One Model, Multiple Tasks: Pathways for Natural Language Understanding

Duyu Tang,∗ Fan Zhang∗, Yong Dai, Cong Zhou, Shuangzhi Wu and Shuming Shi

Tencent

Reporter: Xiachong Feng

Authors

Duyu Tang

Shuming Shi

https://scholar.google.com/citations?user=9uz-D-kAAAAJ&hl=zh-CN
https://scholar.google.com/citations?user=Lg31AKMAAAAJ&hl=zh-CN

Background: SuperGLUE LeaderBoard

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
+	1	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
	2	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
	3	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
+	4	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	5	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	6	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	7	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
	8	Descartes Team	frozen T5 1.1 + SPoT		89.2	91.1	95.8/97.6	95.6	87.9/61.9	93.3/92.4	92.9	75.8	93.8	66.9	83.1/82.6
	9	SPoT Team - Google	Frozen T5 1.1 + SPoT		89.2	91.1	95.8/97.6	95.6	87.9/61.9	93.3/92.4	92.9	75.8	93.8	66.9	83.1/82.6
+	10	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4

https://super.gluebenchmark.com/leaderboard

Background: Generation of Artificial Intelligence

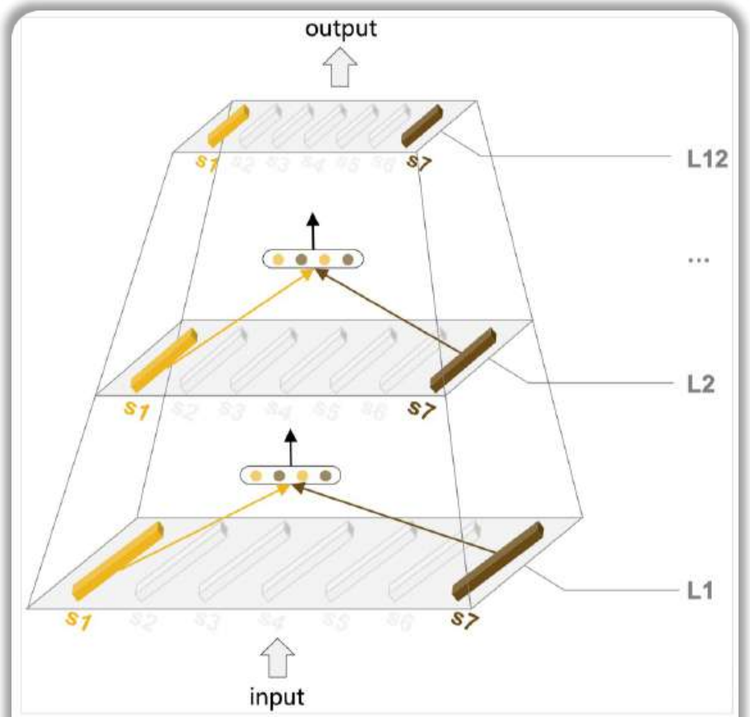
Backgrounds

- Today's AI models are typically **trained to do only one thing.** Pathways will enable us to train a single model to do thousands or millions of things.
- Today's models mostly **focus on one sense**. Pathways will enable multiple senses.
- Today's models are **dense and inefficient.** Pathways will make them sparse and efficient.

SkillNet

Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

Table 1: Examples of skills and descriptions.

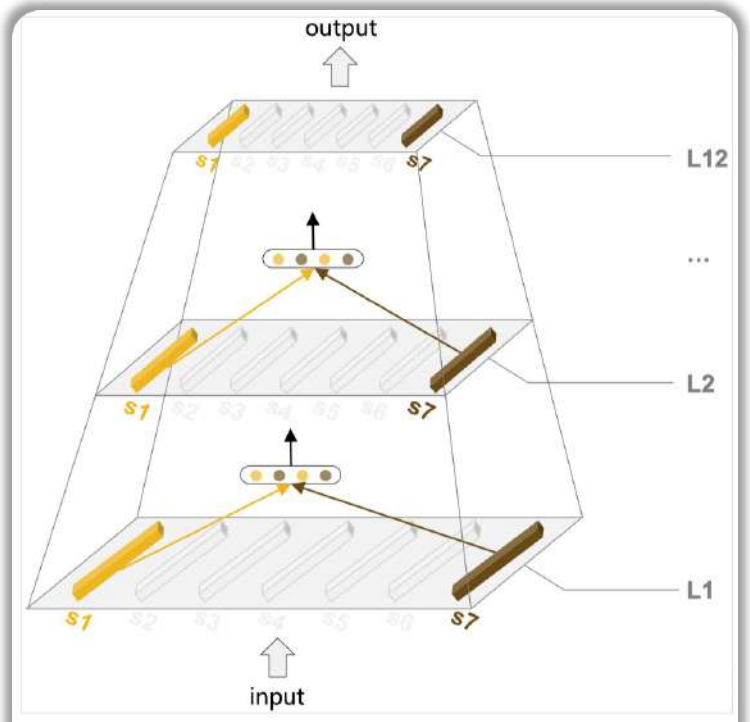


(a) SkillNet for text classification. s1 and s7 activated.

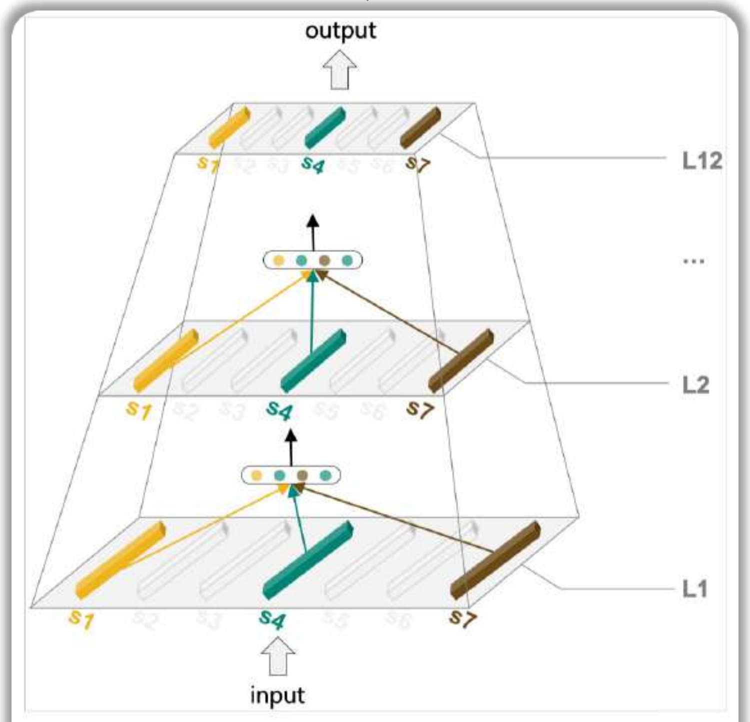
SkillNet

Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

Table 1: Examples of skills and descriptions.



(a) SkillNet for text classification. s1 and s7 activated.

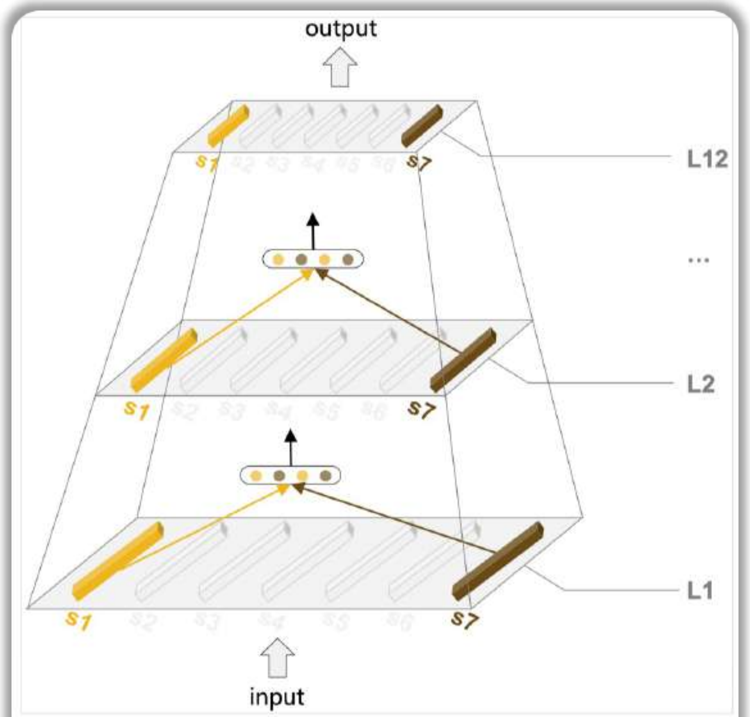


(b) SkillNet for sentiment classification. s1, s4, s7 activated.

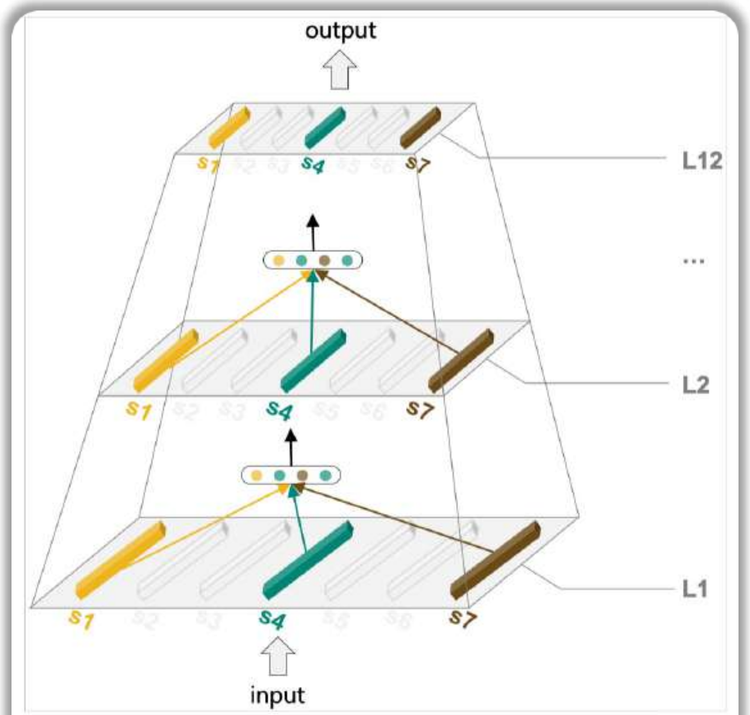
SkillNet

Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

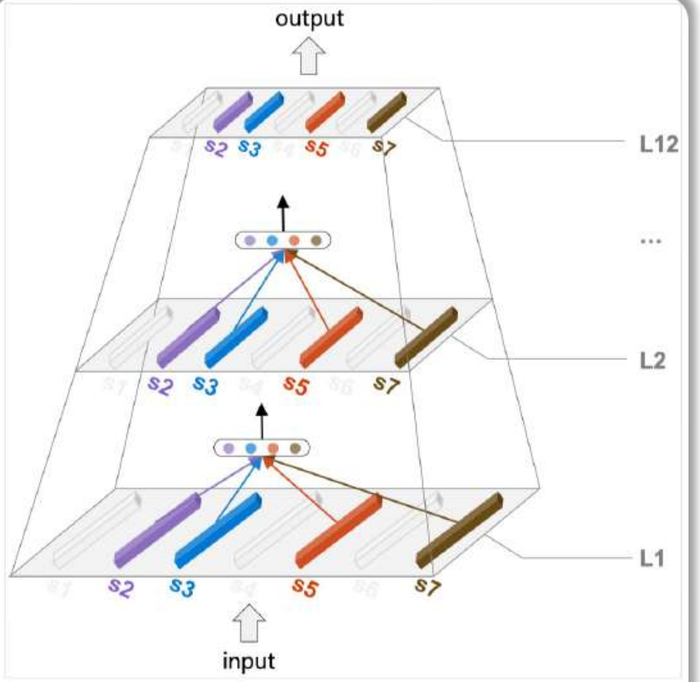
Table 1: Examples of skills and descriptions.



(a) SkillNet for text classification. s1 and s7 activated.



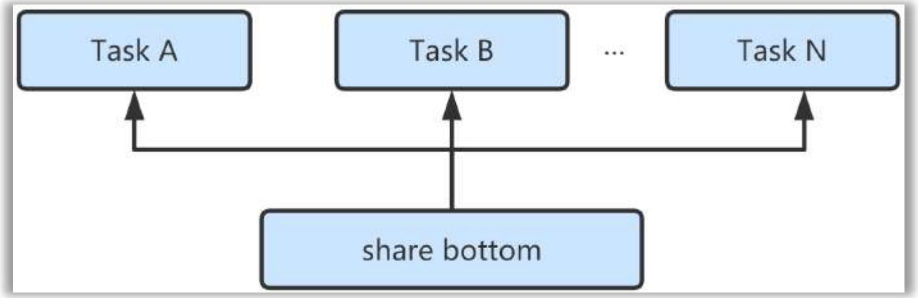
(b) SkillNet for sentiment classification. s1, s4, s7 activated.



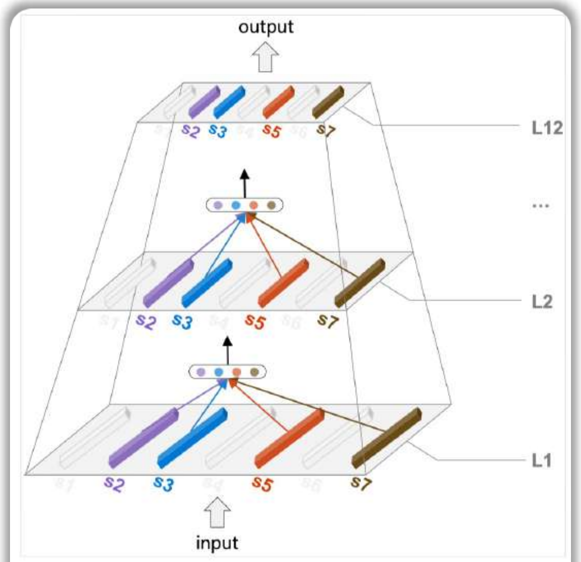
(c) SkillNet for machine reading comprehension. s2, s3, s5 and s7 activated.

SkillNet vs Multi-task Learning

- Multi-task learning methods typically have **one shared feature representation layer** (e.g., Transformer) plus multiple task-specific prediction layers.
- It is **unclear** what types of knowledge or skills are learned in the feature representation layer.



Multi-task Learning

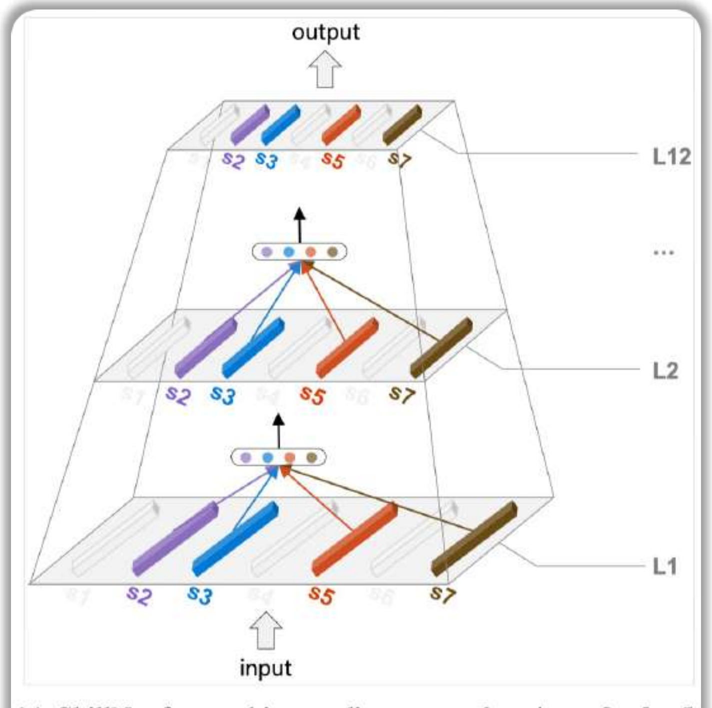


(c) SkillNet for machine reading comprehension. s2, s3, s5 and s7 activated.

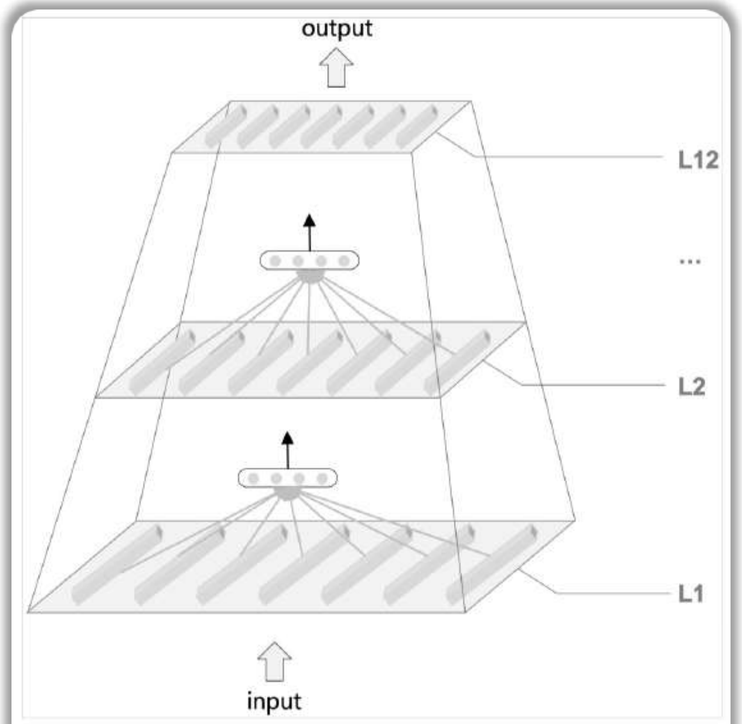
SkillNet

SkillNet vs MoE

- MoE: **fully activate** all the experts or **partially activate** a part of experts guided by an additional parameterized gating module.

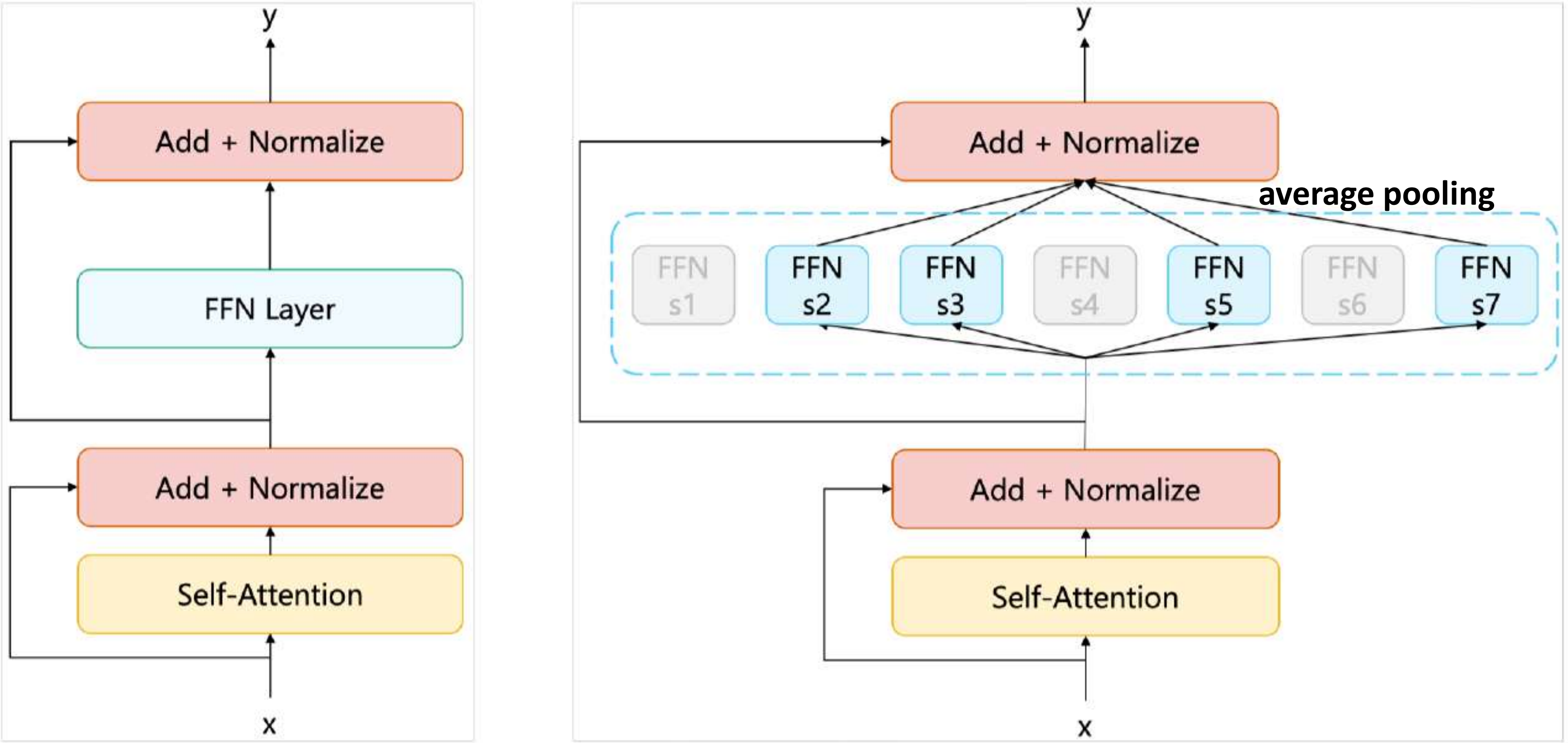


(c) SkillNet for machine reading comprehension. s2, s3, s5 and s7 activated.



(d) Mixture of experts.

Model Architecture



(a) Each layer in Transformer

(b) Each layer in SkillNet

Figure 2: A simple implementation of SkillNet (b) with comparison to the standard Transformer (a). This example illustrates the application of SkillNet to machine reading comprehension, where s2, s3, s5 and s7 are activated.

Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

Table 1: Examples of skills and descriptions.

Each stands for one particular skill. When being applied to one task, only the FFN layers corresponding to relevant skills are activated.

Tasks

Task Id	Task	Skills							Dataset
		s1	s2	s3	s4	s5	s6	s7	
T1	Sentiment Analysis	✓			✓			✓	ChnSentiCorp (9.6k / 1.2k)
T2	Natural Language Inference	✓		✓				✓	OCNLI (50k / 3k)
T3	Semantic Similarity	✓		✓			✓	✓	AFQMC (34.3k / 4.3k)
T4	Text Classification	✓						✓	TNEWS (53.3k / 10k)
T5	Named Entity Recognition		✓					✓	OntoNotes (15.7k / 4.3k)
T6	Machine Reading Comprehension		✓	✓		✓		✓	CMRC 2018 (10k / 3.4k)

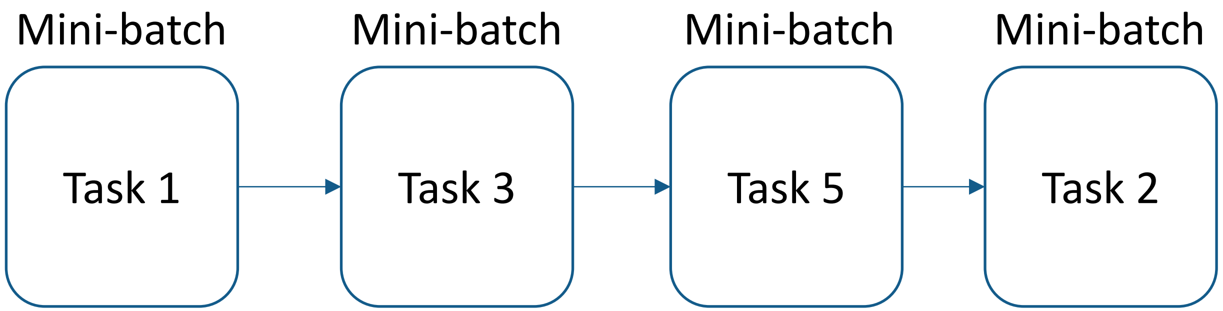
Table 2: Tasks and datasets used to train the multi-task model. Relevant skills (defined in Table 1) for each dataset is marked with a tick. The numbers of training and evaluation instances in each dataset are given in parentheses.

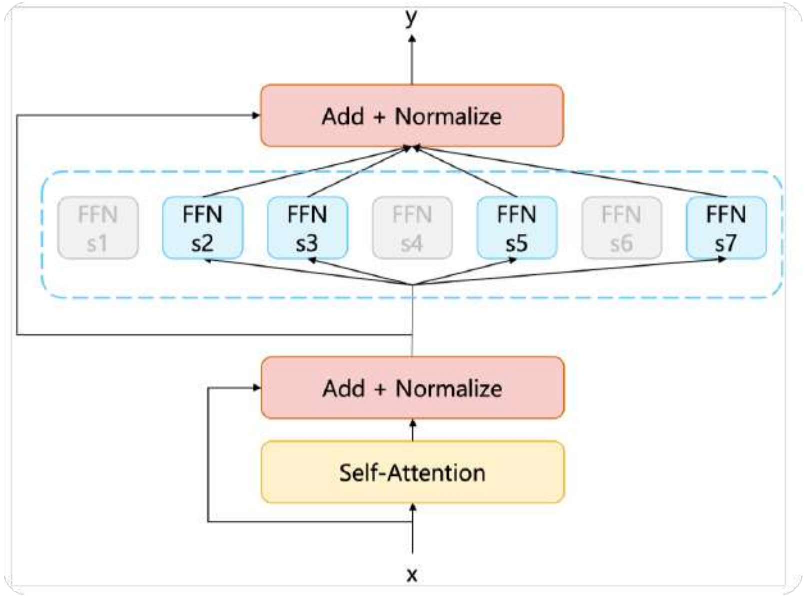
Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

Table 1: Examples of skills and descriptions.

Model Training

- The model is trained on the concatenation of training samples from these tasks.
- In each iteration, a **minibatch** is selected from one task, and the model parameters are updated according to the task-specific objective.





Adaptation to New Tasks

Skills considered in the multi-task training stage are sufficient to tackle the new task

Open domain question answering

Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

Table 1: Examples of skills and descriptions.

The new task may need new skills that are unseen in the multi-task training stage

Chinese medical question-answer matching

①

Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

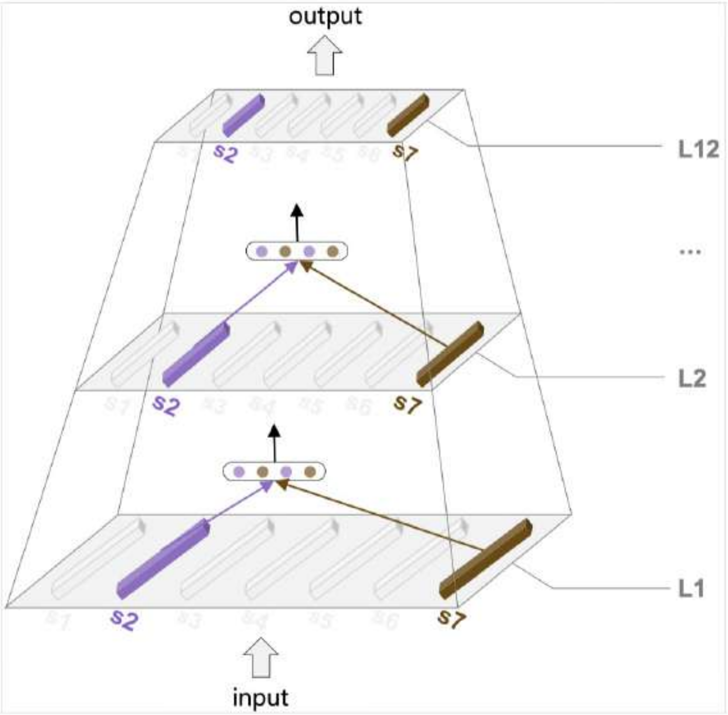
Table 1: Examples of skills and descriptions.

②

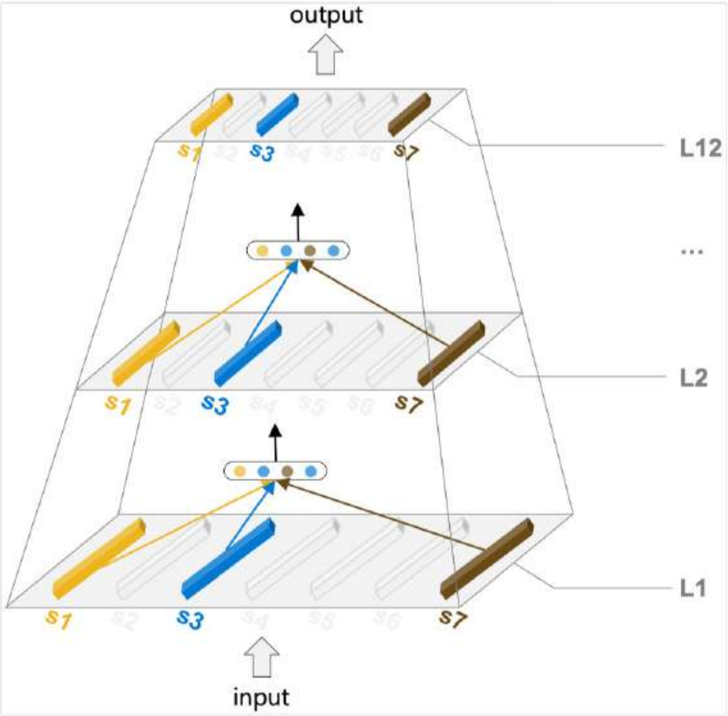
Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill
S8	understanding texts in the medical domain

Table 1: Examples of skills and descriptions.

Pre-training



(a) pre-training with masked language modeling. s2 and s7 activated.



(b) pre-training with next sentence prediction. s1, s3, s7 activated.

Figure 3: An illustration of how SkillNet (our Pathways model) is pre-trained with masked language modeling and next sentence prediction. The model is sparsely activated during pre-training. Skills are defined in Table 1.

Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

Table 1: Examples of skills and descriptions.

Experiments

- **Task-specific fine-tuning**
  - We fine-tune all the parameters of our **BERT model for each task individually**. Therefore, we have a total of six task specific models in our experiments.

- **Joint fine-tuning (Dense) --> Multi-task Learning**
  - We adopt our BERT as a shared model to obtain feature representation and then feed it to **multiple task-specific prediction layers**. The parameters of the BERT model and all the top layers are learned jointly on the six tasks.

- **Joint fine-tuning (MoE) --> Mixture of Experts**
  - Following Shazeer et al. (2017), we set the number of the FFNs in each layer **as seven** and activate the top-2 FFNs for each token**, determined by a gating module**. The parameters of these FFNs are initialized with our BERT model and updated with the task-specific prediction layers.

Results

Task Id	Task
T1	Sentiment Analysis
T2	Natural Language Inference
T3	Semantic Similarity
T4	Text Classification
T5	Named Entity Recognition
T6	Machine Reading Comprehension

	T1	T2	T3	T4	T5	T6	**Avg**
BERT Fine-tuning	**94.7**†	74.6†	**74.2**‡	56.1‡	78.2*	84.5†	77.1
Task-specific fine-tuning	94.3	75.0	72.3	56.9	79.2	84.8	77.1
Joint fine-tuning (Dense)	93.4	75.1	71.0	**57.4**	78.2	83.8	76.5
Joint fine-tuning (MoE)	94.0	74.0	71.4	57.3	78.8	84.5	76.7
SkillNet w/o sparse pre-training	94.1	**75.3**	72.1	56.9	81.2	84.6	77.4
SkillNet w/ sparse pre-training	94.4	75.0	73.9	57.0	**81.5**	**85.7**	**77.9**

Table 3: Evaluation results on the six tasks during multi-task training. We report accuracy for T1 $\sim$ T4 and F1 for T5 $\sim$ T6. **Avg** is the average score of all tasks. Results with †, ‡ and * are based on google BERT from Cui et al. (2021), Xu et al. (2020) and our experiments, respectively.

Results on New Tasks

- Open domain question answering

Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

Table 1: Examples of skills and descriptions.

	#Params Activated	Dev	Test
BERT Fine-tuning†	102M	80.7	80.8
Task-specific fine-tuning (BERT-base)	102M	80.3	80.9
Task-specific fine-tuning (RoBERTa-large)	326M	82.7	83.2
Joint fine-tuning (Dense)	102M	80.7	81.6
Joint fine-tuning (MoE)	159M	81.0	82.4
SkillNet w/o sparse pre-training	272M	81.5	83.2
SkillNet w/ sparse pre-training	272M	**83.9**	**84.4**

Table 4: Evaluation results on the NLPCC-DBQA dataset. We report the F1 score on the dev and test set. Results with † are based on google BERT from Sun et al. (2019).

Results on New Tasks

- Chinese medical question-answer matching

Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

S8 understanding texts in the medical domain

Table 1: Examples of skills and descriptions.

	Update Old Skills	#Params Activated	Dev	Test
BERT Fine-tuning†		110M	78.6	78.2
Task-specific fine-tuning (BERT-base)		102M	78.4	78.1
Task-specific fine-tuning (RoBERTa-large)		326M	78.9	78.7
Joint fine-tuning (Dense)		102M	78.5	78.3
Joint fine-tuning (MoE)		159M	78.7	78.4
No New Skills				
SkillNet w/o sparse pre-training	Y	272M	78.8	78.6
SkillNet w/ sparse pre-training	Y	272M	79.0	78.9
Injecting New Skills				
SkillNet w/o sparse pre-training	N	57M	77.8	77.1
SkillNet w/ sparse pre-training	N	57M	78.6	78.2
SkillNet w/o sparse pre-training	Y	329M	79.2	79.0
SkillNet w/ sparse pre-training	Y	329M	**79.5**	**79.3**

Table 5: Evaluation results on the cMed dataset. We report the top-1 accuracy on the dev and test set. Results with † are based on google BERT from Cui and Han (2020).

Ablation Study and Analysis

- Average score decrease when any skill is removed in the SkillNet model.
- There is a significant drop when deleting the general skill s7.
- The task performance drops sharply when some closely related skills are removed
  - s4 “understand sentiment” for T1 “sentiment analysis”
  - S5 “understand questions” for T6 “MRC”
  - S6 “understand texts in finance domain” for T3 “Semantic Similarity”
- Removing s2 significantly affects the performance on NER and MRC

Task Id	Task
T1	Sentiment Analysis
T2	Natural Language Inference
T3	Semantic Similarity
T4	Text Classification
T5	Named Entity Recognition
T6	Machine Reading Comprehension

Skill	Description
s1	get the semantic meaning of a sequence
s2	get the semantic meaning of a token
s3	understand how two text segments interact
s4	understand the sentiment of texts
s5	understand natural language questions
s6	understand texts in finance domain
s7	generic skill

Table 1: Examples of skills and descriptions.

	T1	T2	T3	T4	T5	T6	Avg
SkillNet	94.08	**75.25**	**72.13**	56.94	**81.19**	**84.64**	**77.37**
– w/o s1	94.06	74.08	70.44	56.57	80.65	84.12	76.65
– w/o s2	**94.24**	75.22	71.34	**57.11**	78.82	83.55	76.71
– w/o s3	93.50	74.07	71.62	57.07	79.84	83.72	76.64
– w/o s4	93.42	74.87	72.06	56.99	78.70	84.08	76.69
– w/o s5	94.15	74.75	71.66	57.08	78.84	83.61	76.68
– w/o s6	93.43	73.63	71.28	56.87	80.86	84.23	76.72
– w/o s7	94.04	74.85	71.99	56.30	78.14	84.22	76.59

Table 6: Ablation results on the six tasks during multi-task training.

Influence of The Sampling Rate

- We can see that the model performs better when the sampling rate α = 1.0, which maintains the natural distribution of the task.

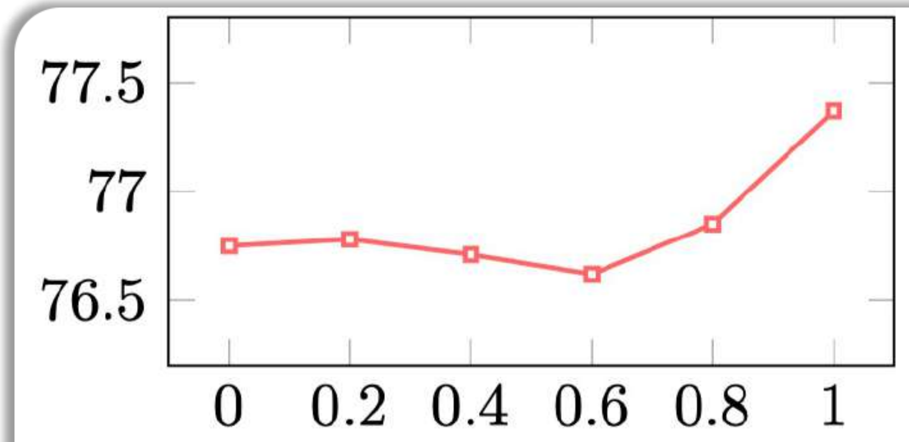


Figure 4: Average score with different α.

Influence of The Number of Top Pathways Layers

- The performance consistently improves as the number grows, demonstrating the effectiveness of our SkillNet model.
- The underlying reason is that when more Pathways layers are incorporated, the skills are better learned as the number of parameters increases.

#Num	#Params Total	Avg
3	187M	76.5
6	272M	76.9
9	357M	77.2
12	422M	77.4

Table 7: The number of total parameters and average score with the different number of top Pathways layers.

Conclusion

- Present a multi-task Pathways model called SkillNet and its application to natural language understanding tasks.
- SkillNet includes a set of parameterized skill modules and sparsely activate some of the modules depending on whether a skill is relevant to the target task.
- The framework is generic and supports both multi-task fine-tuning and pre-training, both with sparse activation.
- Results demonstrate that the approach performs better than baseline systems on both old and new tasks, and sparse pre-training brings further improvements.

Thanks~