

---

# One Model, Multiple Modalities: A Sparsely Activated Approach for Text, Sound, Image, Video and Code

---

Yong Dai\*, Duyu Tang\*, Liangxin Liu\*, Minghuan Tan\*, Cong Zhou\*, Jingquan Wang\*,  
Zhangyin Feng, Fan Zhang, Xueyu Hu, Shuming Shi

Tencent AI Lab

## Abstract

People perceive the world with multiple senses (e.g., through hearing sounds, reading words and seeing objects). However, most existing AI systems only process an individual modality. This paper presents an approach that excels at handling multiple modalities of information with a single model. In our “**SkillNet**” model, different parts of the parameters are specialized for processing different modalities. Unlike traditional dense models that always activate all the model parameters, our model sparsely activates parts of the parameters whose skills are relevant to the task. Such model design enables SkillNet to learn skills in a more interpretable way. We develop our model for five modalities including text, image, sound, video and code. Results show that, SkillNet performs comparably to five modality-specific fine-tuned models. Moreover, our model supports self-supervised pretraining with the same sparsely activated way, resulting in better initialized parameters for different modalities. We find that pretraining significantly improves the performance of SkillNet on five modalities, on par with or even better than baselines with modality-specific pretraining. On the task of Chinese text-to-image retrieval, our final system achieves higher accuracy than existing leading systems including Wukong<sub>VIT-B</sub> and Wenlan 2.0 while using less number of activated parameters.

## 1 Introduction

In recent years, Transformer [40] and Transformer-based pretrained models [12, 35] have revolutionized natural language processing [33] and there have been growing interests in extending the successful paradigm to broader artificial intelligence areas including computer vision [8, 23, 32], speech processing [4] and program analysis [18]. Researchers from different communities have no communication barrier and typically repeat the same process: pretraining for each modality and finetuning all the model parameters for each task.

Despite the remarkable progress made in artificial intelligence, existing methods differ from human learning in the following three aspects [1]. First, we human perceive the world using multiple senses. We know that the word “dog”, the bark of a dog and the image/video of a dog all refer to the same concept. However, most existing methods only process one modality of information. Second, the human brain has around 100 billion neurons, of which different parts are specialized for different skills. When we accomplish a task, we only call upon a small fraction of neurons that are relevant to the task. However, most existing methods activate all the model parameters. Third, when we solve a new problem or learn a new skill, we don’t learn from nothing but combine old skills to learn

---

\*Correspondence to: Duyu Tang (duyutang@tencent.com), \* indicates equal contribution

new things quickly. However, existing methods typically learn for each task from scratch (or from a general or foundation model), resulting in hundreds of models for hundreds of tasks.

In this work, we propose a multitask multimodal approach called SkillNet. We use a single model to handle multiple tasks that require the understanding of different modalities of information. In SkillNet, different parts of the parameters are specialized for different skills. When the model is applied to a downstream task, unlike traditional “dense” models that always activate all the model parameters, it “sparsely” activates parts of the parameters whose skills are relevant to the target task. For example, we could define five modality-related skills  $\{s_{text}, s_{image}, s_{sound}, s_{video}, s_{code}\}$ , which are specialized for understanding text, image, sound, video and code, respectively. Consider the task of automatic speech recognition (ASR), which only relates to the skill of auditory understanding (i.e.,  $s_{sound}$ ). When SkillNet is applied to ASR, model parameters related to other four skills (i.e.,  $\{s_{text}, s_{image}, s_{video}, s_{code}\}$ ) are deactivated. Similarly, for text-to-image retrieval, which is to find semantically related images given texts, only  $s_{text}$  and  $s_{image}$  are activated. Figure 1 gives high-level illustrations of the aforementioned situations. There are many different ways to implement SkillNet. In this work, we provide a simple implementation on top of Transformer [40]. Instead of producing general  $K/Q/V$  vectors for each token, we activate different modality-specific parameters to produce different modality-specific  $K/Q/V$  vectors before conducting multi-head attention. The intuition is that we expect the model to call upon different parts as needed to process different types of signals and combine information from multiple senses to form our understanding about a concept (like the aforementioned example about the concept of dog).

We conduct experiments on tasks of five modalities, including text classification, automatic speech recognition, text-to-image retrieval, text-to-video retrieval and text-to-code retrieval. Results show that, SkillNet performs comparably to five modality-specific models with only one model file. Furthermore, after being pretrained, SkillNet performs better than systems with modality-specific pretraining on three of five modalities. On the task of Chinese text-to-image retrieval, SkillNet obtains higher accuracy than existing systems (e.g., Wukong<sub>ViT-B</sub> and Wenlan 2.0) while using less number of activated parameters. Our work demonstrates the feasibility of developing one general model that is both accuracy and efficient to tackle multiple tasks of different modalities.

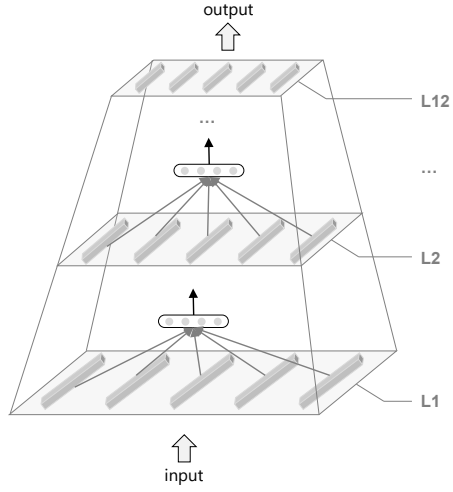
## 2 Comparison to Existing Methods

We describe the connections and differences of this work to related multimodal, multitask and mixture-of-experts methods.

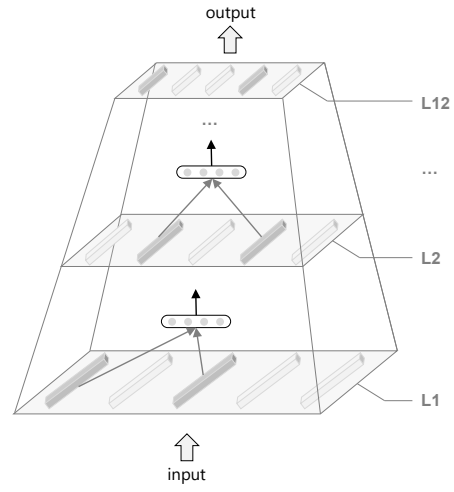
**Multimodal** Since there are large amounts of multimodal works, we only describe the closely related ones. Omnivore [19] uses a single model to process multiple visual modalities, including single-view 3D data, images and videos. VATT [1] learns multimodal representations on raw signals for video, audio and text. Compared to Omnivore and VATT, our work studies more modalities and our approach is sparse. Data2vec [5] is a general learning objective that manipulates over latent representations instead of modality-specific tokens. The same learning objective is used to learn for text, speech and vision. However, they don’t perform multitask training. Our work is orthogonal to Data2vec and it is interesting to combine the advantages of Data2vec and SkillNet.

**Multitask** This work also relates to multitask learning methods. Systems built upon Transformer typically use shared feature encoder plus task-specific prediction layers for understanding tasks [29] and use natural language prompts to steer encoder-decoder model for generation tasks [37]. Most existing multitask methods are dense—all the model parameters are activated. An exception is SkillNet-NLU and SkillNet-NLG [28, 46], recently introduced sparse models that perform multitask learning on text. This work can be viewed as an extension to the multimodal situation.

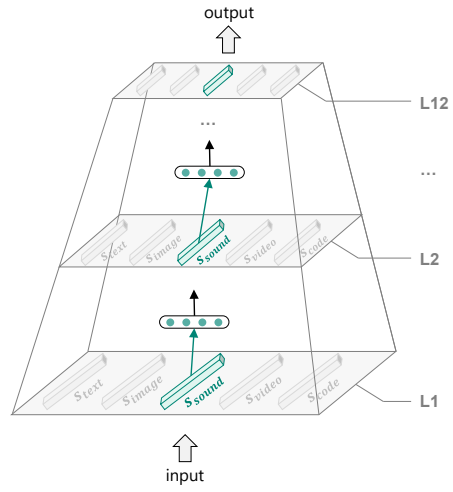
**Mixture-of-Expert (MoE)** Transformer-based MoE methods typically include multiple homogeneous neural networks (called experts), which can be fully activated or partially activated guided by an additional gating function [15, 16, 27, 38]. However, it is unclear what type of knowledge is learned in each expert and why an expert is activated. From this point of view, our approach can be viewed as a sparse multimodal MoE. Unlike traditional MoE methods, each expert in our model has a clear definition and the activation of each expert has a clear reason (judged by human experts).



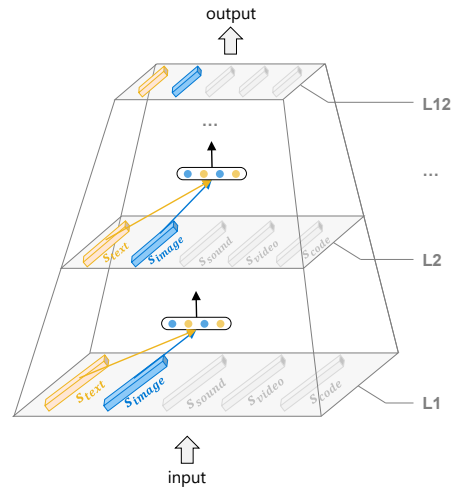
(a) Fully activated dense model



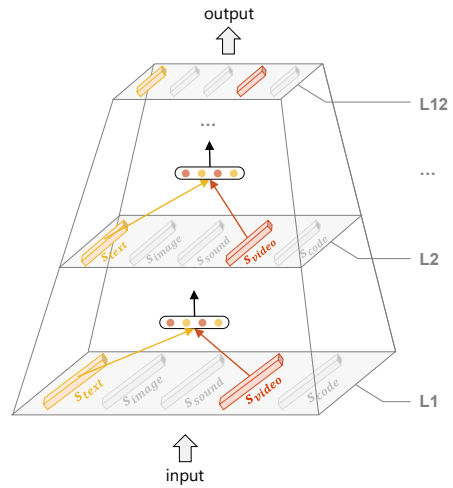
(b) Sparsely activated MoE



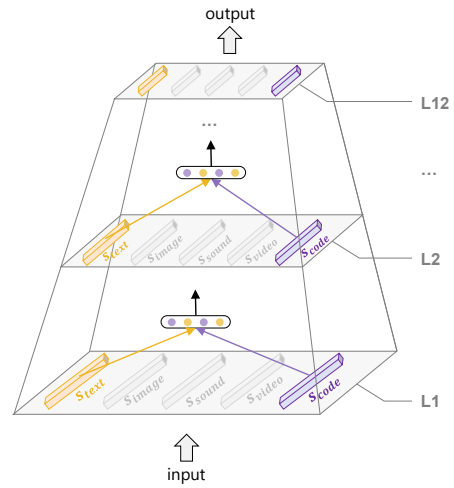
(c) SkillNet for ASR.



(d) SkillNet for text-to-image retrieval



(e) SkillNet for text-to-video retrieval



(f) SkillNet for text-to-code retrieval

Figure 1: In SkillNet, each pillar refers to a skill. Pillars filled in color (e.g., yellow, blue, green, purple and red) are activated.

For example, the expert corresponding to  $s_{text}$  is responsible for understanding text signal and it is activated only if the input signal is text.

### 3 Method

This section gives our sparsely activated model SkillNet. We first give a brief background on Transformer (§3.1). Then, we describe the model architecture of SkillNet (§3.2). Finally, we describe how to produce the embeddings for different modalities (§3.3).

#### 3.1 Background on Transformer

To make our paper self-contained, we briefly describe Transformer here. Transformer [40] is a commonly used model architecture with multiple layers, each of which consists of a multi-head attention layer followed by a feed-forward network (FFN) layer. The multi-head attention mechanism concatenates the output vectors of  $H$  different heads and then linearly projects them by  $W^O$ :

$$\text{Multi-Head}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_H)W^O, \quad (1)$$

where  $Q$  (Query),  $K$  (Key),  $V$  (Value) are the hidden representations of the previous layer. In each head,  $Q$ ,  $K$  and  $V$  are transformed with projection matrices before being fed to the attention function:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \quad (2)$$

where  $W_i^Q$ ,  $W_i^K$  and  $W_i^V$  are model parameters, and  $i$  denotes the  $i$ -th head. The attention function computes the dot products of the query with all keys, and uses softmax to obtain the weights on values:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (3)$$

where  $d_k$  is the dimension of key. Finally, the FNN layer is applied to obtain the final representations.

Residual connection [22] and layer normalization [3] are adopted for both multi-head attention layer and FFN layer. Since Transformer is prevalent, we exclude the details and refer readers to the original paper.

We use image search via Siamese network as the running example to show how to apply Transformer to downstream tasks. As shown in Figure 2, two Transformers are used to encode the text and the image, respectively. For each side, we take the vector of the first token ([CLS]) to represent the input. The semantic similarity between text and image is computed using dot product or cosine.

#### 3.2 Architecture of SkillNet

We build our SkillNet model by using Transformer [40] as the backbone. Specifically, we modify the multi-head attention of each Transformer layer as follows. Instead of producing general  $K/Q/V$  vectors for each token, we activate different modality-specific parameters to produce different modality-specific  $K/Q/V$  vectors before conducting multi-head attention. Take  $Q$  as an example. Instead of having only one projection matrix  $W_i^Q$  for all queries, we have five projection parameter matrices  $\{W_i^{Q_{text}}, W_i^{Q_{image}}, W_i^{Q_{sound}}, W_i^{Q_{video}}, W_i^{Q_{code}}\}$ , of which each item stands for a skill of understanding the information of a particular modality. When the model is applied to a task, we only activate the corresponding projection matrices of relevant skills. Similar modifications are made for keys and values. The computation of a head is modified as follows.

$$\text{head}_i^{\text{skill}} = \text{Attention}(QW_i^{Q_{Activated}}, KW_i^{K_{Activated}}, VW_i^{V_{Activated}}) \quad (4)$$

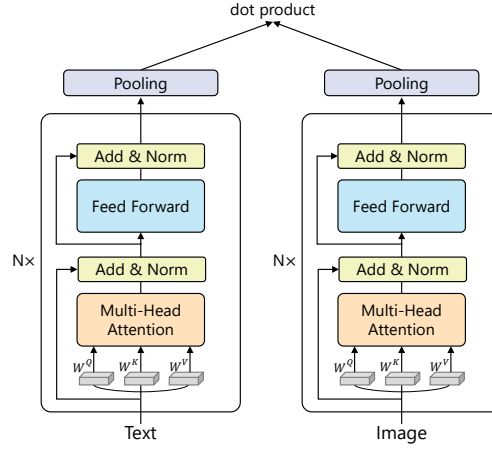


Figure 2: An illustration of image search with Transformer-based Siamese network.

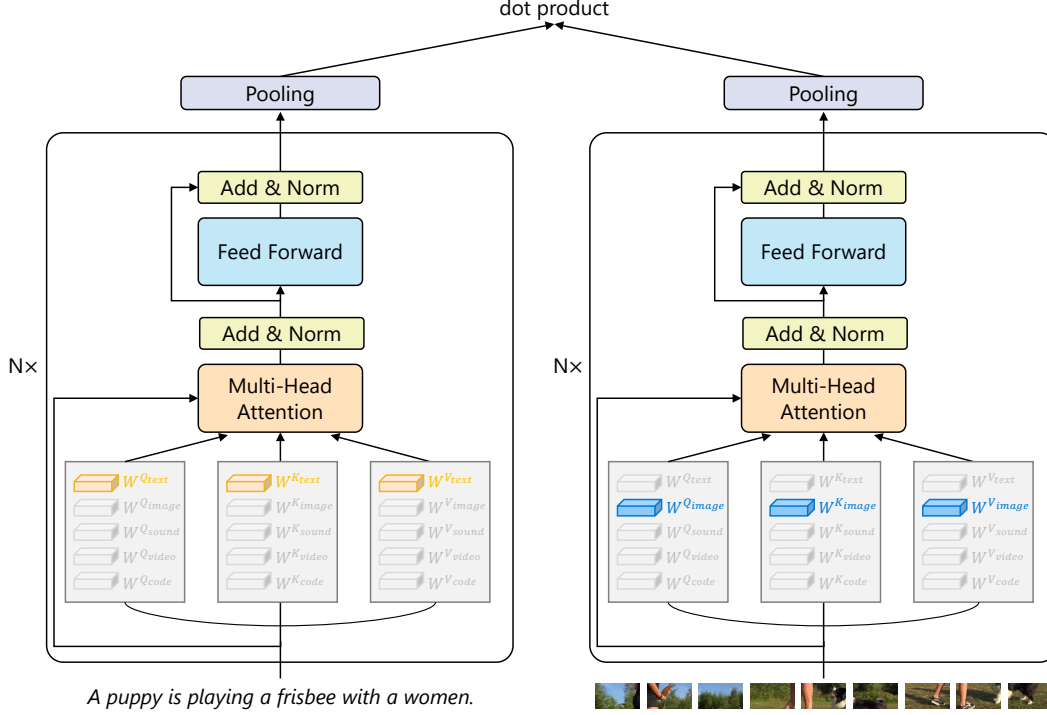


Figure 3: Architecture of SkillNet for image retrieval. Text encoder and image encoder are two pathways of one shared model —  $s_{text}$  and  $s_{image}$  are activated for the text encoder and the image encoder, respectively.

$$W_i^{Q_{Activated}} = \begin{cases} W_i^{Q_{text}} & \text{if } s_{text} \text{ is activated} \\ W_i^{Q_{image}} & \text{if } s_{image} \text{ is activated} \\ W_i^{Q_{sound}} & \text{if } s_{sound} \text{ is activated} \\ W_i^{Q_{video}} & \text{if } s_{video} \text{ is activated} \\ W_i^{Q_{code}} & \text{if } s_{code} \text{ is activated} \end{cases} \quad (5)$$

As shown in Figure 4, we only need one model to handle the the task of image retrieval, where we activate  $s_{text}$  and  $s_{image}$  for the text encode and image encoder, respectively.

### 3.3 Embeddings

We describe how to produce the embeddings for different modalities.

**Text** Following BERT [12], we tokenize a text into a sequence of wordpiece tokens [45] and build the embedding of each wordpiece by adding up its token embedding, position embedding and segment embedding. We also add a special classification token  $[CLS_{text}]$  at the beginning of a sequence to produce the representation of the sequence. If the input includes two segments, we add a special token  $[SEP]$  between the two segments.

**Sound** Given a raw waveform as the input, we follow wav2vec [4] and use convolutional network to produce a sequence of vectors as the embeddings. Specifically, we use seven convolutions with 512 channels, strides of (5,2,2,2,2,2,2) and kernel widths of (10,3,3,3,3,2,2) to generate a vector sequence from a 20ms framerate sampled at 16KHz. After that, we adopt a 1D convolutional network to transform the vector sequence to 768 dimensional embeddings, which are summed up with their corresponding position embeddings as the final sound embeddings.

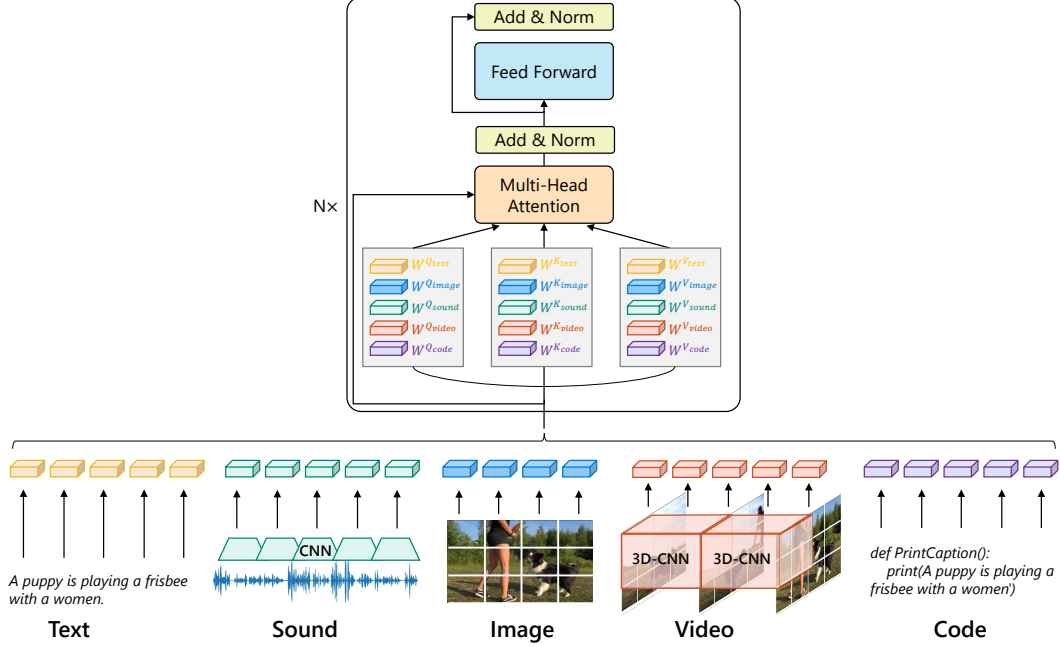


Figure 4: An illustration of the pipeline and the embeddings of different modalities.

**Image** Following Vision Transformer (ViT) [13], we build patch embeddings for each image. We first reshape each image of  $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$  into 2D patches of  $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$ , where  $(H, W)$  is the image resolution,  $(P, P)$  is the resolution of each patch,  $N$  is the number of patches and  $C$  is the number of image channels (e.g. 3 for RGB). Then, a 2D convolutional network is applied to transform patch pixels to 768 dimensional embeddings, which are added with the corresponding position embeddings as the final patch embeddings.<sup>2</sup> We add a special token  $[\text{CLS}_{\text{image}}]$  at the beginning of each sequence to produce the representation of the image.

**Video** We follow Vivit [2], an extension of ViT for video, to produce video embeddings. Given a video  $V \in \mathbb{R}^{T \times H \times W \times C}$ , where  $T$  is the number of sampled frames, we extract  $\lceil T/t \rceil \cdot \lceil H/h \rceil \cdot \lceil W/w \rceil$  non-overlapping, spatio-temporal “tubes” and use a 3D convolution to produce a representation for each tube. We further add  $\lceil T/t \rceil + \lceil H/h \rceil + \lceil W/w \rceil$  positional embeddings and concatenate a special token  $[\text{CLS}_{\text{video}}]$  at the beginning of each sequence to represent the whole video input.

**Code** We follow CodeBERT [18] to produce code embeddings. We tokenize a code snippet to a sequence of code-specific wordpiece tokens. The final embedding of each token is the sum of token embedding, position embedding and segment embedding. A special token  $[\text{CLS}_{\text{code}}]$  is added to the beginning of each sequence to produce the embedding of the entire code.

## 4 Tasks

In this section, we first describe downstream tasks involving five modalities in §4.1. Each modality relates to an active research area that covers many tasks. We select one task for each modality with preferences for well recognized tasks (e.g., ASR) and tasks relate to multiple modalities (e.g., video/code retrieval). Since our framework also supports sparsely activated pretraining, we conduct multimodal pretraining to initialize the model parameters. The pretraining tasks are described in §4.2

Task Id	Task	Skills				
		$s_{text}$	$s_{image}$	$s_{sound}$	$s_{video}$	$s_{code}$
T1	Text Classification	✓				
T2	Automatic Speech Recognition			✓		
T3	Text-to-Image Retrieval	✓	✓			
T4	Text-to-Video Retrieval	✓			✓	
T5	Text-to-Code Retrieval	✓				✓

Table 1: Relations between tasks and skills. Relevant skills for each task are marked with ticks.

#### 4.1 Downstream Tasks

**Text** Text classification is a classic and fundamental text understanding task [34]. Given a sentence as the input, the task is to predict which category the sentence belongs to. Following BERT [12], we add a  $[CLS_{text}]$  token at the beginning of each sentence to represent the meaning of the whole sentence. For the task of text classification, only the parameters that relate to  $s_{text}$  are activated.

**Sound** Automatic speech recognition (ASR) is to convert speech to text [24]. Following wave2vec [4], we produce speech features and generate a transcription by performing token-level classification. Connectionist temporal classification loss [20] is adopted for model training. For the task of ASR, only the parameters that relate to  $s_{sound}$  are activated.

**Image** We consider text-to-image retrieval. Given a text as the query, the task is to find the target image from a set of candidates. Considering the efficiency of the inference stage, we use two separate passes (like Siamese Network) to produce text and image vectors separately with no cross-modality attention. Notably, we use the same model with different activation configurations (i.e.,  $s_{text}$  is activated for text and the  $s_{image}$  is activated for image) to produce text and image vectors. The semantic similarity between a text and an image is calculated with dot product or cosine function.

**Video** We consider text-to-video retrieval. Given a text as the query, the task is to find the target video from a set of candidates. The framework is similar to the aforementioned image retrieval. We use the same model with different activated parameters (i.e.,  $s_{text}$  is activated for text and  $s_{video}$  is activated for video) to produce text and video vectors separately.

**Code** We consider natural language code retrieval. Given a text as the query, the task is to find the most relevant code from a set of candidates. We use the same model with different activated parameters (i.e.,  $s_{text}$  for text and  $s_{code}$  for code) to produce text and code vectors separately. The framework is similar to image retrieval.

#### 4.2 Pretraining Tasks

Recap that our approach also supports multimodal pretraining with sparse activation. We describe the pretraining tasks for each modality here.

**Text** We adopt masked language modeling (MLM) as the pre-training task [12, 30]. Given a text, we randomly mask 15% of the tokens. Each masked token is replaced with a special [MASK] token 80% of the time, a random token 10% of the time, and left unchanged for the remaining 10% of the time.

**Sound** We develop a simplified version of HuBERT [25] and pretrain through predicting the categories of the masked sound tokens, whose target labels are produced with an offline clustering process. We set the number of clusters to 100 and use k-mean clustering with Mel-Frequency Cepstral Coefficients (MFCCs) acoustic features. We use the same masking strategies of wav2vec2 [4], where about 5% of the time-steps are randomly sampled as start indices and the subsequent 10 time-steps are masked.

<sup>2</sup>In this work, we use different positional embeddings for different modalities.



Method	Text	Image	Sound	Video	Code
Modality-specific models	0.48	69.63	0.20	63.18	53.97
Dense multimodal baseline	0.48	55.70	0.23	19.46	57.59
MoE multimodal baseline	0.49	60.93	0.19	64.81	50.04
SkillNet w/o pretraining	0.48	68.76	0.20	66.49	60.14
Baselines with modality-specific pretraining	0.56*	71.70 <sup>†</sup>	0.17	77.31	66.33
<b>SkillNet</b>	<b>0.57</b>	<b>73.59</b>	<b>0.17</b>	<b>81.77</b>	<b>70.66</b>

Table 2: Results on five tasks. Evaluation metrics for five modalities are accuracy, Recall@10, CER (lower is better), Recall@10 and Recall@10, respectively. The result tagged with \* is from [46], whose pretraining text corpus is the superset of our work. The result tagged with <sup>†</sup> is from the previous best system for Chinese text-to-image retrieval [21], whose pretraining image corpus is also the superset of our image pretraining data.

**Image** We follow CLIP [36] and use contrastive objectives for pretraining. We use the same architecture for image retrieval as illustrated in §4.1 and adopt in-batch negative sampling.

**Video** Similar to the configuration of image pretraining, we consider a contrastive pretraining task of text-video matching. In-batch negative sampling is adopted.

**Code** Like CodeBERT [18], we concatenate code and text, separate them with [SEP] and randomly mask 15% of the tokens. The pretraining task is to predict the masked tokens.

## 5 Experiments

### 5.1 Setup

We compare to the following baselines.

- **Modality-specific models.** We train five different models for different modalities. The model architecture for each modality is the standard Transformer.
- **Dense multimodal baseline.** We train a multimodal model that jointly learns for five modalities. This is a dense model in that all these modalities share a common standard Transformer architecture, which is equivalent to SkillNet with only one skill and that skill is always activated.
- **MoE multimodal baseline.** We train a Mixture-of-Expert (MoE) [27] baseline and set the number of experts as the number of skills of SkillNet (i.e., 5). There is a gating function to selectively active top-2 experts for each token.

We implement SkillNet on top of HuggingFace’s Transformers [42]. We conduct experiment with 12 Transformer encoder layers and 768 hidden state dimensions and leave the extension to larger model scales to the future. Since the parameters of SkillNet can be pretrained (as described in §4.2), we have two model configurations, depending on whether the parameters are pretrained in the same sparsely activated manner. We also compare to baselines with modality-specific pretraining. For text, we compare to [46], which uses the superset of our text pretraining corpus to pretrain BERT. For image, we compare to Wukong<sub>ViT-B</sub> [21], which has the similar model scale (with 12 Transformer layers) and is pretrained with a superset of our image pretraining data. For speech, video and code, we pretrain modality-specific models with the same amount of pretraining data of SkillNet.

Details about the datasets and training process are given in the Appendix.

### 5.2 Results and Analysis

Table 2 gives the results on five tasks. Systems in the first group are not pretrained. We can see that SkillNet performs comparably to modality-specific models. An interesting finding is that the joint model with a dense encoder is not friendly to the low-resource task like text-to-video, but this phenomenon does not exist in either MoE system or SkillNet. The second group includes systems with modality-specific pretraining or joint pretraining. We can see that pretraining consistently



improves the performance of SkillNet on all five tasks, even better than modality-specific pretraining on image, video and code.

On the task of text-to-image retrieval, SkillNet achieves better accuracy compared to existing leading systems but using less number of activated parameters. Numbers are given in Table 3. Since the parameters of Wenlan 2.0 and Wukong<sub>ViT-B</sub> are not reported in their papers, we calculate their parameters based on their model descriptions. The parameters of Wenlan 2.0 [17] include three parts, an image encoder consisting of an EfficientNet-B7 [39] (66M) and four Transformer encoder layers (50M), a text encoder RoBERTa-Large [10] (326M) and a cross-modal projection layer with two fully-connected layers (3M). Wukong<sub>ViT-B</sub> [21] includes a Vision Transformer (ViT) [14] (86M) as the image encoder, a standard decoder-only transformer (110M) as the text encoder and a linear cross-modal projection layer (0.6M).

Method	Number of Activated Params	R@1	R@10
Wenlan 2.0 [17]	445M	34.1	69.1
Wukong <sub>ViT-B</sub> [21]	197M	36.7	71.7
SkillNet	124M	37.0	73.6

Table 3: Performance and activated model parameters of text-to-image retrieval methods.

Figure 5 shows the learning curves of SkillNet with or without pretraining on different tasks. We can see that in general pretraining gives the model a good starting point and leads to better accuracy.

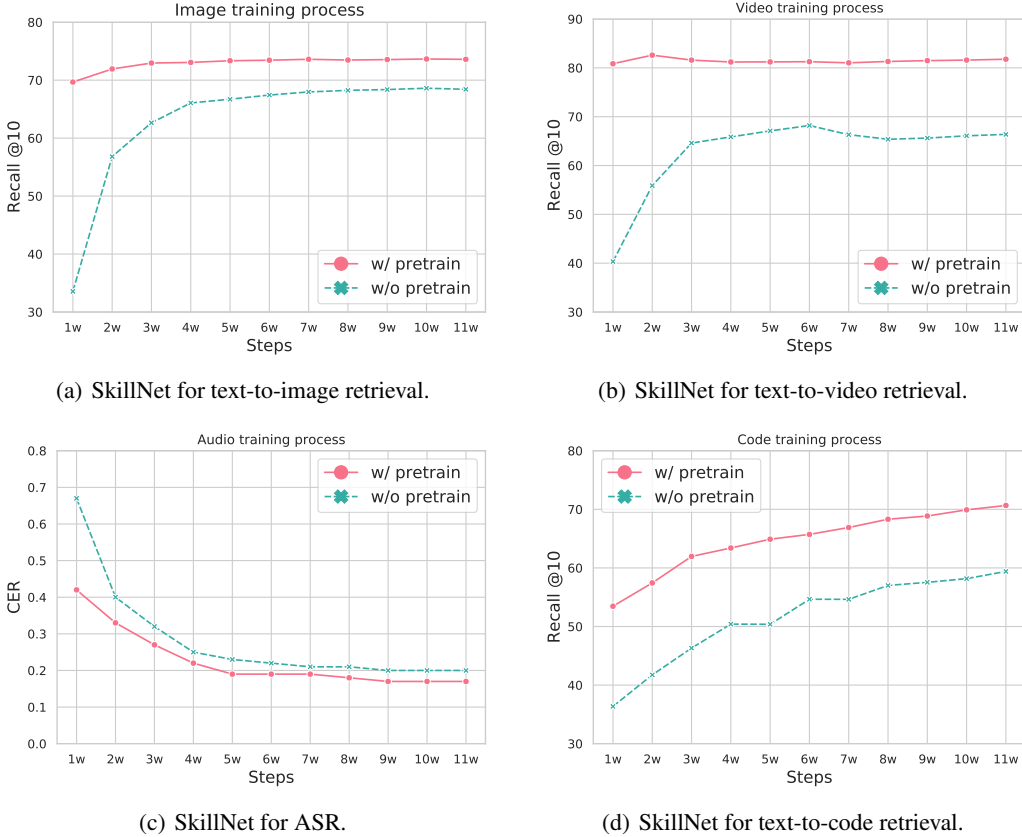


Figure 5: Performance of SkillNet with different finetuning steps. X-axis stands for the training steps. Y-axis stands for the evaluation metric (lower is better for CER).

Figure 6, 7 and 8 give case studies on image, video and code retrieval, respectively.








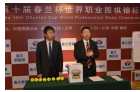
Query	Model Outputs			Ground Truth
湍急的河水里有一群穿着救生衣的人在划橡皮艇 (Trans: A group of people in life jackets are rowing a rubber dinghy in a fast river)	 score = 34.91	 score = 30.88	 score = 23.377	
一个戴着墨镜的男人牵着一个穿着白色裙子的人走在道路上 (Trans: A man in sunglasses walks down the road holding a woman's hand in a white dress)	 score = 33.03	 score = 30.69	 score = 29.40	
一个背着包的女人走在人来人往的街道上 (Trans: A woman with a bag is walking on a busy street)	 score = 32.48	 score = 30.90	 score = 29.65	
一个双臂抬起的运动员跪在绿茵茵的球场上 (Trans: An athlete with raised arms kneels on a green field)	 score = 32.60	 score = 28.20	 score = 28.01	
展板前的桌子前一个戴着眼镜的男人旁有一个双手相握的男人在讲话 (Trans: A man with clasped hands is speaking next to a man with glasses at a table in front of a display board)	 score = 34.04	 score = 29.50	 score = 28.29	

Figure 6: Case study for text-to-image retrieval. For each query, we show top-3 returned images and the relevance scores returned by SkillNet.



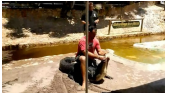















Query	Model Outputs		
一个穿红色衣服的男人坐在鳄鱼身上用手摸着它的嘴巴 (Trans: A man in red sits on a crocodile and touches its mouth with his hands)			 score = 29.63 (ground truth)
			 score = 25.66
			 score = 23.344
一个穿着红色衣服的人正在和人演戏 (Trans: A person in red is acting with others)			 score = 30.21 (ground truth)
			 score = 27.61
			 score = 26.39

Figure 7: Case study for text-to-video retrieval. For each query, we show the top-3 returned videos returned by SkillNet and provide three frames for each video.

## 6 Conclusion

This paper presents a sparsely activated multimodal multitask approach called SkillNet. We demonstrate the feasibility of using one model to achieve comparable performance compared to multiple modality-specific models. We further show that sparse pretraining gives a better initialized parameters which leads to improved accuracy, even better than modality-specific pretraining on three of five

Query	Model Outputs		
删除所有缓存的图像 (Trans: delete all cached images)	<pre>def clear(self):     self._cache.clear()     self._currentsize = 0</pre> <p>score = 138.16</p>	<pre>def clear_cache(grip_class=None)     if grip_class is None:         grip_class = Grip     grip_class(stdinReader()).clear_cache()</pre> <p>score = 108.37</p>	<pre>def clear(self):     self.display(Image.new(self.mode,         self.size))</pre> <p>score = 107.30</p>

Figure 8: Case study for text-to-code retrieval. For each query, we show top-3 returned codes and the relevance scores returned by SkillNet.

modalities. On Chinese text-to-image retrieval, our final system yields better accuracy with less activated parameters compared to existing leading systems. Our approach is modality-agnostic and task-agnostic. We leave the extension to larger number of modalities and tasks to the future.

## References

- [1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.
- [2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.
- [5] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.
- [6] Mehdi Bahrami, NC Shrikanth, Shade Ruangwan, Lei Liu, Yuji Mizobuchi, Masahiro Fukuyori, Wei-Peng Chen, Kazuki Munakata, and Tim Menzies. Pytorrent: A python library corpus for large-scale language models. *arXiv preprint arXiv:2110.01710*, 2021.
- [7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.
- [8] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [9] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5. IEEE, 2017.
- [10] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online, November 2020. Association for Computational Linguistics.
- [11] Jeff Dean. Introducing pathways: A next-generation ai architecture. In *Google Blog*, 2021. URL <https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture/>.
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

- [15] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. *arXiv preprint arXiv:2112.06905*, 2021.
- [16] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021.
- [17] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. Wenlan 2.0: Make ai imagine via a multimodal foundation model. *arXiv preprint arXiv:2110.14378*, 2021.
- [18] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. Codebert: A pre-trained model for programming and natural languages. *arXiv preprint arXiv:2002.08155*, 2020.
- [19] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. *arXiv preprint arXiv:2201.08377*, 2022.
- [20] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.
- [21] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and a foundation framework. *arXiv preprint arXiv:2202.06767*, 2022.
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.
- [24] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.
- [25] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.
- [26] Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. Transmart: A practical interactive machine translation system. *arXiv preprint arXiv:2105.13072*, 2021.
- [27] Dmitry Lepikhin, Hyoungho Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.
- [28] Junwei Liao, Duyu Tang, Fan Zhang, and Shuming Shi. Skillnet-nlg: General-purpose natural language generation with a sparsely activated approach, 2022. URL <https://arxiv.org/abs/2204.12184>.
- [29] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.
- [30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [32] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.
- [33] Christopher D Manning. Human language understanding & reasoning. *Daedalus*, 151(2):127–138, 2022.
- [34] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning-based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3): 1–40, 2021.
- [35] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

- [37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.
- [38] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- [39] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [41] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.
- [42] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2020.
- [43] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipai Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017.
- [44] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*, 2020.
- [45] W Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
- [46] Fan Zhang, Duyu Tang, Yong Dai, Cong Zhou, Shuangzhi Wu, and Shuming Shi. Skillnet-nlu: A sparsely activated model for general-purpose natural language understanding. *arXiv preprint arXiv:2203.03312*, 2022.

## A Appendix

**Configurations for Downstream Tasks** We describe the datasets and configurations for the downstream tasks as described in §4.1. For **text**, we use TNEWS [44], a benchmark dataset for Chinese text classification with 15 categories. It includes 53,300 sentences for training, 10,000 for development, and 10,000 for testing. Evaluation metric is accuracy. For **sound**, we adopt the dataset of AISHELL [9] for automatic speech recognition. It includes 170 hours of speech data in Mandarin. The evaluation metric is character error rate (CER), which means the percentage of characters that are incorrectly predicted (the lower the better). For **image**, we use AIC-ICC dataset [43], a benchmark dataset for text-to-image retrieval. It includes 210,000 image-text pairs for training and 30,000 for evaluation. We follow Wukong [21] and consider the first 10,000 images and 50,000 texts from the validation set as the final testing set. Evaluation metric is Recall @K (e.g., K = 1 and 10). For **video**, we carry out text-to-video retrieval on VATEX [41], which includes 25,991 videos for training and 3,000 for validation. Since some videos are unavailable for they are deleted or hidden by either YouTube or the users, we actually obtain 23,453 videos for training and 2,709 videos for validation. We randomly select 1,500 videos from validation set as our testing set and use the remaining videos as the development set. In the original dataset, there are 10 sentences in Chinese and 10 sentences in English to describe each video. In this work, we only utilize the Chinese captions. Similar to image retrieval, we use Recall @K (e.g., K = 1, 5 and 10) as the evaluation metrics. For **code**, since there is no publicly available dataset for Chinese language, we create a dataset by translating the PyTorch [6] dataset. It contains 218,814 Python package libraries from PyPI and Anaconda environment. We translate English docstrings to Chinese by a translation toolkit Transmart [26]. We delete duplicate code-text pairs and remove instances with low translation quality. We mix the original training set, development set and test set. At last, we shuffle the mixed set and randomly select 100,000/20,000/30,000 for training, validation and testing, respectively. The evaluation metric is Recall @K.

The model configurations are given as follows. For **text**, the max length is 512, and a special text padding token is padded if the input is shorter. For **sound**, we truncate each waveform to no more than 20ms, which leads to the max length of the sound embedding being 1,000. If the input is shorter, the remaining positions are filled with a special sound padding token. Other configurations can be found in §3.3. For **image**, same



with ViT-B/16 from CLIP [36], we first resize and normalize each image to  $224 \times 224$ . Then, we split each image into 196 patches with the patch size of  $16 \times 16$ , which are sent into a 3 in-channel and 768 out-channel 2D-convolution with kernel size of (16, 16) and stride step of (16, 16). For **video**, we truncate each video to no more than 10 seconds and transform each video into frames by 3 frames per second. Then, we randomly sample 6 frames for each video. At last, 6 video frames after cropping and normalizing are sent into a 3 in-channel and 768 output-channel 3D-convolution with a kernel size of (3, 16, 16) and stride step (3, 16, 16). For **code**, we set the whole max length as 512 with the limit of text max length of 64. We use AdamW [31] optimizer and linear scheduler with 1,000 warmup steps. There are different ways to initialize the model parameters. To accelerate the training process, instead of training from random initialization, we use ViT-B/16 from CLIP [36] to initialize image-related parameters and initialize other parameters from scratch. Since different modalities have different memory costs, we set the batch sizes as 512/1024/3072/1024/512 for text/sound/image/video/code to maximize the memory usage of GPUs. We observe that sound and code modalities require longer training steps to converge and the data scale of video is smaller than other modalities. Therefore, we sample instances for text/sound/image/video/code modalities with the ratio of 2/4/2/1/4. For each update, we only sample instances from one modality, which makes the learning process more stable. We update our model for 200,000 steps in total.

**Configurations for Pretraining Tasks** We describe the datasets and configurations for the pretraining tasks as described in §4.2. For **text**, we crawl a collection of raw Chinese texts containing Wikipedia, novels, news, lyrics and poems. We clean the data and finally obtain a dataset of about 300 gigabytes. For **sound**, we collect audio datasets from an open-source platform<sup>3</sup>, which includes about 1,200 hours of Chinese speech data. For **image**, we download the Wukong dataset [21] which originally includes 101,483,885 text-image pairs and filter out low-quality instances that with no Chinese words, too many illegal symbols and the length of captions is less than 4. We finally use about 84,000,000 text-image pairs for pretraining. For **video**, we use WebVid-2M [7], which comprises of over two million video-text pairs scraped from the internet. We translate the original English texts to Chinese by the translation tool Transmart and use the translated data for pretraining. For **code** pretraining, we hold out 800,000 code-text pairs from the aforementioned code dataset translated from PyTorrent, which have no overlaps with the datasets used for the downstream task of text-code retrieval.

We pretrain SkillNet using the AdamW [31] optimizer with the learning rate  $1e-5$  and a linear scheduler with 10,000 warmup steps. Same with the configuration for downstream tasks, we use one-modality data to pretrain our model in each update. The model is pretrained for 1,000,000 steps in total with batch sizes of 1024/512/8192/2048/512 for text/sound/image/video/code, respectively. Pretraining takes about 14 days on 64 A100 GPUs of 40GB memory size.

---

<sup>3</sup><https://blog.ailemon.net/2018/11/21/free-open-source-chinese-speech-datasets/>