An Introduction to MT-DNN:

Multi-Task Deep Neural Networks for Natural Language Understanding

GUO YAWEN 04/14/2019

Backgrounds

· ELMo、OpenGPT、BERT、MT-DNN

· Learning vector-space representations of text, e.g., words and sentences, is fundamental to many natural language understanding (NLU) tasks.

· MTL provides an effective way of leveraging supervised data from many related tasks.

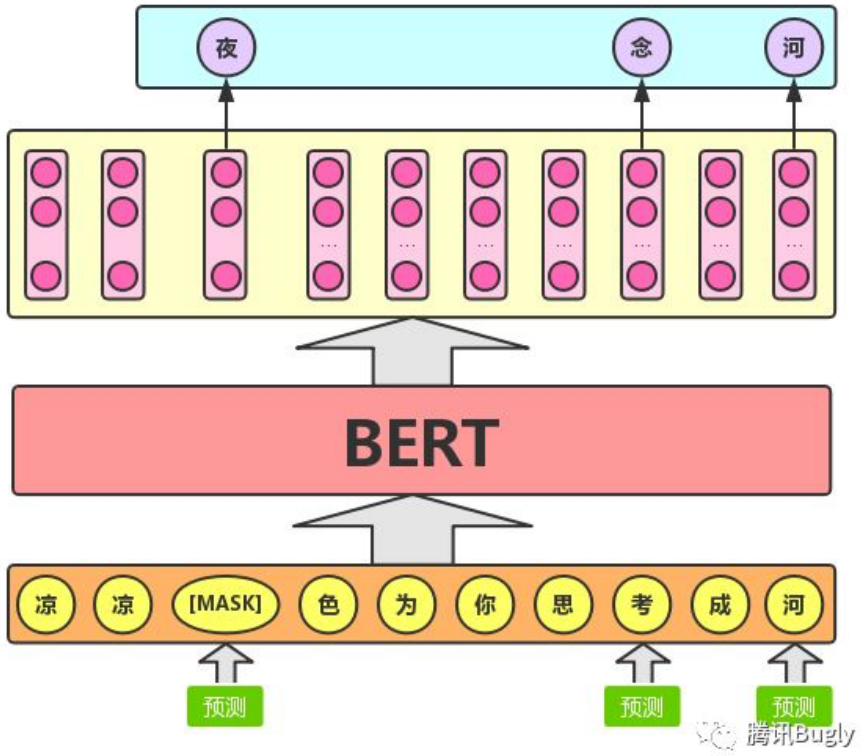
· The use of multi-task learning profits from a regularization effect via alleviating overfitting to a specific task, thus making the learned representations universal across tasks.

Why *MT- DNN*?

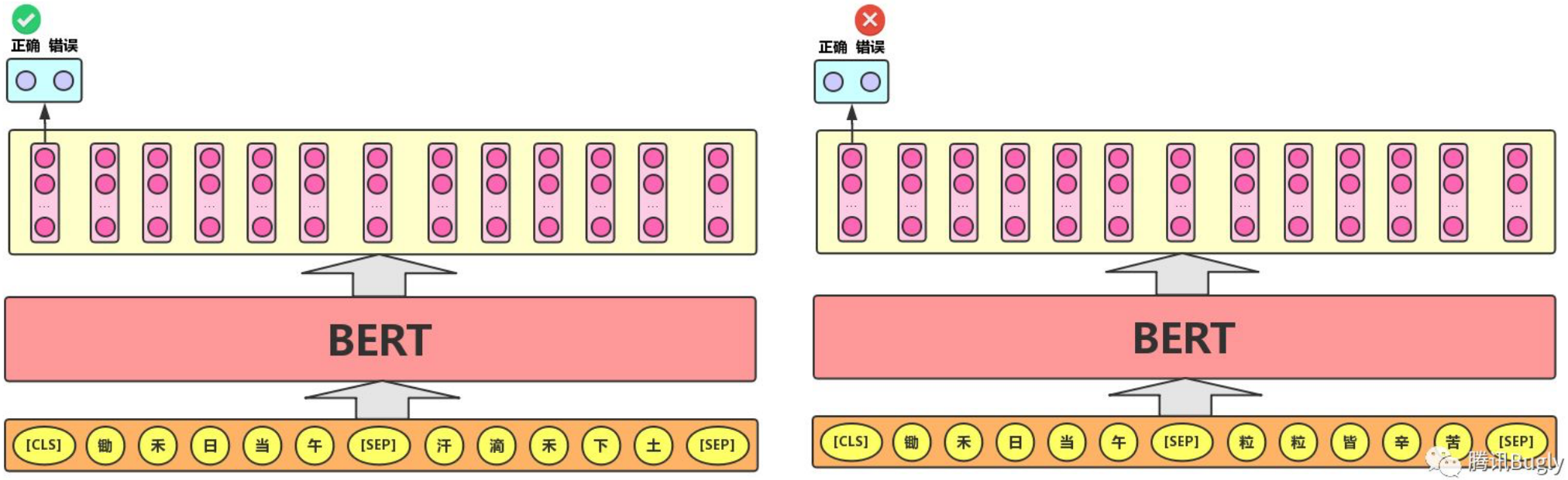
· MT- DNN to combine *multi-task learning* and *language model pre-training* for language representation learning.

· MT-DNN obtains new state-of- the-art results on ten NLU tasks across three popular benchmarks: SNLI, SciTail, and GLUE. MT- DNN also demonstrates an exceptional generalization capability in domain adaptation experiments.

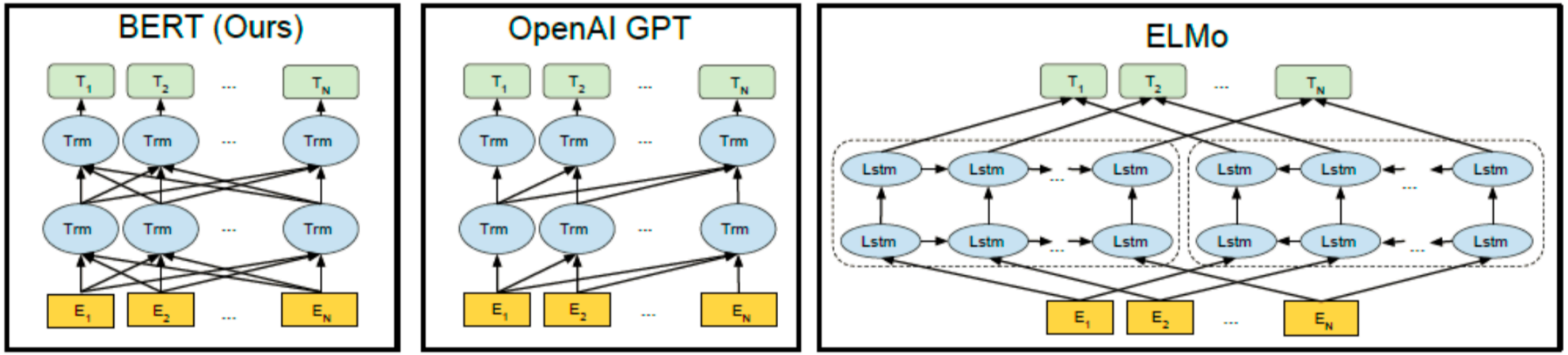
BERT – Masked LM



BERT – NextSentence Prediction

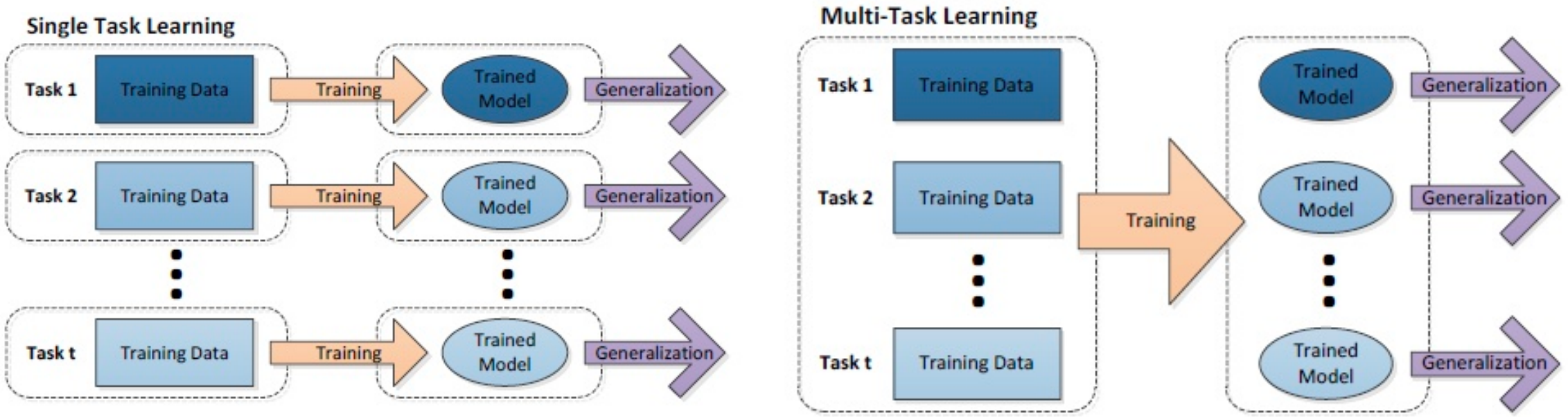


BERT-Bidirectional Encoder Representations from Transformers

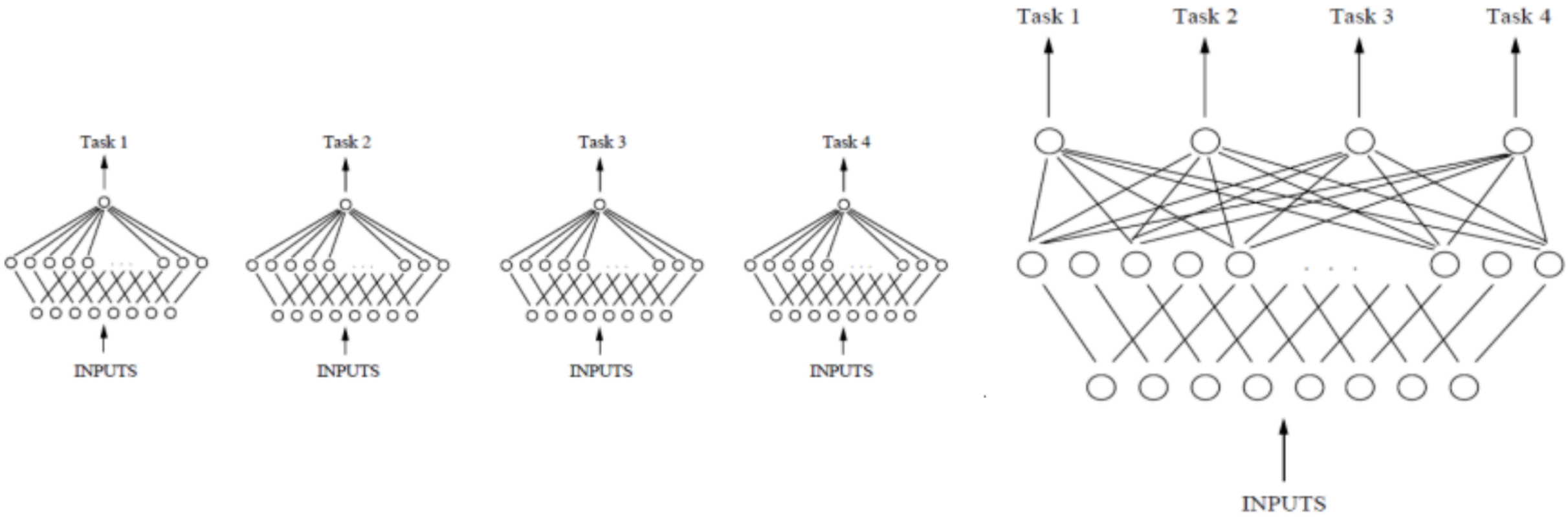


Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

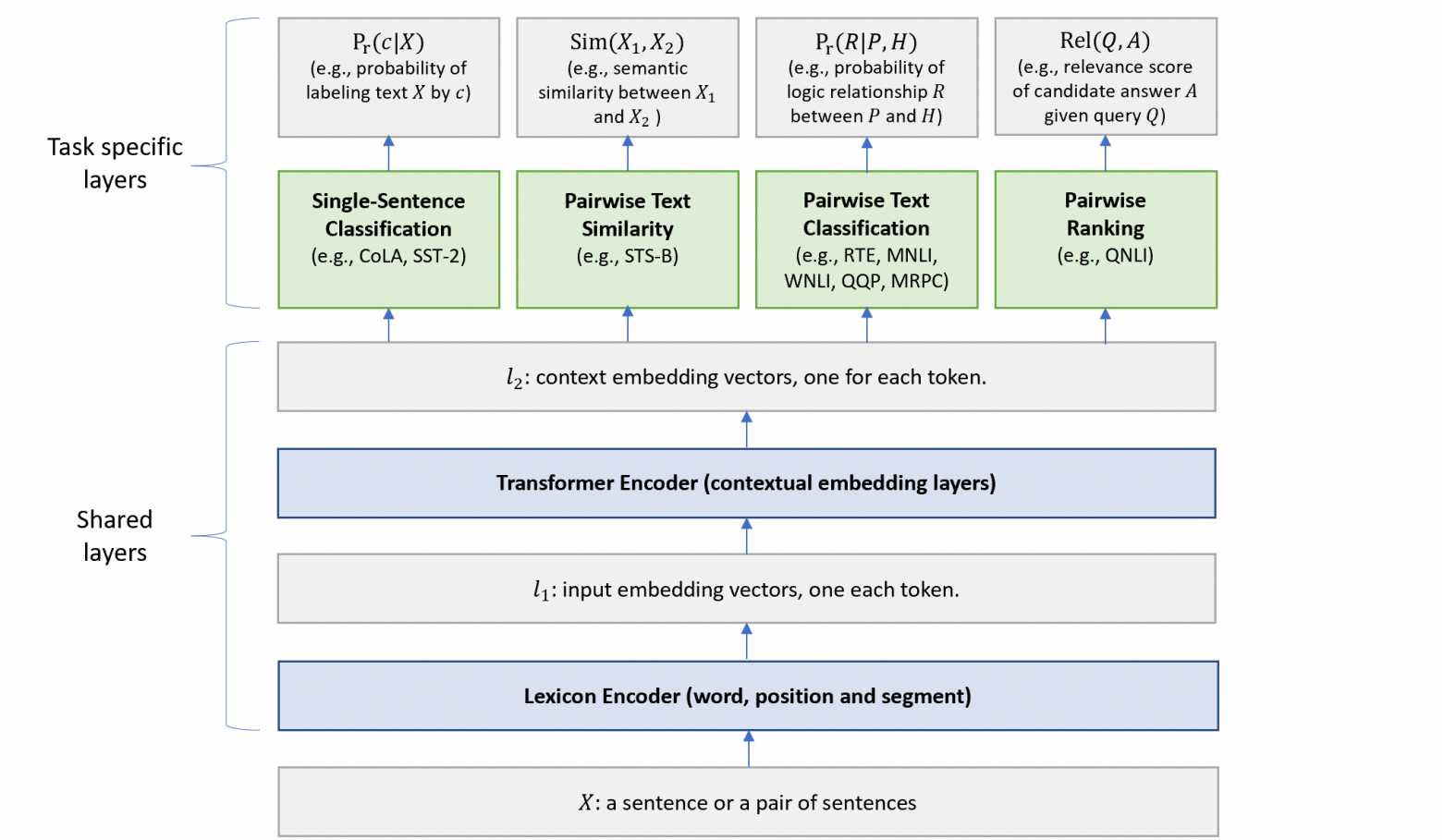
Multi- task Learning



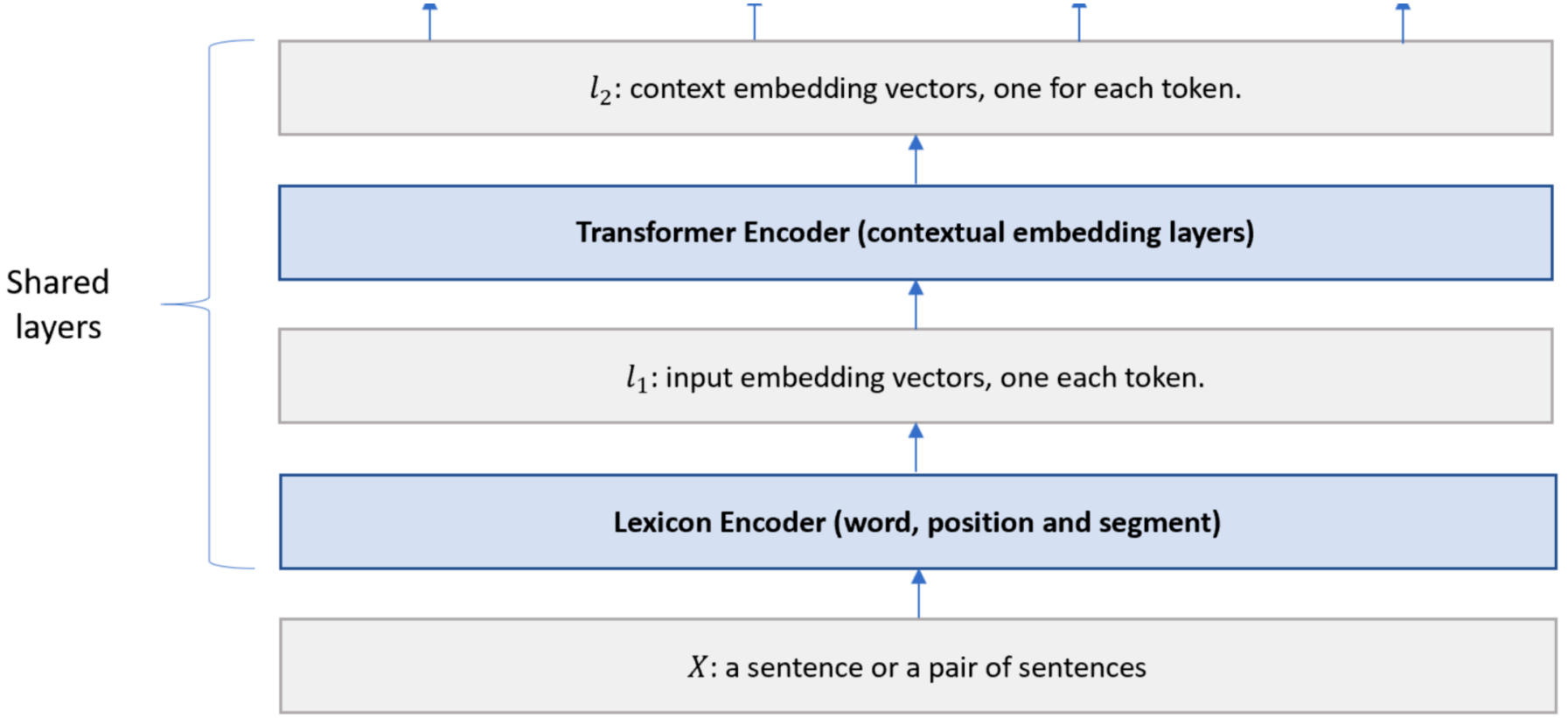
Multi-task Learning（shared representation）



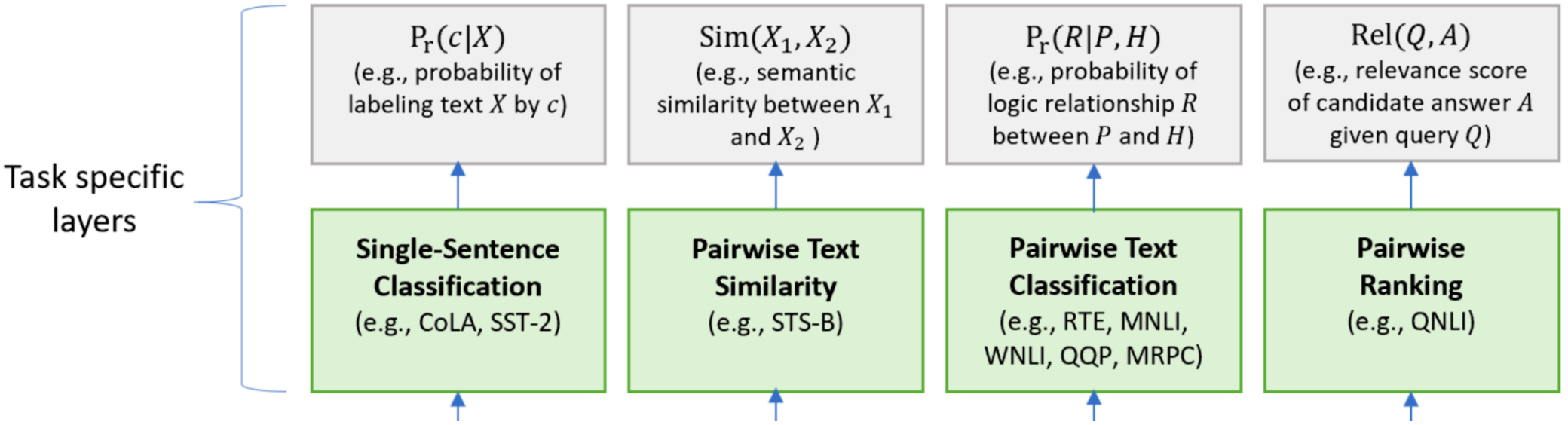
MT-DNN Model


Task specific layers
$\mathrm{P_r}(c|X)$
(e.g., probability of labeling text X by c)
$\mathrm{Sim}(X_1, X_2)$
(e.g., semantic similarity between X_1 and X_2)
$\mathrm{P_r}(R|P, H)$
(e.g., probability of logic relationship R between P and H)
$\mathrm{Rel}(Q, A)$
(e.g., relevance score of candidate answer A given query Q)
Single-Sentence Classification
(e.g., CoLA, SST-2)
Pairwise Text Similarity
(e.g., STS-B)
Pairwise Text Classification
(e.g., RTE, MNLI, WNLI, QQP, MRPC)
Pairwise Ranking
(e.g., QNLI)
Shared layers
l_2: context embedding vectors, one for each token.
Transformer Encoder (contextual embedding layers)
l_1: input embedding vectors, one each token.
Lexicon Encoder (word, position and segment)
X: a sentence or a pair of sentences

MT-DNN Model（Lower layers）



MT- DNN Model (Top layers)



Datasets

MT-DNN obtains new state-of- the-art results on ten NLU tasks across three popular benchmarks: SNLI, SciTail, and GLUE.

Corpus	Task	#Train	#Dev	#Test	#Label	Metrics
Single-Sentence Classification (GLUE)						
CoLA	Acceptability	8.5k	1k	1k	2	Matthews corr
SST-2	Sentiment	67k	872	1.8k	2	Accuracy
Pairwise Text Classification (GLUE)						
MNLI	NLI	393k	20k	20k	3	Accuracy
RTE	NLI	2.5k	276	3k	2	Accuracy
WNLI	NLI	634	71	146	2	Accuracy
QQP	Paraphrase	364k	40k	391k	2	Accuracy/F1
MRPC	Paraphrase	3.7k	408	1.7k	2	Accuracy/F1
Text Similarity (GLUE)						
STS-B	Similarity	7k	1.5k	1.4k	1	Pearson/Spearman corr
Relevance Ranking (GLUE)						
QNLI	QA/NLI	108k	5.7k	5.7k	2	Accuracy
Pairwise Text Classification						
SNLI	NLI	549k	9.8k	9.8k	3	Accuracy
SciTail	NLI	23.5k	1.3k	2.1k	2	Accuracy

Table 1: Summary of the three benchmarks: GLUE, SNLI and SciTail.

Tasks

Single-Sentence Classification

CoLA :to predict whether an English sentence is grammatically plausible.

SST-2: whether the sentiment of a sentence extracted from movie reviews is positive or negative.

Text Similarity

The model predicts a real-value score indicating the semantic similarity of the two sentences

Tasks

Pairwise Text Classification

Given a pair of sentences, the model determines the relationship of the two sentences based on a set of pre-defined labels.

Relevance Ranking

Given a query and a list of candidate answers, the model ranks all the candidates in the order of relevance to the query

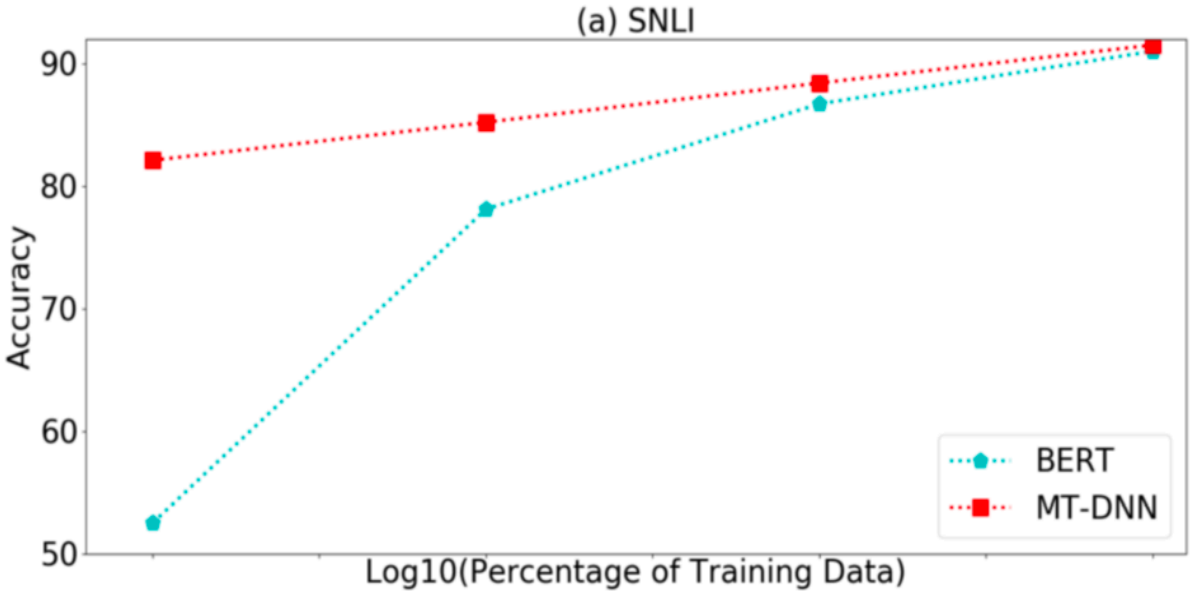
Results - GLUE

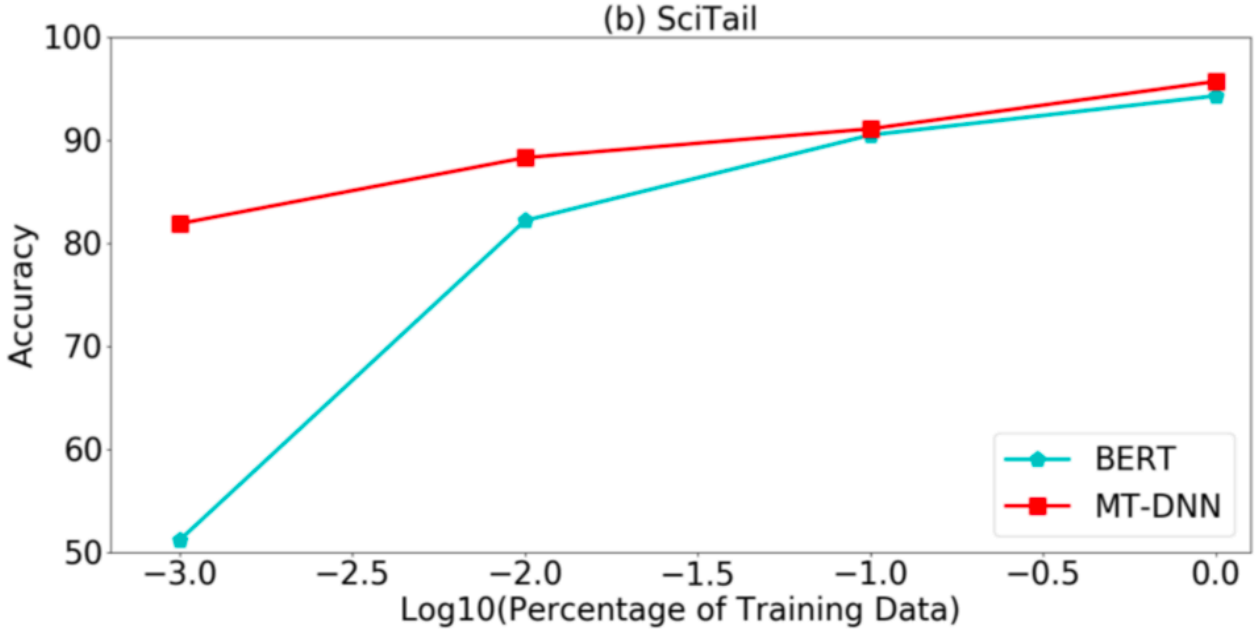
Model	CoLA 8.5k	SST-2 67k	MRPC 3.7k	STS-B 7k	QQP 364k	MNLI-m/mm 393k	QNLI 108k	RTE 2.5k	WNLI 634	AX	**Score**
BiLSTM+ELMo+Attn	36.0	90.4	84.9/77.9	75.1/73.3	64.8/84.7	76.4/76.1	79.9	56.8	65.1	26.5	70.5
Singletask Pretrain Transformer	45.4	91.3	82.3/75.7	82.0/80.0	70.3/88.5	82.1/81.4	88.1	56.0	53.4	29.8	72.8
GPT on STILTs	47.2	93.1	87.7/83.7	85.3/84.8	70.1/88.1	80.8/80.6	87.2	69.1	65.1	29.4	76.9
$\text{BERT}_{\text{LARGE}}$	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	91.1	70.1	65.1	39.6	80.4
MT-DNN	**61.5**	**95.6**	**90.0/86.7**	**88.3/87.7**	**72.4/89.6**	**86.7/86.0**	**98.0**	**75.5**	**65.1**	40.3	**82.2**

Result – SNLI & SciTail

Model	MNLI-m/mm	QQP	MRPC	RTE	QNLI	SST-2	CoLA	STS-B
$\text{BERT}_{\text{BASE}}$	84.5/84.4	90.4/87.4	84.5/89.0	65.0	88.4	92.8	55.4	89.6/89.2
ST-DNN	84.7/84.6	91.0/87.9	86.6/89.1	64.6	94.6	-	-	-
MT-DNN	**85.3/85.0**	**91.6/88.6**	**86.8/89.2**	**79.1**	**95.7**	**93.6**	**59.5**	**90.6/90.4**

Results - Domain Adaptation





Conclusion

- MT- DNN combines multi-task learning and language model pre-training for language representation learning. MT-DNN obtains new state-of- the-art results on ten NLU tasks across three popular benchmarks: SNLI, SciTail, and GLUE. MT- DNN also demonstrates an exceptional generalization capability in domain adaptation experiments.
- There are many future areas to explore to improve MT-DNN, including a deeper understanding of model structure sharing in MTL, a more effective training method that leverages relatedness among multiple tasks, and ways of incorporating the linguistic structure of text in a more explicit and controllable manner.

Thank you !

GUO YAWEN 04/14/2019