task vs. multi-task
  - **GLUE** dataset (similar with **MTDNN**)
  - **10** random seeds
  - Use **1 STDEV** as the threshold for performance **IMPROVED** and DROPPED Otherwise, NEUTRAL.

MNLI	RTE	MRPC	STS-B	QQP	QNLI	SST-2	CoLA

Some auxiliary dataset might be **unhelpful** or even **harmful**!

Negative Transfer !

Less Auxiliary Data is Possible?

The first question: Are all auxiliary data **beneficial**?

- Single-task vs. multi-task
  - **GLUE** dataset (similar with **MTDNN**)

MNLI	RTE	MRPC	STS-B	QQP	QNLI	SST-2	CoLA
—	👍	👍	👍	—	—	👎	👎

Some auxiliary dataset might be **unhelpful** or even **harmful**!

Negative Transfer !

Outline

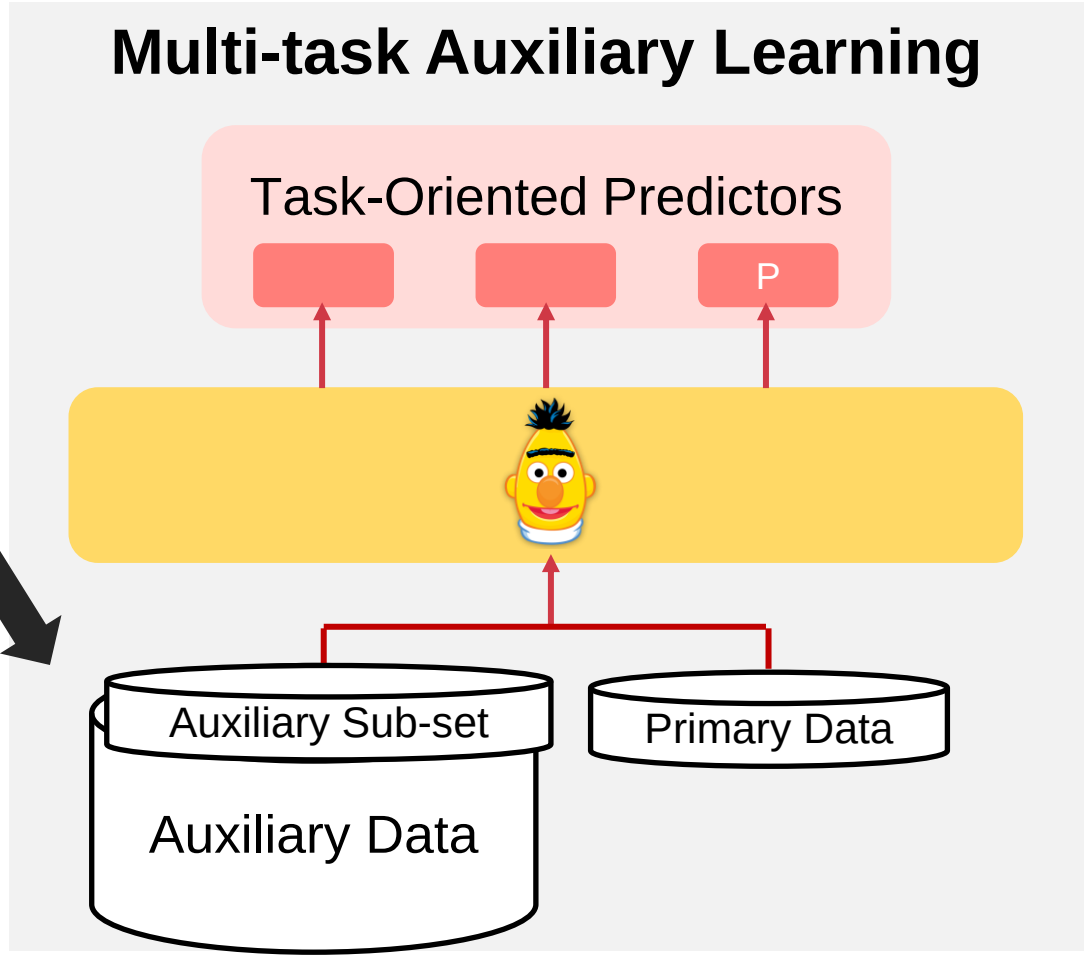
- Background
- Two-Stage Multi-Task Auxiliary Learning
  - We need a **data-sampling** method to shrink the size of the auxiliary data
- Efficient Data Selection
- Experiments
- Conclusion

Two-Stage Multi-Task Auxiliary Learning

Sampling Method

Select the most **beneficial** auxiliary data

Goal: **reducing** the cost of training auxiliary data



Prior Work: AutoSeM (Guo et al., 2019)

- Idea: automatically select the most **beneficial (related)** auxiliary tasks
  - Beta-Bernoulli multi-armed bandit with Thompson Sampling
- Decide the **mixing ratio** of auxiliary tasks
  - Gaussian Process
  - Trial and error

Avoid **negative transfer** and further **improve the performance**!

Target 1: Reduce auxiliary dataset size ? ○

Target 2: Reduce the total computing cost ? ✕

Why?

The sampling method itself is **Computationally Expensive**!

Challenges of the Sampling Method

Prior work

Our target

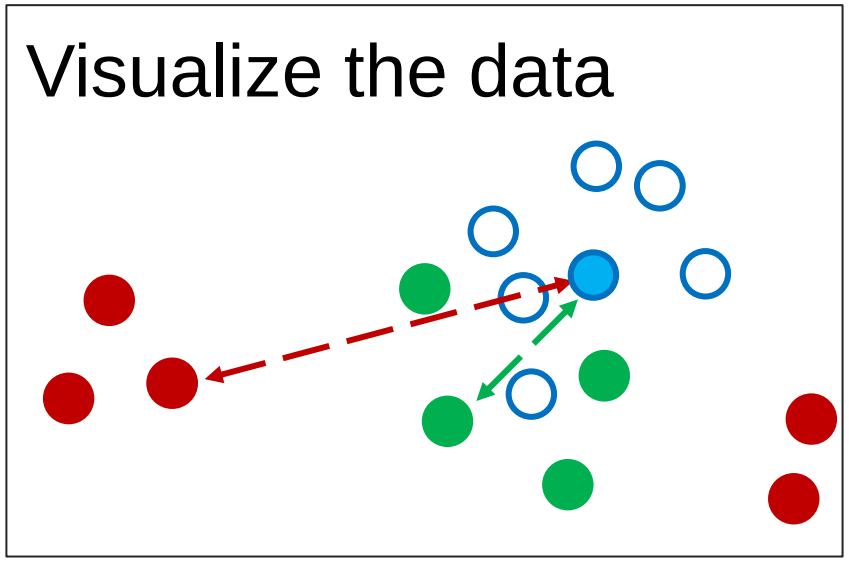
Outline

- Background
- Two-Stage Multi-Task Auxiliary Learning
- Efficient Data Selection
  - Select the most beneficial auxiliary data by **feature similarity**
- Experiments
- Conclusion

Feature Similarity Assumption

Assumption: More **similar** is an auxiliary data to the primary task, more **benefit** it can bring.

- Primary task data
- Pseudo centroid of primary task data
- Useful auxiliary data
- Neutral/Harmful auxiliary data

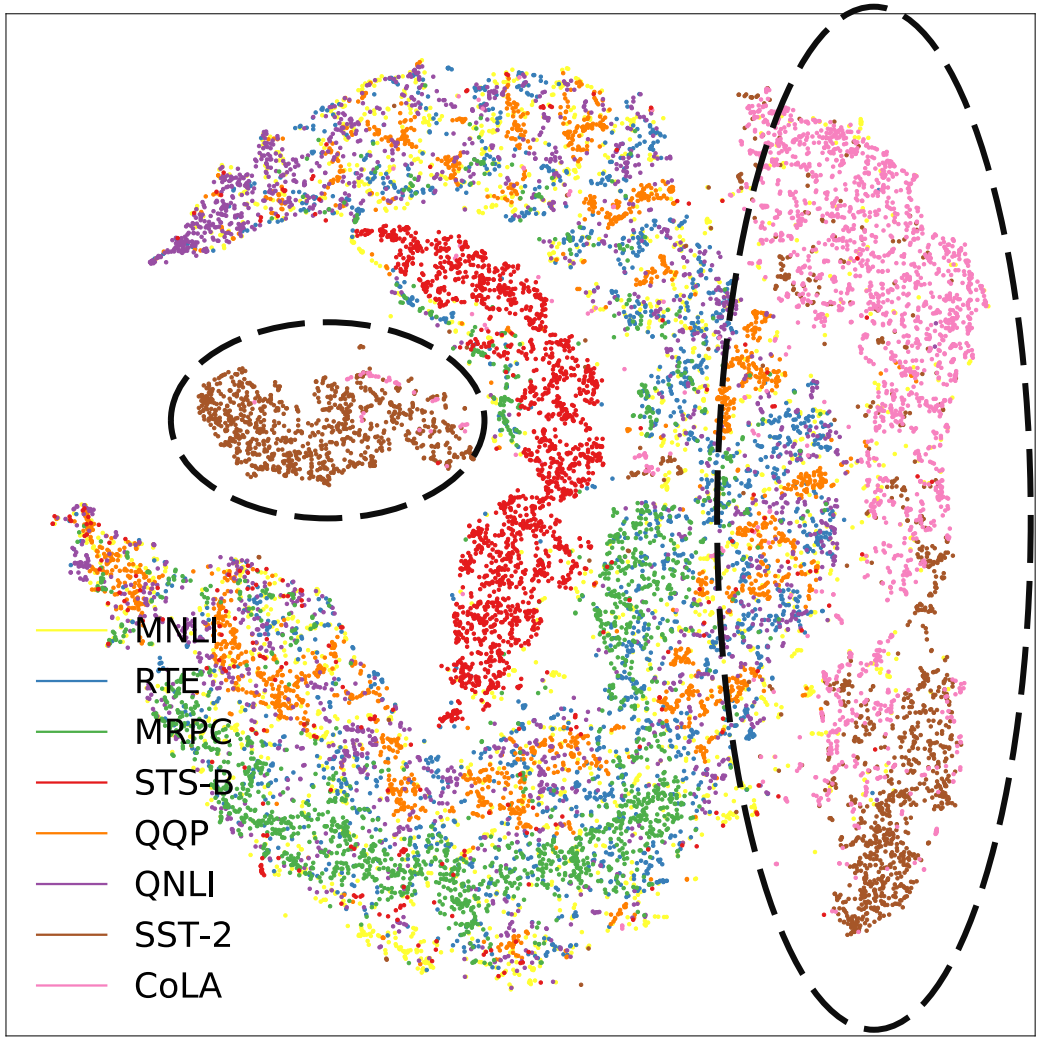


Feature Similarity Assumption

- Toy experiment
  - MT-DNN setting: multi-task train **500** data for each **GLUE** task
- T-SNE Visualization
  - **Last hidden state features** of BERT

MNLI	RTE	MRPC	STS-B	QQP	QNLI	SST-2	CoLA
—	👍	👍	👍	—	—	👎	👎

SST-2 and CoLA are more separate from others.



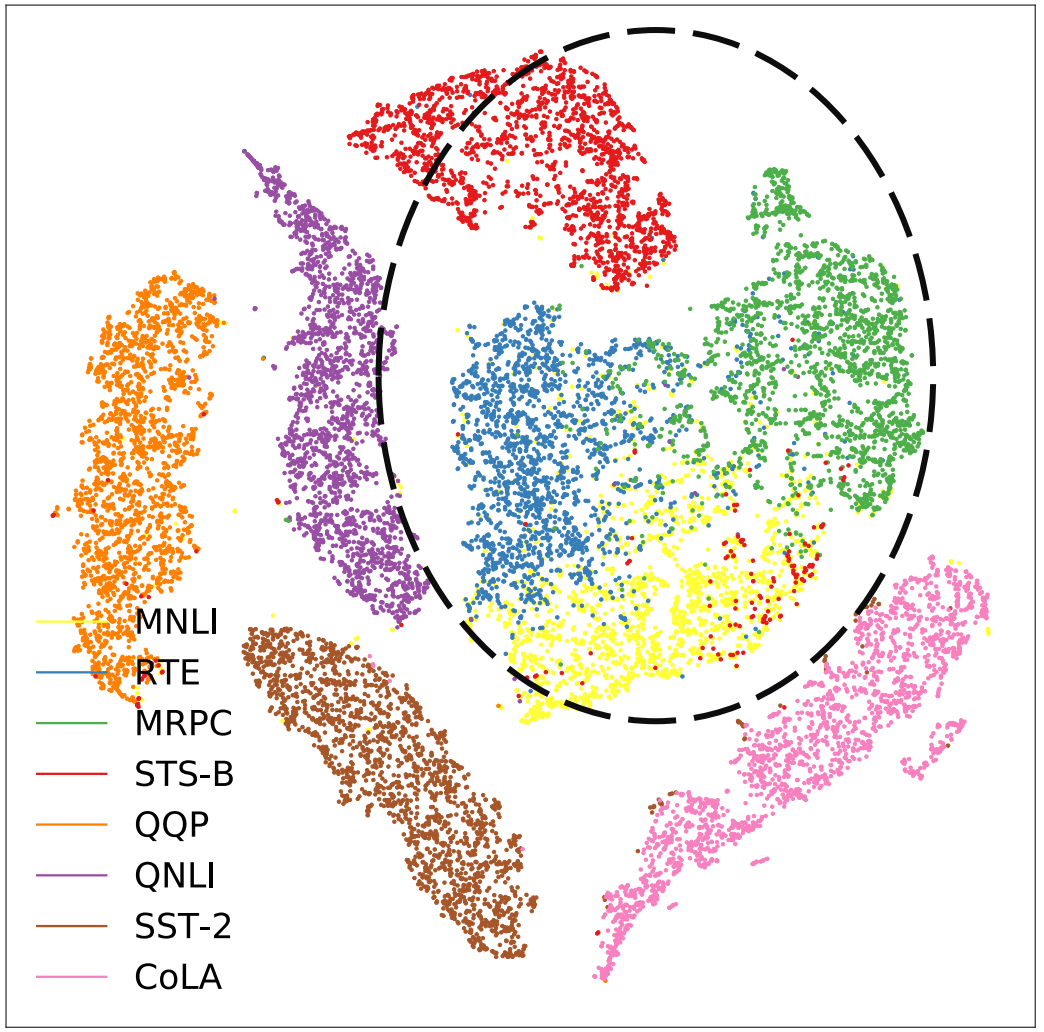
Feature Similarity Assumption

- Adding task-discriminator
  - Force the model to encode more task information into features

MNLI	RTE	MRPC	STS-B	QQP	QNLI	SST-2	CoLA
—	👍	👍	👍	—	—	👎	👎

STS-B, RTE, MRPC and MNLI have more data **overlap** with each other.

The tasks with **more similar auxiliary data** improve most!



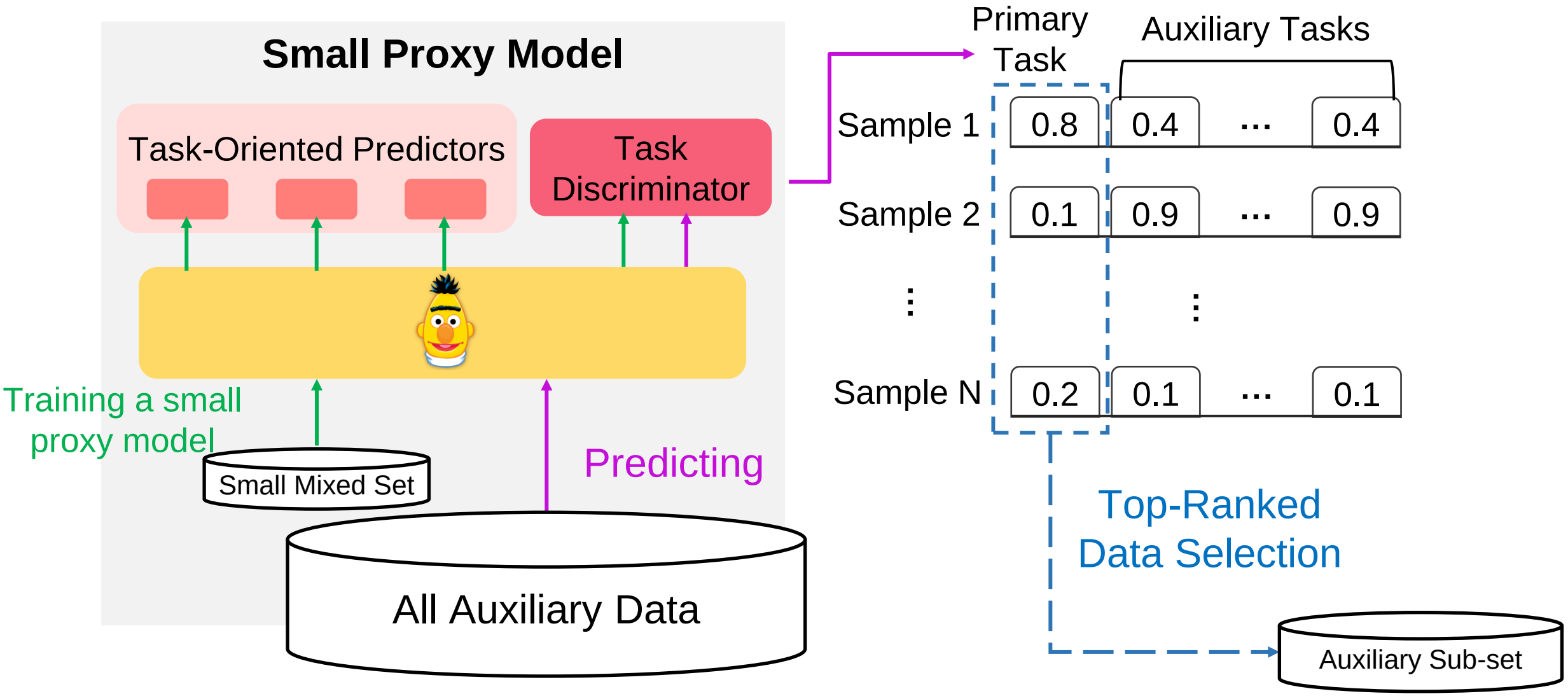
Usefulness of Auxiliary Data

- **Feature similarity** may indicate the **usefulness** of auxiliary data to a primary task.

How do we get the rank the **feature similarity**?

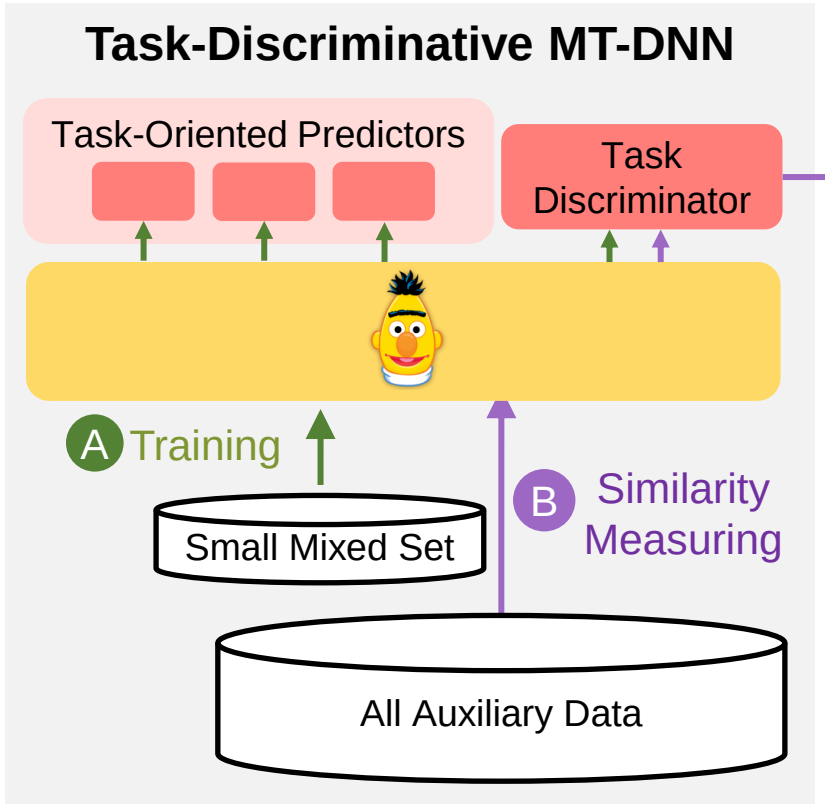
Train a **small proxy model** with **a task discriminator** to predict similarity.

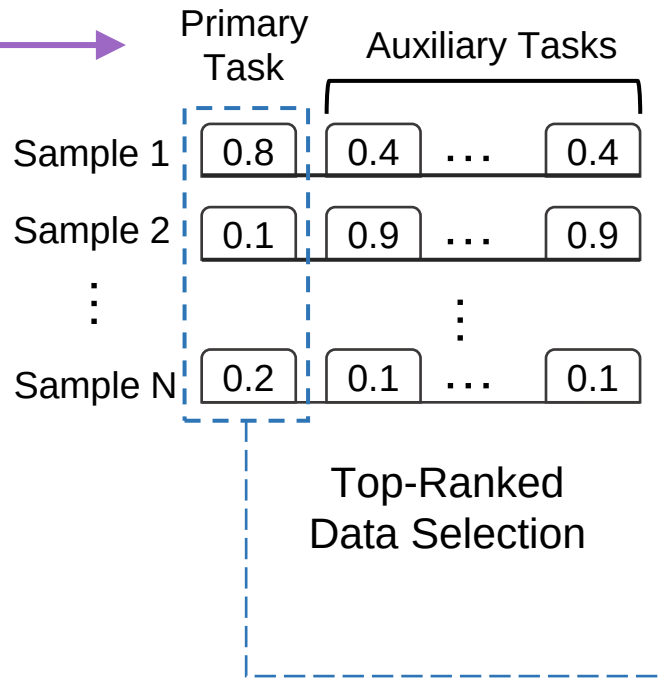
Data Selection: Similarity Ranking



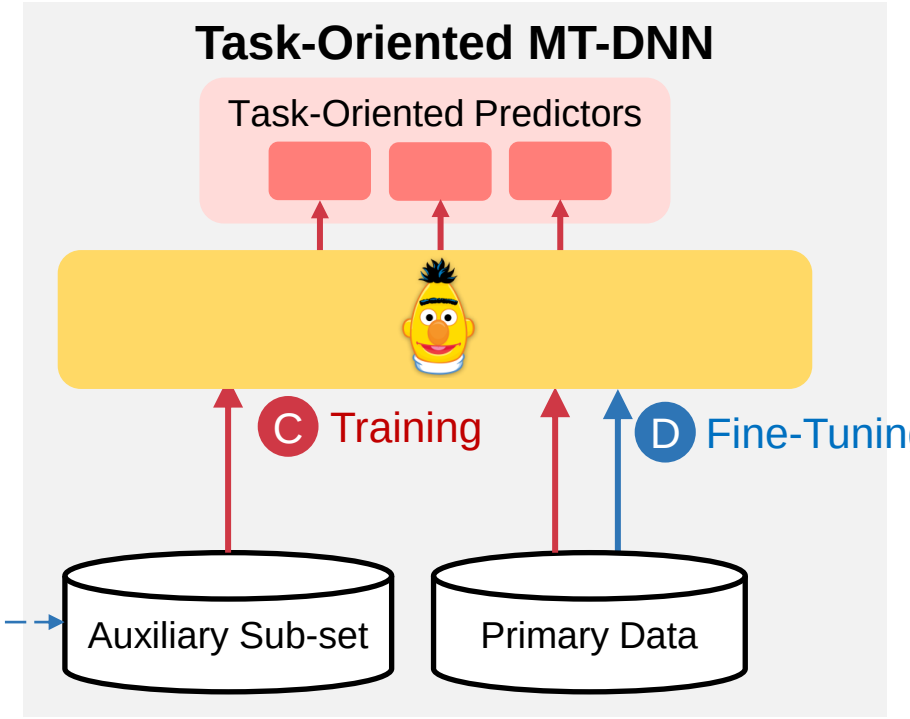
Two-Stage Multi-Task Auxiliary Learning

Stage 1: Similarity Ranking





Stage 2: Multi-task Auxiliary Learning & Fine-tuning



Goal: multi-task auxiliary learning on **less auxiliary data** but **comparable (or even improved) performance**

Efficient!

Outline

- Background
- Two-Stage Multi-Task Auxiliary Learning
- Efficient Data Selection
- Experiments
- Conclusion

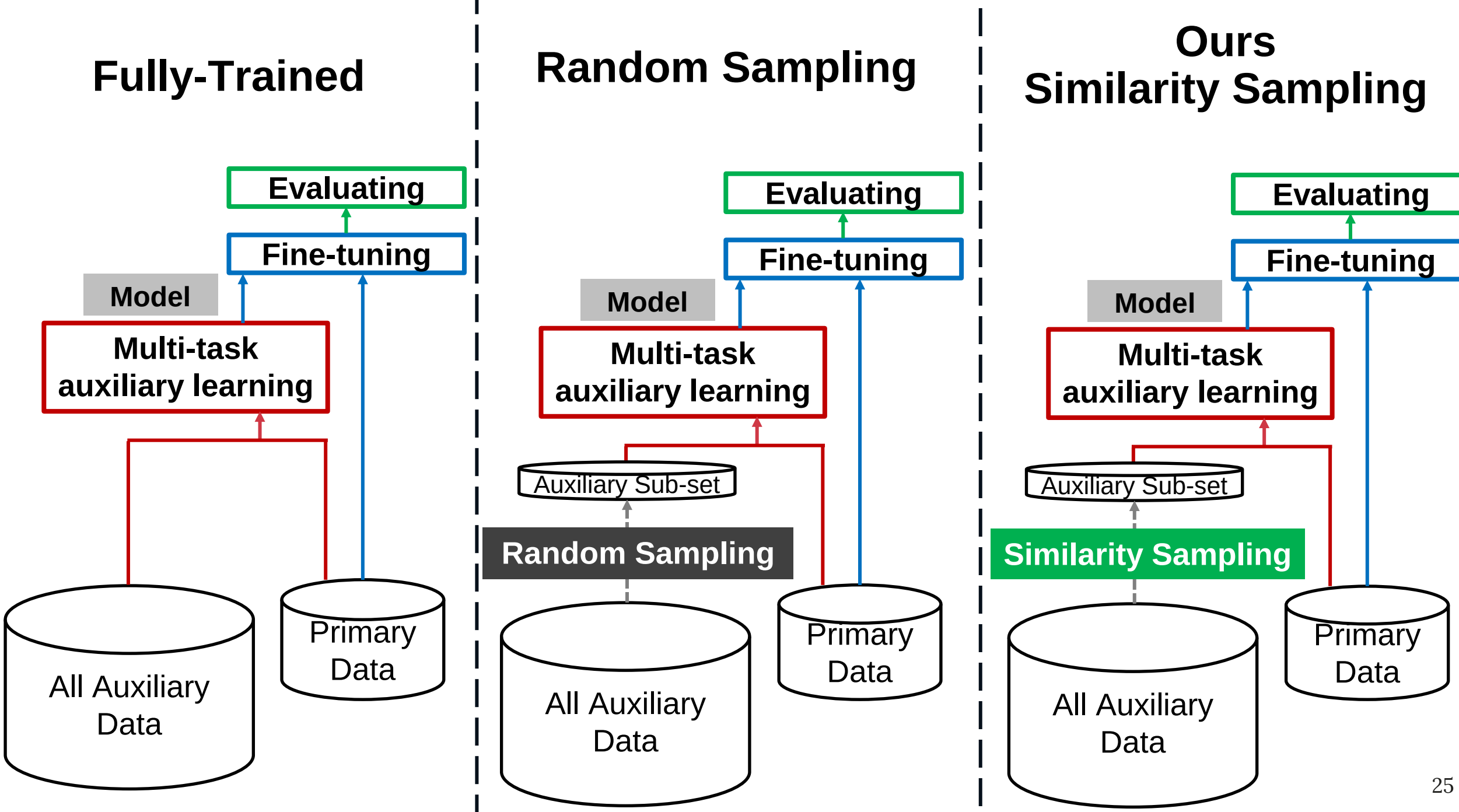
Setting

- Similar with **MT-DNN**
- Data: GLUE (**960K**)
- Model: Bert-base
- Baselines:
  - **No-MTL** (Weak)
  - **Random Sampling** (Surprisingly Strong)
  - **Fully-trained** (Strong)

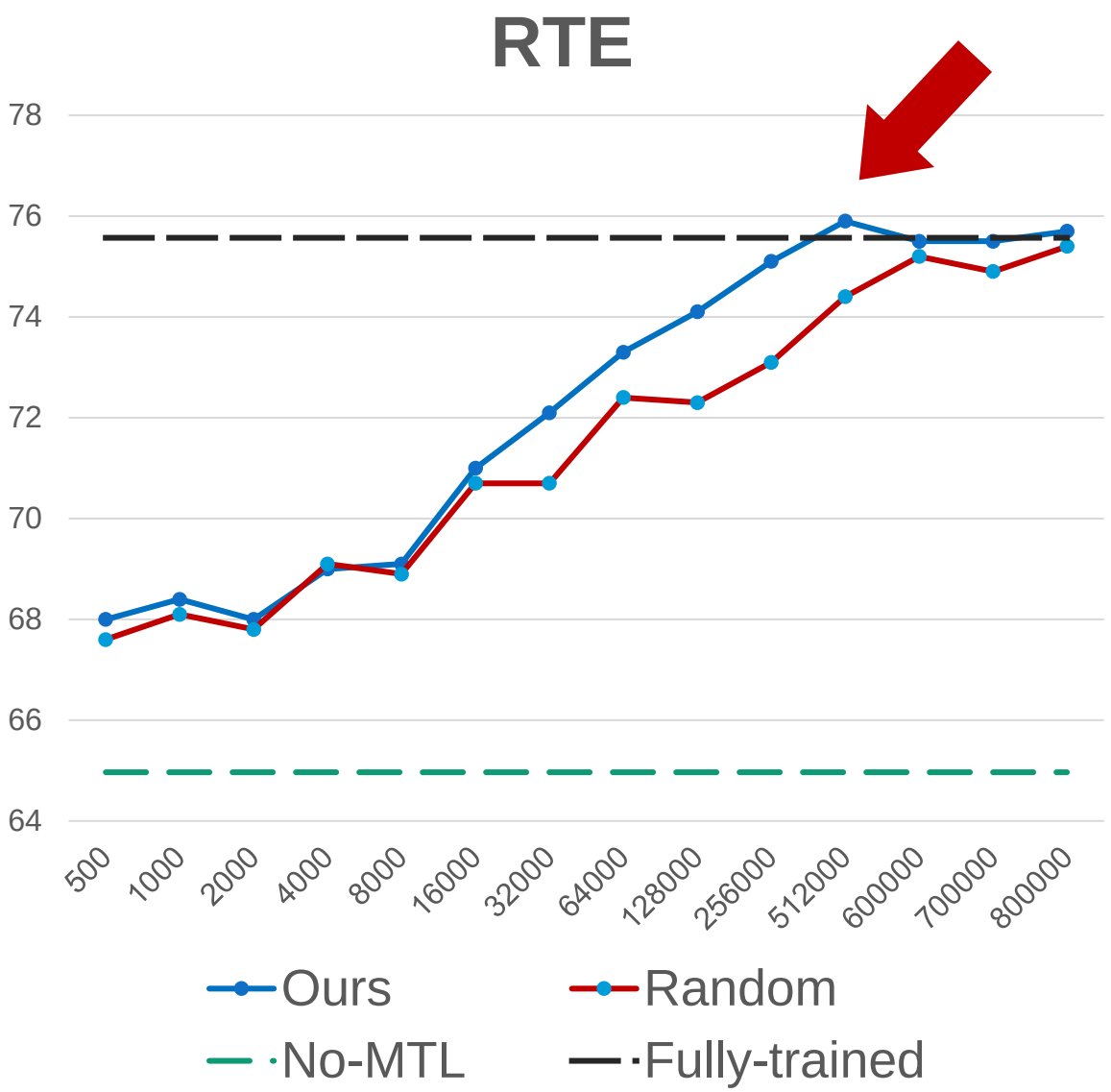
Primary tasks

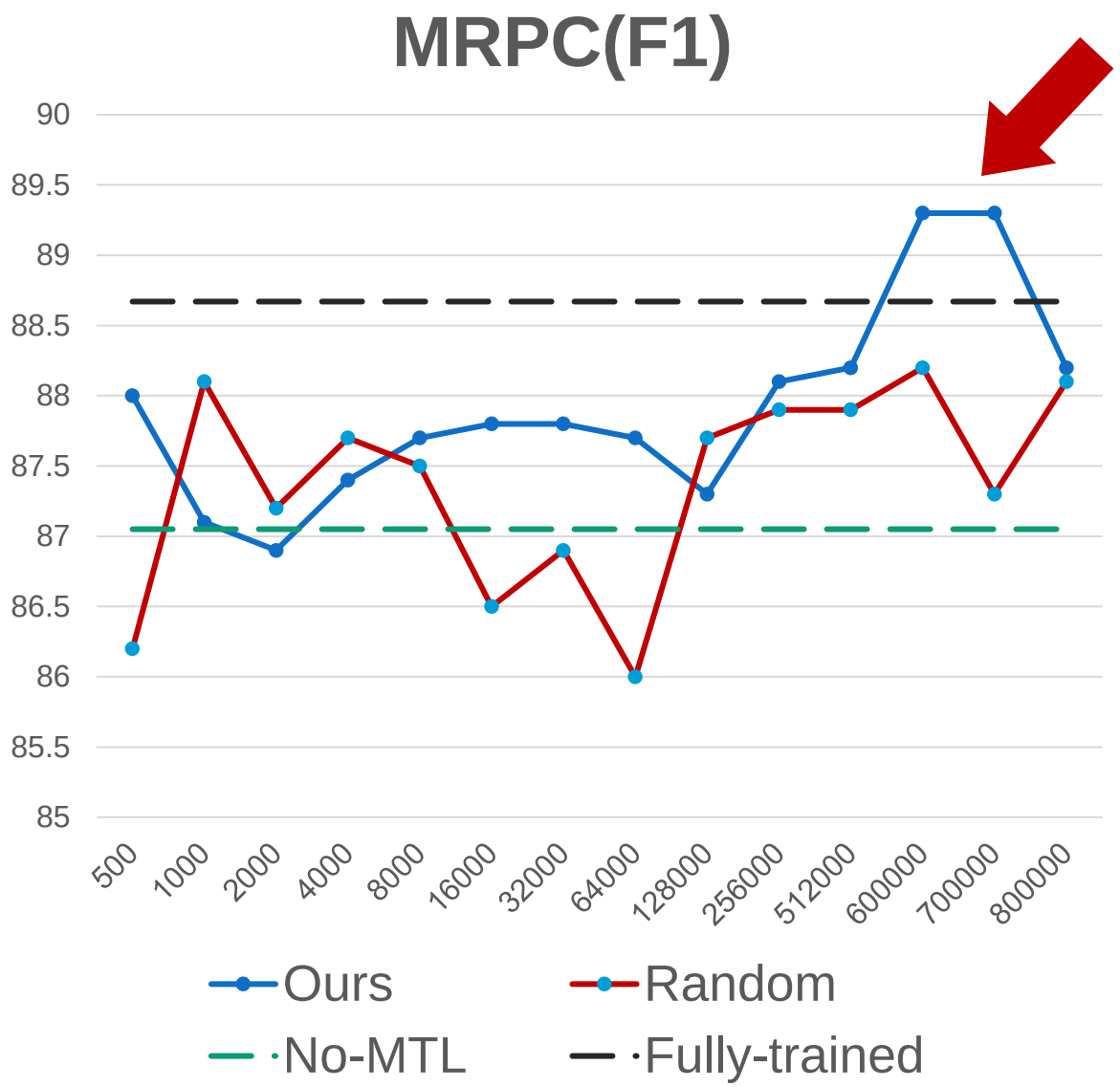
MNLI	RTE	MRPC	STS-B	QQP	QNLI	SST-2	CoLA
—	👍	👍	👍	—	—	👎	👎

They are improved by MTL, so there exist useful data in auxiliary tasks

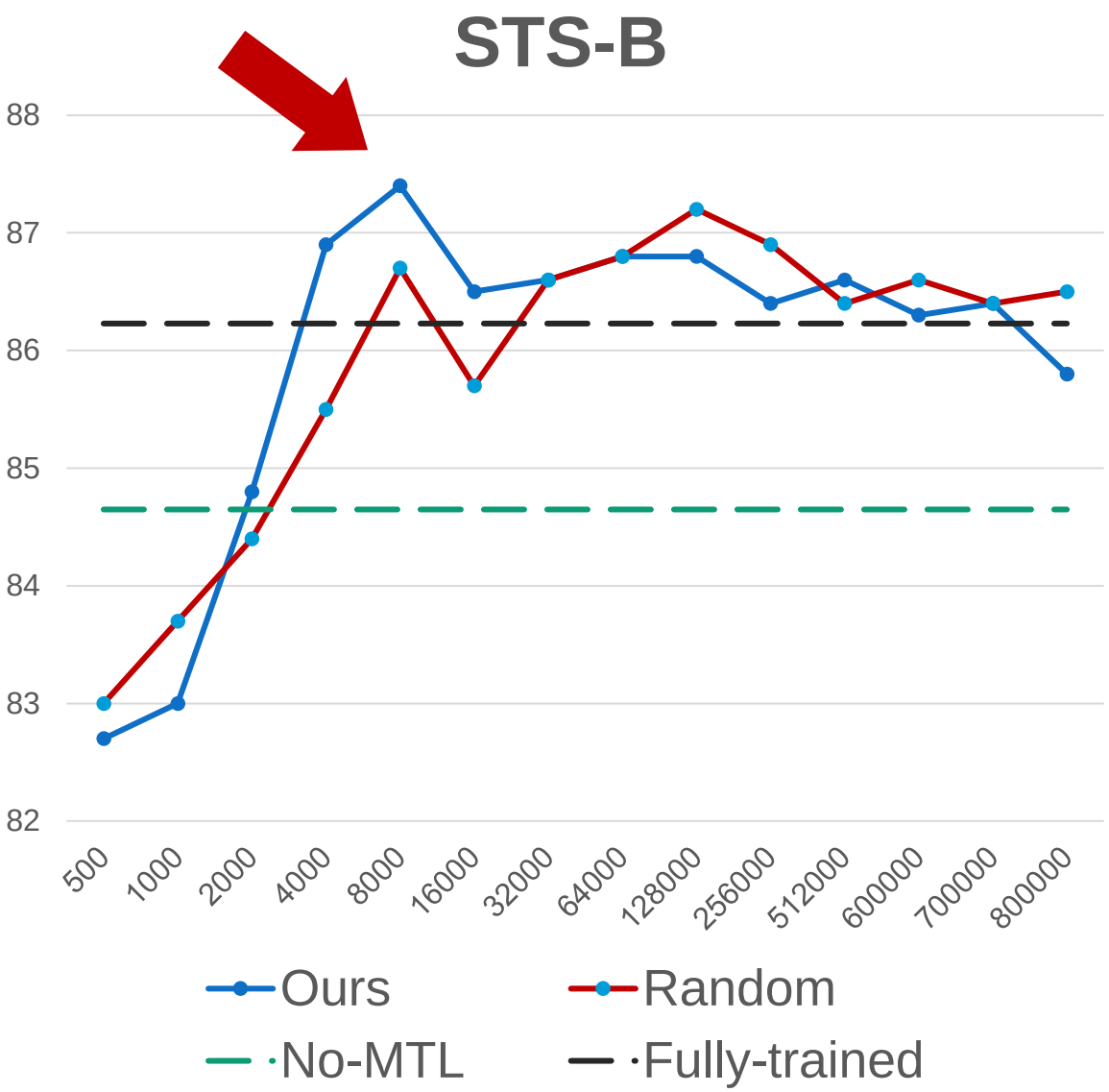

Fully-Trained
Evaluating
Fine-tuning
Model
Multi-task auxiliary learning
All Auxiliary Data
Primary Data
Random Sampling
Evaluating
Fine-tuning
Model
Multi-task auxiliary learning
Auxiliary Sub-set
Random Sampling
All Auxiliary Data
Primary Data
Ours
Similarity Sampling
Evaluating
Fine-tuning
Model
Multi-task auxiliary learning
Auxiliary Sub-set
Similarity Sampling
All Auxiliary Data
Primary Data

Results


RTE
78
76
74
72
70
68
66
64
500
1000
2000
4000
8000
16000
32000
64000
128000
256000
512000
600000
700000
800000
Ours
Random
No-MTL
Fully-trained


MRPC(F1)
90
89.5
89
88.5
88
87.5
87
86.5
86
85.5
85
500
1000
2000
4000
8000
16000
32000
64000
128000
256000
512000
600000
700000
800000
Ours
Random
No-MTL
Fully-trained

Results



Findings

Ours > Random

Ours > Fully-Trained

Random > Fully Trained (STS-B)

Our method can use **less data** to achieve **better performance**!

Efficiency Evaluation

- **How many auxiliary data is needed to surpass fully-trained (RTE, MRPC, STS-B)?**

  Ours: 50%, 60%, 0.05% Random: 100%, 100%, 1%

STS-B Runtime(s)	Similarity Sampling		Multi-Task Auxiliary Learning		Total	How much faster?
	Training small proxy model	Predict similarity	MTL	Finetuning		
Fully-trained	--		**15801**	190	15991	--
Random			260		450	35x
Ours	95	775			670	23x

Outline

- Background
- Two-Stage Multi-Task Auxiliary Learning
- Efficient Data Selection
- Experiments
- Conclusion

Conclusion

- Address the **efficiency importance** in multi-task auxiliary learning
- Propose a **data sampling** method to shrink the size of the auxiliary data → computing cost reduction
- First use **feature similarity** to determine the data usefulness
- Our method outperforms random sampling and further surpass fully-trained model using less data

First work for **time-efficiency of multi-task auxiliary learning**:

http://github.com/MiuLab/FastMTL