Large-scale Multi-task Pre-training

Zhuosheng Zhang
zhangzs@sjtu.edu.cn
https://bcmi.sjtu.edu.cn/~zhangzs

Introduction

Zhuosheng Zhang

- **Education Background**
  - 2012-2016: B.S in Wuhan University
  - 2016-2020: M.S. in Shanghai Jiao Tong University
  - 2020-2023: Ph.D. in Shanghai Jiao Tong University, Advisor: Prof. Hai Zhao
  - 2022: Internship at MSR from Feb. 2022 to Now, Mentor: Shuohang Wang
- **Research Interests**
  - Natural Language Processing
  - Pre-trained Language Models
  - Machine Reading Comprehension

Overview: Large-scale Multi-task Pre-training

- **Goals**
  - Bridge self-supervised pre-training with task requirements by leveraging **large-scale supervised tasks**
  - Use a unified model to solve a wide range of tasks
- **Benchmark Tasks**
  - Commonsense Reasoning (Rainbow)
  - Legal Language Understanding (LexGLUE)
- **Scientific Questions**
  - How to capture task relationships in large-scale multi-task pre-training
- **Contributions**
  - A unified **encoder-only** multi-task pre-trained langauge model trained on 40 tasks
  - A probing tool of using task prefix to explore the **task relationships** in large-scale MTL
  - **Human-parity performance** on commonsense reasoning leaderboards.

Benchmark Tasks

- **Rainbow:** develop models that use **commonsense knowledge** to answer multiple-choice questions.

Dataset	Goal
ANLI	Abductive reasoning in narratives. It asks models to identify the best explanation among several connecting a beginning and ending
COSMOSQA	asks commonsense reading comprehension questions about everyday narratives
HELLASWAG	requires models to choose the most plausible ending to a short context
PIQA	a multiple-choice question answering benchmark for physical commonsense reasoning
SOCIALIQA	evaluates commonsense reasoning about social situations and interactions.
WINOGRANDE	a large-scale collection of Winograd schema-inspired problems requiring reasoning about both social and physical interactions.

goal (string)	sol1 (string)	sol2 (string)	label (class label)
When boiling butter, when it's ready, you can	Pour it onto a plate	Pour it into a jar	1 (1)
To permanently attach metal legs to a chair, you can	Weld the metal together to get it to stay firmly in place	Nail the metal together to get it to stay firmly in place	0 (0)
how do you indent something?	leave a space before starting the writing	press the spacebar	0 (0)
how do you shake something?	move it up and down and side to side quickly.	stir it very quickly.	0 (0)
Clean tires	Pour water, cape off caked on dirt. Use speed wool to clean out crevices and sparrow spaces.	Pour water, scrape off caked on dirt. Use a steel wool to clean out crevices and narrow…	1 (1)
how do you taste something?	smell it enough to taste it.	place it in your mouth to taste.	1 (1)

Benchmark Tasks

- **LexGLUE:** a benchmark dataset for legal language understanding in English

Dataset	Source	Sub-domain	Task Type	Training/Dev/Test Instances	Classes
ECtHR (Task A)	Chalkidis et al. (2019a)	ECHR	Multi-label classification	9,000/1,000/1,000	10+1
ECtHR (Task B)	Chalkidis et al. (2021c)	ECHR	Multi-label classification	9,000/1,000/1,000	10+1
SCOTUS	Spaeth et al. (2020)	US Law	Multi-class classification	5,000/1,400/1,400	14
EUR-LEX	Chalkidis et al. (2021a)	EU Law	Multi-label classification	55,000/5,000/5,000	100
LEDGAR	Tuggener et al. (2020)	Contracts	Multi-class classification	60,000/10,000/10,000	100
UNFAIR-ToS	Lippi et al. (2019)	Contracts	Multi-label classification	5,532/2,275/1,607	8+1
CaseHOLD	Zheng et al. (2021)	US Law	Multiple choice QA	45,000/3,900/3,900	n/a

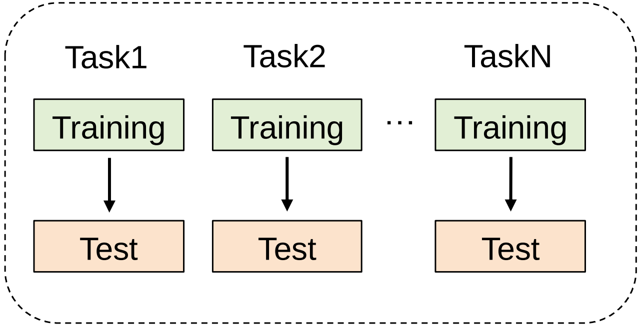
context (string)	endings (json)	label (class label)
Drapeau’s cohorts, the cohort would be a “victim” of making the bomb. Further, firebombs are inherently…	["holding that possession of a pipe bomb is a crime of violence for purposes of 18 usc 3142f1", "holding…	0 (0)
Colameta used customer information that he took from Protégé. Additionally, Colameta admits to having take…	["recognizing that even if a plaintiff claims certain information constitutes trade secrets its claim may…	1 (1)
property tax sale. In reviewing section 6323(b)(6), this Court noted that it provides that a county’s tax…	["holding that where there is a conflict between statutes the more recent statute is controlling and a…	4 (4)

Language Understanding Needs Diverse Skills

- ❑ **Different tasks may share common patterns (required skills)**
- ❑ **It is potential to build a unified foundation model and adapt it to different tasks**

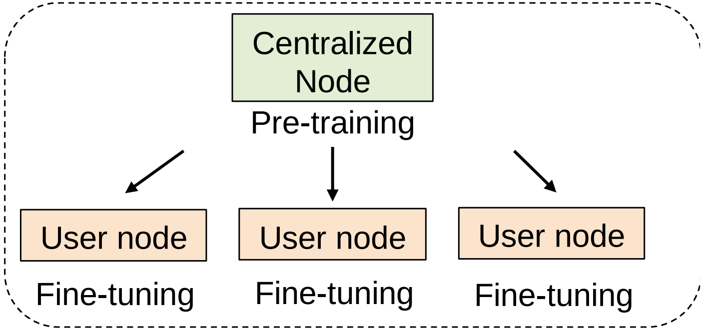
Skill	Description (Example)
Capable of	Whether an object is capable of performing an action (“A watch is capable of telling the past time”)
Long-tail knowledge	The question contains factual long-tail information (“Washington DC is located further south than Washington State”)
Plausibility	Quantifiers or always-never relations (“The peak of a mountain almost always reaches above the the tree line”)
Comparison	Comparison between two objects (“The end of a baseball bat is larger than the handle”)
Physical	Physical commonsense (“Do you build the walls on a house before putting on the roof?”)
Causality	Cause and effect relations (“If you get into an accident because you have been drinking alcohol you will be arrested?”)
Temporal	Temporal understanding (“None had ever reached the top of Mount Everest before 1977?”)
Negation	The question includes a negation phrase (“A mock trial is something with no legal consequence”)
Strategy	Reasoning steps are implicit and should be inferred using a strategy (“Blood banks almost never take cash or checks as deposits”)
Event chain	Question is about order of events (“Putting on shoes is done in this order normally: person ties shoelaces then slips shoes onto feet”)

From Individual Task Modeling to Centralized Training



Previous

Each user trains individual machine learning models for each task.



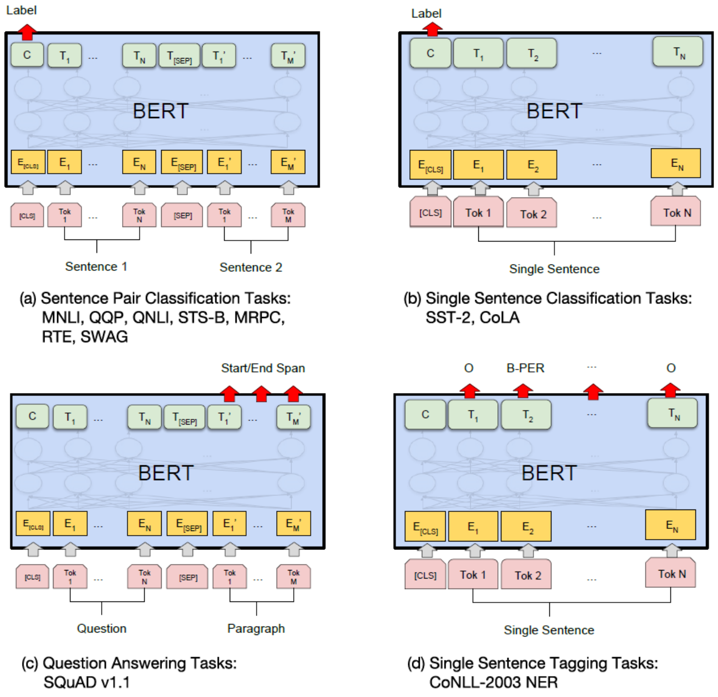
Now

The central node pre-trains the generalized language model and provides the model to users for task-specific fine-tuning.

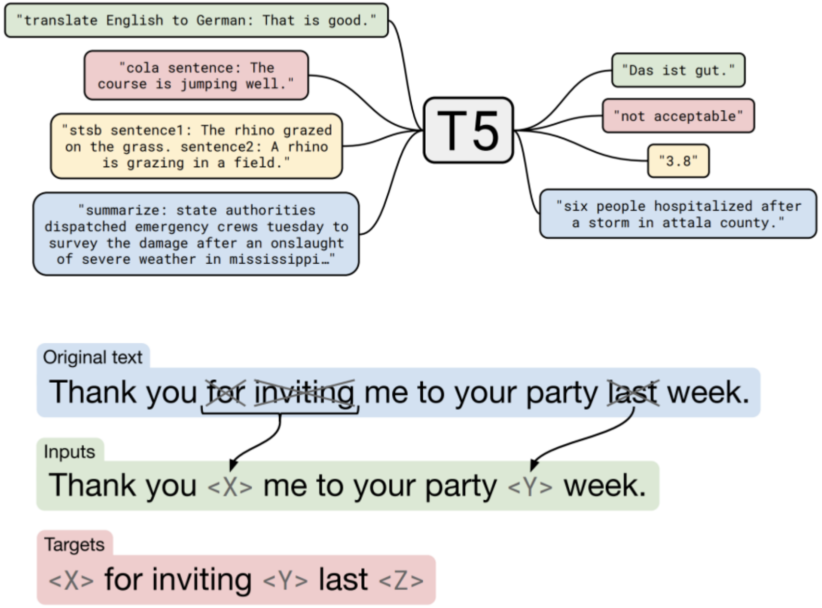
Individual training → Centralized pre-training + individual fine-tuning

**Extreme case : GPT3 gives predictions directly, eliminating the fine-tuning process*

Towards Multi-task Pre-training: Unified Modeling of Tasks



(a) Different formats of tasks



(b) Unified text-to-text format

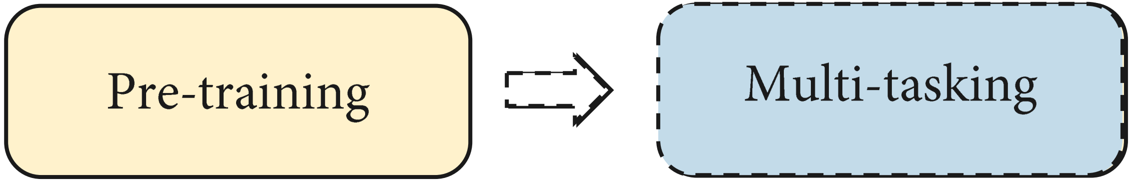
Large-scale Multi-task Pre-training

- ❑ **Theme:** Leveraging task-aware annotated data as supervised signals to assist with self-supervised learning on large-scale unlabeled data
- ❑ **Trend:** extreme scaling of task numbers, with little attention paid to the relationships between tasks
- ❑ **Challenges**
  - ➢ Catastrophic Forgetting
  - ➢ Negative Transfer

Challenge: Catastrophic Forgetting

Additional large-scale learning stage between **pre-training** and **fine-tuning**

Also known as **multi-task pre-fine-tuning** or **sequential training**



Challenge: Negative Transfer

Observation: tasks in different families may have side effects between each other.

	SUM	DLG	NLI	CLS	SEM	CMNS	CBQA	RC	Δ_{AVG}
SUM	27.89 29.36	37.81	60.45	77.10	78.25	61.92	7.84	65.37	-6.9%
DLG	29.05	38.56 39.76	63.62	77.10	75.55	64.05	13.39	64.75	+0.1%
NLI	28.61	40.60	64.91 67.23	77.29	77.72	67.60	15.24	66.40	+4.3%
CLS	29.52	40.16	66.69	77.14 77.47	76.05	65.29	12.93	65.20	+1.4%
SEM	29.30	38.86	62.46	76.83	72.09 72.79	57.84	12.44	63.52	-2.5%
CMNS	29.28	39.27	65.08	77.05	76.29	68.24 68.35	16.48	66.01	+4.7%
CBQA	29.75	39.29	64.96	77.66	75.21	66.84	14.68 19.98	66.37	+1.2%
RC	29.45	38.12	63.70	77.14	76.98	66.62	10.26	62.94 65.60	-2.4%
$AVG_{\backslash diag}$	29.28	39.16	63.77	77.17	76.43	64.31	12.65	65.37	

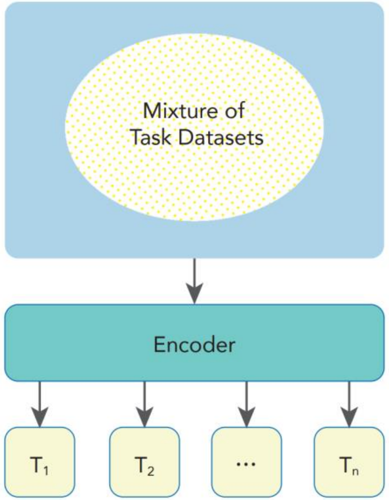
Summarization tasks generally seem to **hurt performance** on **dialogue system, natural language inference**, and **commonsense reasoning**

[6] Aribandi, Vamsi, et al. "ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning." International Conference on Learning Representations. 2021.

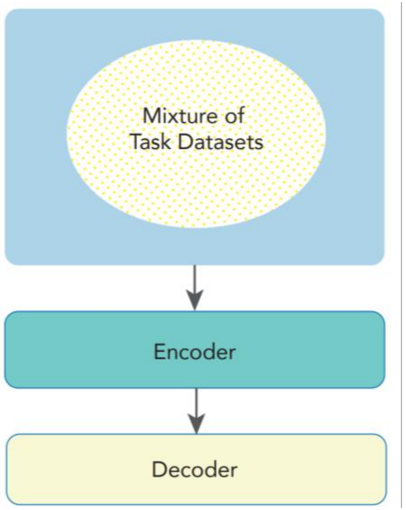
Previous Multi-task Language Models

a) Traditional methods: MT-DNN

b) Unified Text-to-text Methods: T5, ExT5, FLAN, T0, etc.



a) Traditional Methods

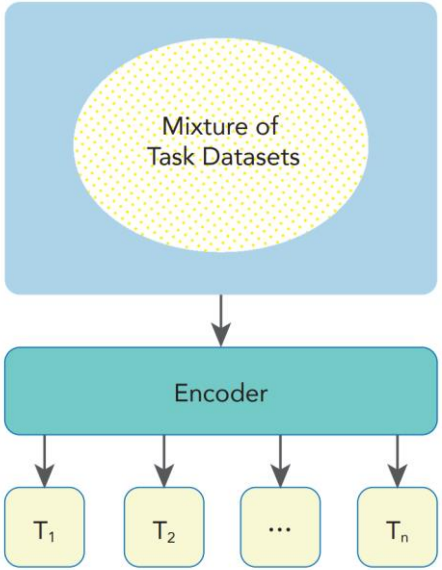


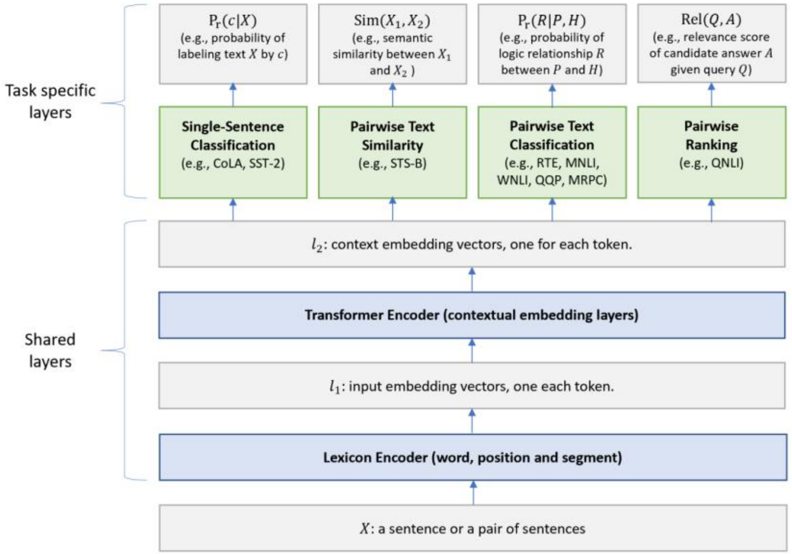
b) Unified Text-to-text Methods

[7] Liu, Xiaodong, et al. "Multi-Task Deep Neural Networks for Natural Language Understanding." ACL. 2019.
[8] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." J. Mach. Learn. Res. 21.140 (2020): 1-67.

Traditional Methods

- **Traditional methods:** MT-DNN
  - Require additional modifications to model architecture and increase model complexity and computation cost
  - Issue of catastrophic forgetting

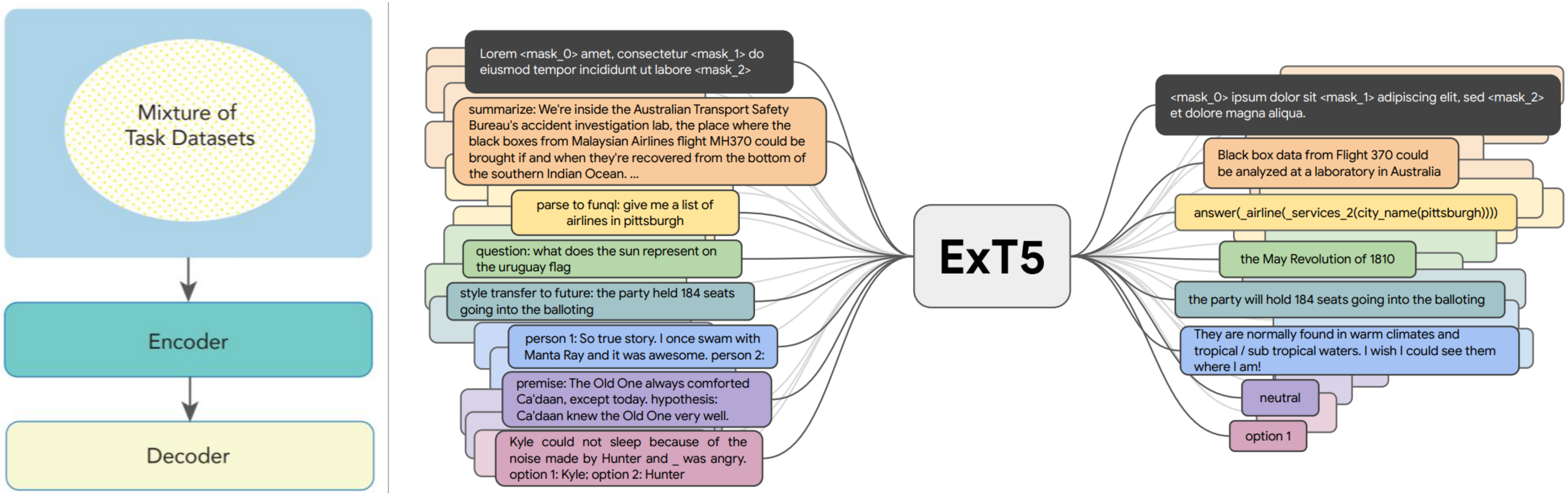




a) Traditional Methods

Unified Text-to-text Methods

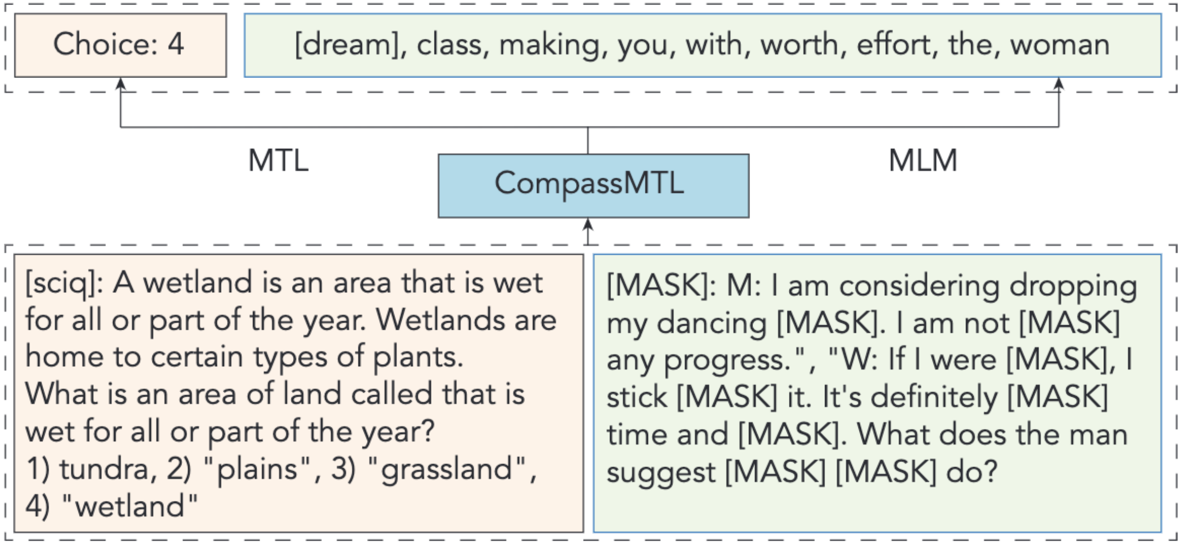
- **Unified Text-to-text Methods:** T5, ExT5, FLAN, T0, etc.
  - Negative transfer between tasks



b) Unified Text-to-text Methods

How to Capture Task Relationships: Our Solution

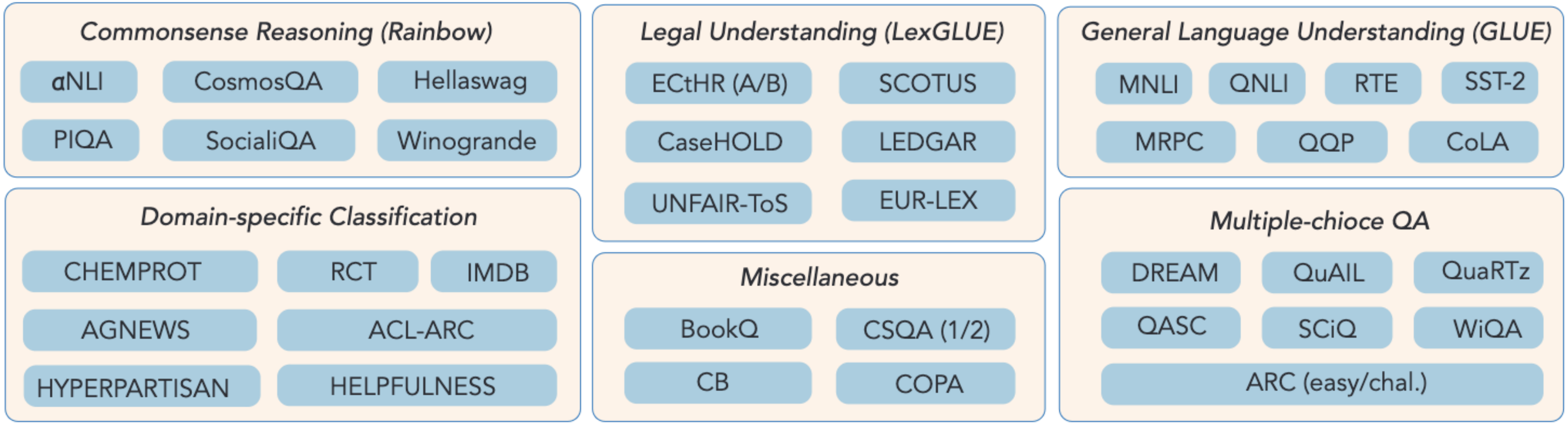
Ours: a task prefix guided multi-task pre-training framework



1) **Data:** Append a task prefix for each data sequence to capture common patterns from the task.
2) **Objective:** Require the model to predict some randomly masked prefixes to capture task differences.

Task Taxonomy

There are 40 datasets used for training our multi-task model, some of which are collected from GLUE SuperGLUE, Rainbow, and LexGLUE



Data Format (conversion)

Basic: Model tasks in a multiple-choice-like format to minimize the format transformation for NLU tasks

Conversion Criteria:

- Ensure that each training data has a specific number of *k* candidate options
- Original pair-wise input texts are regarded as context and question in the view of multiple-choice problem

If the number of candidate options > *k*	**the redundant options will be randomly discarded**
If the number of candidate options < *k*	add "N/A" placeholder options
If the ground-truth is a list	randomly select a correct option from the gold list and randomly sample *k-1* negative options from the held-out set
If ground-truth is a list and there is an empty choice	construct the truth option manually; the negative examples are constructed as the same as 3)

As a result, each training example will be formed as a sequence like
{ [Prefix]: context, question, option }

Data Format (Examples)

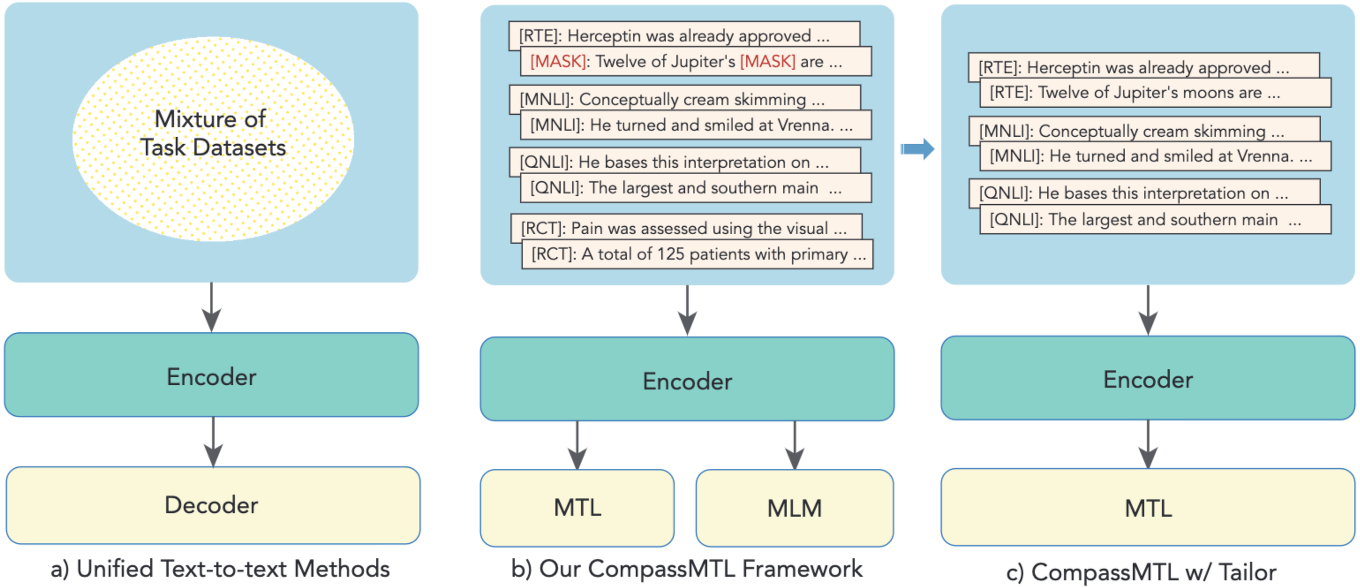
Context	Question	Option(s)
[sciq] A wetland is an area that is wet for all or part of the year. Wetlands are home to certain types of plants.	What is an area of land called that is wet for all or part of the year?	["tundra", "plains", "grassland", "wetland"]
[commonsense_qa] revolving door	A revolving door is convenient for two direction travel, but it also serves as a security measure at a what?	["bank", "library", "department store", "mall", ~~"new york"~~]
[dream] M: I am considering dropping my dancing class. I am not making any progress.", "W: If I were you, I stick with it. It's definitely worth time and effort.	What does the man suggest the woman do?	["Consult her dancing teacher.", "Take a more interesting class.", "Continue her dancing class.", "N/A"]
[scotus] The Interstate Commerce Commission, acting under § 19a of the Interstate Commerce Act, ordered the appellant to furnish certain inventories, schedules, maps and charts of its pipe line property ...	-	["Unions", "Economic Activity", "Judicial Power", "Federalism"]
[unfair_tos] you must provide accurate and complete data during the registration and update your registration data if it changes .	-	["there is no unfair contractual term", "Limitation of liability", "Unilateral termination", "Arbitration"]

Model Architecture

Backbone: Encoder-only, based on the DeBERTa architecture

Training Objectives: Multi-task Learning (MTL) + Masked Language Modeling (MLM)

Usages: Unified Foundation Model + Probing Tool



Model Architecture

Data-centric: without modification of model architecture. It can be regarded as an efficient implementation of the traditional MTL method composed of a **shared representation** module and **task-aware modules**.

the prefix is supposed to reflect the common patterns from the dataset

the model is required to predict randomly masked prefixes to capture task differences.

[sciq]: A wetland is an area that is wet for all or part of the year. Wetlands are home to certain types of plants. What is an area of land called that is wet for all or part of the year? 1) tundra, 2) "plains", 3) "grassland", 4) "wetland"

[MASK]: M: I am considering dropping my dancing [MASK]. I am not [MASK] any progress.", "W: If I were [MASK], I stick [MASK] it. It's definitely [MASK] time and [MASK]. What does the man suggest [MASK] [MASK] do?

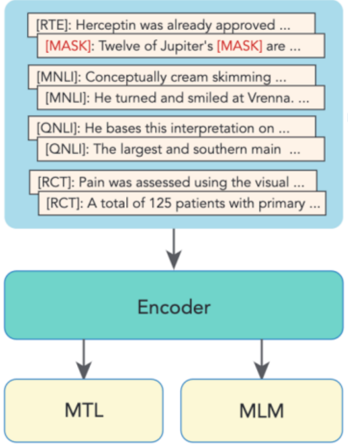
Model Architecture

Model Evolution

CompassMTL → Relationship Probing → CompassMTL w/ Tailor

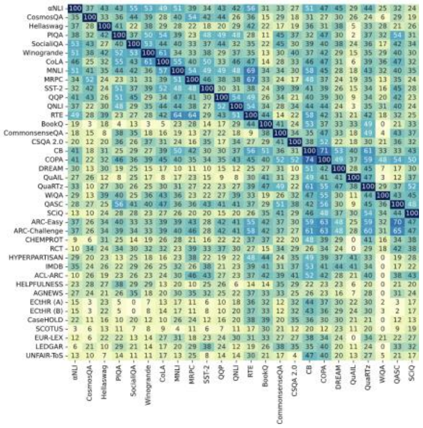
CompassMTL

Trained on the 40 tasks



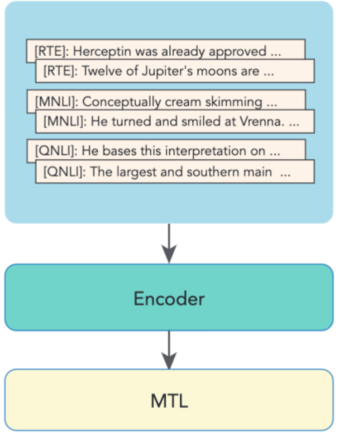
Relationship Probing

1) Only uses the MLM
2) Input data without options



CompassMTL w/ Tailor

Trained with selected tasks



Main Results

1. CompassMTL models outperform the related public models in general
2. Our **encoder-only models** yield better performance than the T5-based encoder-decoder models.
3. It is potential to achieve better performance by multi-task learning with **related tasks** (w/ Tailor)

Model	Arch.	Tasks	Params.	αNLI	CosmosQA	HellaSwag	PIQA	SocialIQA	Winogrande	Average
UNICORN	Enc-Dec	6	770M	79.5	83.2	83.0	82.2	75.5	78.7	80.4
ExT5	Enc-Dec	107	770M	82.3	85.9	89.0	85.0	79.7	82.5	84.1
ExDeBERTa	Enc only	40	567M	87.9	85.3	83.6	85.5	79.6	87.0	84.8
CompassMTL	Enc only	40	567M	91.7	87.8	95.6	87.3	81.7	89.6	89.0
w/ Tailor	Enc only	14	567M	**92.5**	**88.8**	**96.1**	**88.3**	**82.2**	**90.5**	**89.7**

Method	ECtHR (A)		ECtHR (B)		SCOTUS		EUR-LEX		LEDGAR		UNFAIR-ToS		CaseHOLD
	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ/m-F_1
BERT	71.2	63.6	79.7	73.4	68.3	58.3	71.4	57.2	87.6	81.8	95.6	81.3	70.8
RoBERTa	69.2	59.0	77.3	68.9	71.6	62.0	71.9	**57.9**	87.9	82.3	95.2	79.2	71.4
DeBERTa	70.0	60.8	78.8	71.0	71.1	62.7	**72.1**	57.4	88.2	83.1	95.5	80.3	72.6
Longformer	69.9	**64.7**	79.4	71.7	72.9	64.0	71.6	57.7	88.2	83.0	95.5	80.9	71.9
BigBird	70.0	62.9	78.8	70.9	72.8	62.0	71.5	56.8	87.8	82.6	95.7	81.3	70.8
Legal-BERT	70.0	64.0	80.4	**74.7**	76.4	66.5	**72.1**	57.4	88.2	83.0	96.0	83.0	75.3
CaseLaw-BERT	69.8	62.9	78.8	70.3	76.6	65.9	70.7	56.6	**88.3**	83.0	96.0	82.3	75.4
ExDeBERTa	-	-	-	-	-	-	-	-	-	-	-	-	74.8
CompassMTL	71.7	60.7	80.6	73.2	**77.7**	**68.9**	67.2	42.1	88.1	82.3	**96.3**	**84.3**	76.1
w/ Tailor	**73.0**	**64.7**	**80.7**	72.3	76.3	68.6	66.9	44.9	**88.3**	**83.2**	96.2	83.2	**78.1**

Main Results

1. CompassMTL models outperform the related public models in general
2. Our **encoder-only models** yield better performance than the T5-based encoder-decoder models.
3. It is potential to achieve better performance by multi-task learning with **related tasks** (w/ Tailor)

Model	Arch.	Tasks	Params.	αNLI	CosmosQA	HellaSwag	PIQA	SocialIQA	Winogrande	Average
UNICORN	Enc-Dec	6	770M	79.5	83.2	83.0	82.2	75.5	78.7	80.4
ExT5	Enc-Dec	107	770M	82.3	85.9	89.0	85.0	79.7	82.5	84.1
ExDeBERTa	Enc only	40	567M	87.9	85.3	83.6	85.5	79.6	87.0	84.8
CompassMTL	Enc only	40	567M	91.7	87.8	95.6	87.3	81.7	89.6	89.0
w/ Tailor	Enc only	14	567M	**92.5**	**88.8**	**96.1**	**88.3**	**82.2**	**90.5**	**89.7**

Method	ECtHR (A)		ECtHR (B)		SCOTUS		EUR-LEX		LEDGAR		UNFAIR-ToS		CaseHOLD
	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ/m-F_1
BERT	71.2	63.6	79.7	73.4	68.3	58.3	71.4	57.2	87.6	81.8	95.6	81.3	70.8
RoBERTa	69.2	59.0	77.3	68.9	71.6	62.0	71.9	**57.9**	87.9	82.3	95.2	79.2	71.4
DeBERTa	70.0	60.8	78.8	71.0	71.1	62.7	**72.1**	57.4	88.2	83.1	95.5	80.3	72.6
Longformer	69.9	**64.7**	79.4	71.7	72.9	64.0	71.6	57.7	88.2	83.0	95.5	80.9	71.9
BigBird	70.0	62.9	78.8	70.9	72.8	62.0	71.5	56.8	87.8	82.6	95.7	81.3	70.8
Legal-BERT	70.0	64.0	80.4	**74.7**	76.4	66.5	**72.1**	57.4	88.2	83.0	96.0	83.0	75.3
CaseLaw-BERT	69.8	62.9	78.8	70.3	76.6	65.9	70.7	56.6	**88.3**	83.0	96.0	82.3	75.4
ExDeBERTa	-	-	-	-	-	-	-	-	-	-	-	-	74.8
CompassMTL	71.7	60.7	80.6	73.2	**77.7**	**68.9**	67.2	42.1	88.1	82.3	**96.3**	**84.3**	76.1
w/ Tailor	**73.0**	**64.7**	**80.7**	72.3	76.3	68.6	66.9	44.9	**88.3**	**83.2**	96.2	83.2	**78.1**

Main Results

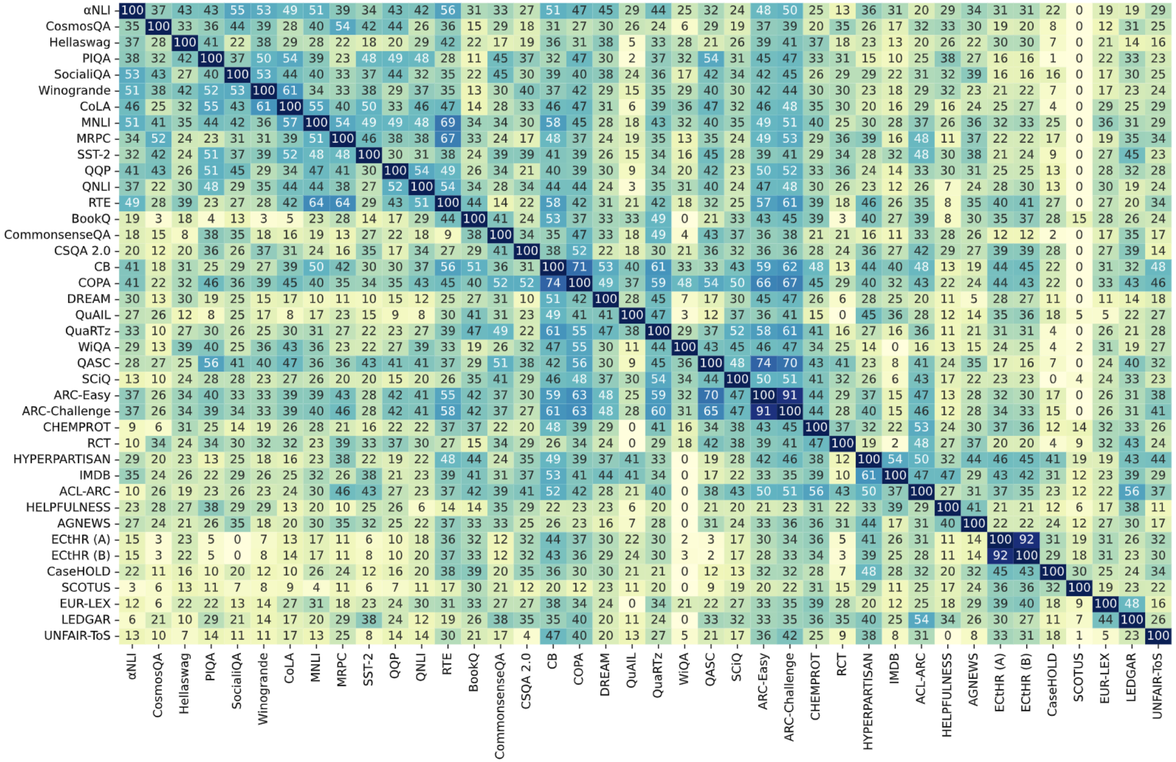
1. CompassMTL models outperform the related public models in general
2. Our **encoder-only models** yield better performance than the T5-based encoder-decoder models.
3. It is potential to achieve better performance by multi-task learning with **related tasks** (w/ Tailor)

Model	**Arch.**	**Tasks**	**Params.**	**αNLI**	**CosmosQA**	**HellaSwag**	**PIQA**	**SocialIQA**	**Winogrande**	**Average**
UNICORN	Enc-Dec	6	770M	79.5	83.2	83.0	82.2	75.5	78.7	80.4
ExT5	Enc-Dec	107	770M	82.3	85.9	89.0	85.0	79.7	82.5	84.1
ExDeBERTa	Enc only	40	567M	87.9	85.3	83.6	85.5	79.6	87.0	84.8
CompassMTL	Enc only	40	567M	91.7	87.8	95.6	87.3	81.7	89.6	89.0
w/ Tailor	Enc only	14	567M	**92.5**	**88.8**	**96.1**	**88.3**	**82.2**	**90.5**	**89.7**

Method	**ECtHR (A)**		**ECtHR (B)**		**SCOTUS**		**EUR-LEX**		**LEDGAR**		**UNFAIR-ToS**		**CaseHOLD**
	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ-F_1	m-F_1	μ/m-F_1
BERT	71.2	63.6	79.7	73.4	68.3	58.3	71.4	57.2	87.6	81.8	95.6	81.3	70.8
RoBERTa	69.2	59.0	77.3	68.9	71.6	62.0	71.9	**57.9**	87.9	82.3	95.2	79.2	71.4
DeBERTa	70.0	60.8	78.8	71.0	71.1	62.7	**72.1**	57.4	88.2	83.1	95.5	80.3	72.6
Longformer	69.9	**64.7**	79.4	71.7	72.9	64.0	71.6	57.7	88.2	83.0	95.5	80.9	71.9
BigBird	70.0	62.9	78.8	70.9	72.8	62.0	71.5	56.8	87.8	82.6	95.7	81.3	70.8
Legal-BERT	70.0	64.0	80.4	**74.7**	76.4	66.5	**72.1**	57.4	88.2	83.0	96.0	83.0	75.3
CaseLaw-BERT	69.8	62.9	78.8	70.3	76.6	65.9	70.7	56.6	**88.3**	83.0	96.0	82.3	75.4
ExDeBERTa	-	-	-	-	-	-	-	-	-	-	-	-	74.8
CompassMTL	71.7	60.7	80.6	73.2	**77.7**	**68.9**	67.2	42.1	88.1	82.3	**96.3**	**84.3**	76.1
w/ Tailor	**73.0**	**64.7**	**80.7**	72.3	76.3	68.6	66.9	44.9	**88.3**	**83.2**	96.2	83.2	**78.1**

Relationship Probing

Probing Model: only uses the MLM objective and is fed without options to alleviate possible shortcuts.

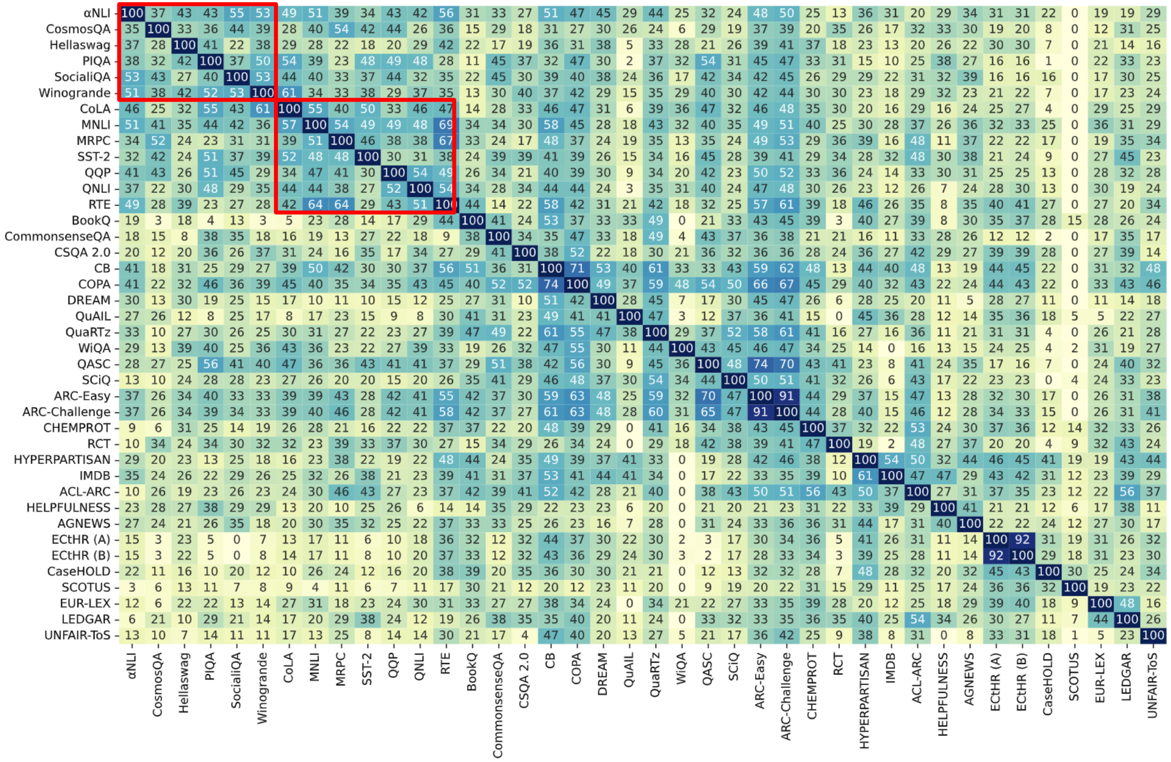


HowTo:

1) Fetch prefix embeddings
2) Calculate the Pearson correlation between each task pair

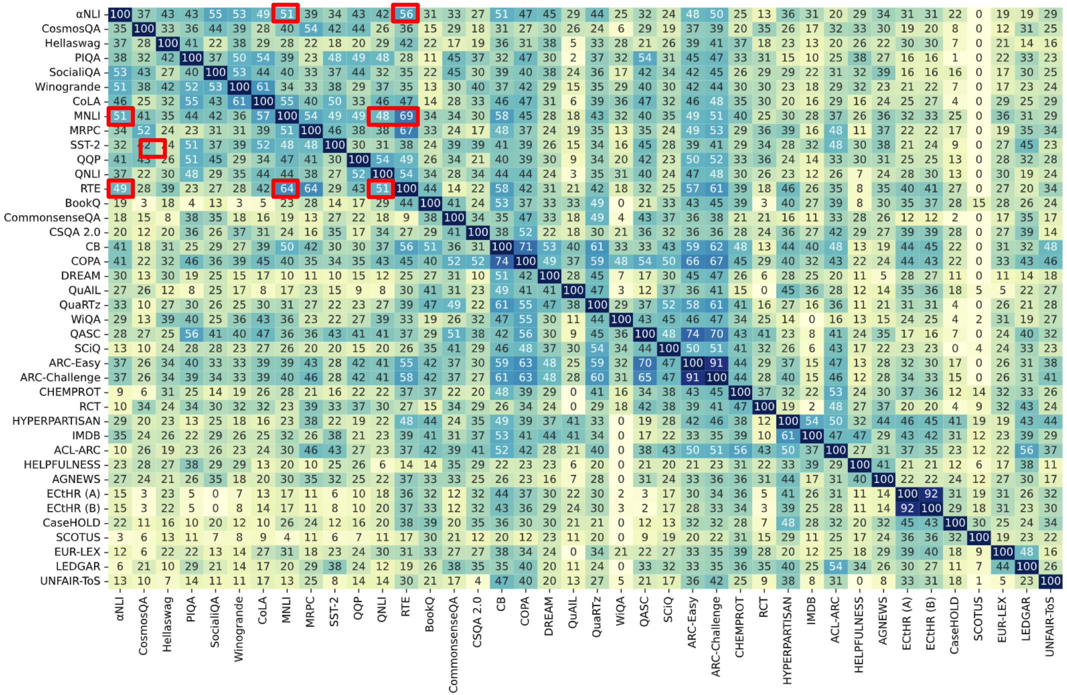
Relationship Probing

1. The datasets inside the same task family (e.g., GLUE and Rainbow) correlate highly with each other.
2. The correlation scores also accord with the common practice of data augmentation.



Relationship Probing

1. The datasets inside the same task family (e.g., GLUE and Rainbow) correlate highly with each other.
2. The correlation scores also accord with the common practice of data augmentation.



1) the NLI datasets (MNLI, QNLI, RTE) share close relevance

2) helpful to initialize from an MNLI model to fine-tune RTE

Relationship Probing

Topic: Whether the relationship scores coordinate with the model performance transferred between tasks

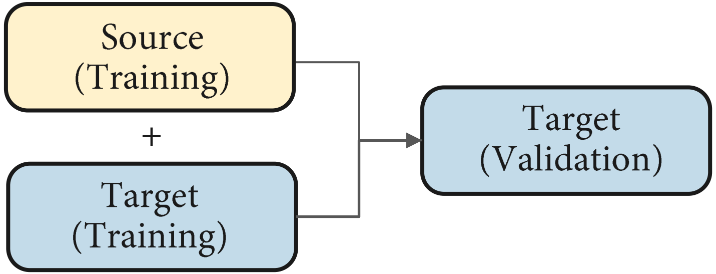
Source tasks: 13 source tasks from GLUE and Rainbow tasks

Target Tasks: 5 target tasks (ANLI, HellaSwag, MRPC, PIQA, QNLI, and RTE)

Dual-task training setup:

Co-training: train individual models using the mixture of training sets from each pair of source & target tasks

Evaluation: then evaluate the model on the validation set of the target dataset.



Relationship Probing

Finally, we have 5 X 13 transfer results.

For each target dataset, we calculate Pearson correlation between relationship scores and transfer accuracy among the source datasets.

Dataset	RTE	MRPC	QNLI	HellaSwag	αNLI
Correlation	0.19	0.22	0.38	0.12	0.51

Table 3: Pearson correlation between the relationship scores and the transfer accuracy.

Result: the relationship scores are positively bound up with the transfer performance

Complementary Transfer

Topic:

1. whether using more datasets always leads to better performance
2. whether using the most related datasets can lead to competitive results.

Data Selection: select a group of datasets to train an MTL model and fine-tuning the model on target datasets.

40-fullset	**the same as our basic setting of CompassMTL**
Top-5	Top-5 ranked dataset according to based on our probed relationship scores
Family	the datasets belonged to the same family with the target dataset
14-subset	the mixture of Rainbow and GLUE datasets

Complementary Transfer

1. **Top-5** variant yields comparable, even better results than the others
2. **Small-scale datasets** (e.g., MRPC and RTE) are more likely to benefit from the complementary transfer

Model	**Tasks**	**RTE**	**MRPC**	**QNLI**	**HellaSwag**	**αNLI**
Single	1	61.4	89.2	95.0	95.1	91.3
40-fullset	40	**92.8**	90.4	95.5	95.6	91.7
Top 5	5	92.4	**91.9**	95.3	95.6	91.6
Family	6/7	91.4	90.2	95.0	95.7	91.9
14-subset	14	91.8	90.3	**95.6**	**96.1**	**92.5**

Complementary Transfer

1. **Top-5** variant yields comparable, even better results than the others
2. **Small-scale datasets** (e.g., MRPC and RTE) are more likely to benefit from the complementary transfer

Model	**Tasks**	**RTE**	**MRPC**	**QNLI**	**HellaSwag**	**αNLI**
Single	1	61.4	89.2	95.0	95.1	91.3
40-fullset	40	**92.8**	90.4	95.5	95.6	91.7
Top 5	5	92.4	**91.9**	95.3	95.6	91.6
Family	6/7	91.4	90.2	95.0	95.7	91.9
14-subset	14	91.8	90.3	**95.6**	**96.1**	**92.5**

Human-parity on Commonsense Reasoning Leaderboards

Models: The submissions are based on the ensemble of three models from complementary transfer.

Results: Compared with public methods that use much larger PrLMs, model ensemble, and knowledge graphs, our models establish new state-of-the-art results and reach **human-parity performance**.

Model	HellaSwag	αNLI
Human Performance	95.60	92.90
Previous SOTA	94.87	92.20
Our Results	95.94	92.80

Human Performance — Accuracy: 0.9560

Download

Rank	Submission	Created	Accuracy
1	UniMTL *Microsoft Azure Cognitive Ser...*	05/11/2022	0.9594
2	DeBERTa Large *EMNLP Paper 3842 Authors*	05/20/2022	0.9557
3	CreAT *Anonymous*	05/03/2022	0.9487
4	DeBERTa MCQ *EMNLP Paper 3842 Authors*	06/03/2022	0.9472
5	DeBERTa Large *Anonymous*	04/14/2022	0.9437
6	UL Test *Google Research*	03/14/2022	0.9413
7	Deberta Testing *Myself*	05/06/2022	0.9394
8	UNICORN *Anonymous*	07/24/2020	0.9385

https://leaderboard.allenai.org/hellaswag/submissions/public
https://leaderboard.allenai.org/anli/submissions/public

Beyond The Unified Format

Topic: whether our model can be used for tasks that are unavailable to be transformed into our format

We evaluate the effectiveness by using the 1) reading comprehension datasets **SQuAD v1.1/2.0** and named entity recognition (NER) dataset **CoNLL 2003**.

Results show that our model is **generally effective across formats**

Model	SQuADv1.1		SQuADv2.0		NER
	EM	F1	EM	F1	F1
Baseline	88.8	94.8	87.1	90.5	96.5
CompassMTL	89.7	95.1	88.5	91.3	96.9

Extension to T5

Our method is **generally applicable to other kinds of PrLMs**, such as encoder-decoder T5.

Model	αNLI	CosmosQA	HellaSwag	PIQA	SocialIQA	Winogrande	Average
T5	68.5	69.6	56.6	67.7	65.1	62.4	65.0
UNICORN	65.3	72.8	56.2	73.3	66.1	61.8	65.9
CompassMTL	**69.1**	**72.6**	**57.7**	**73.6**	**66.6**	**64.9**	**67.4**

Table 9: Results on the Rainbow validation sets by using T5-base as the backbone model.

Conclusions

- **A unified task prefix guided multi-task method**
  - Strong foundation backbone for a wide range of NLU tasks
  - A probing tool for analyzing task relationships
- **Effectiveness**
  - Generalizable advances over tasks in diverse formats
  - Establishes human-parity results on commonsense reasoning tasks
- **Findings**
  - prefixes reflect task relationships, which correlate with transfer learning performance between tasks
  - suggest directions for data augmentation of complementary tasks

Prospects for Future Studies

1) Collaborative multi-task learning of PrLMs

The recipe of using **task prefixes + prefix prediction in MLM** has shown effective for MTL pre-training.

2) **Suggestive choice for data augmentation**

The probed task relationships have shown informative in **finding complementary tasks**, which help obtain better performance for a target task, especially for small-scale datasets.

3) **Guidance for skill-aware model evaluation**

The discovery of task relationships may help determine redundant datasets that assess similar patterns of models to **avoid evaluation redundancy and save computation**.

Thanks & QA

Zhuosheng Zhang
zhangzs@sjtu.edu.cn
https://bcmi.sjtu.edu.cn/~zhangzs