Switch-BERT: Learning to Model Multimodal Interactions by Switching Attention and Input

Qingpei Guo[1], Kaisheng Yao[2], and Wei Chu[1]

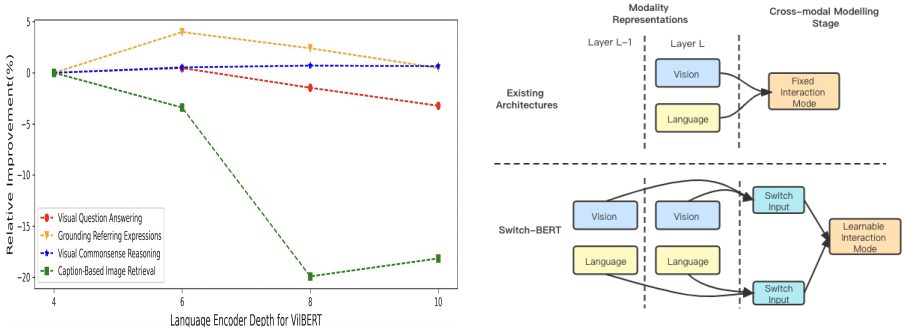
[1] Ant Group {qingpei.gqp, weichu.cw}@antgroup.com
[2] Amazon AWS AI kaishey@amazon.com

Abstract. The ability to model intra-modal and inter-modal interactions is fundamental in multimodal machine learning. The current state-of-the-art models usually adopt deep learning models with fixed structures. They can achieve exceptional performances on specific tasks, but face a particularly challenging problem of modality mismatch because of diversity of input modalities and their fixed structures. In this paper, we present **Switch-BERT** for joint vision and language representation learning to address this problem. Switch-BERT extends BERT architecture by introducing learnable layer-wise and cross-layer interactions. It learns to optimize attention from a set of attention modes representing these interactions. One specific property of the model is that it learns to attend outputs from various depths, therefore mitigates the modality mismatch problem. We present extensive experiments on visual question answering, image-text retrieval and referring expression comprehension experiments. Results confirm that, whereas alternative architectures including ViLBERT and UNITER may excel in particular tasks, Switch-BERT can consistently achieve better or comparable performances than the current state-of-the-art models in these tasks. Ablation studies indicate that the proposed model achieves superior performances due to its ability in learning task-specific multimodal interactions.

Keywords: multimodal interactions, cross-layer interaction, switch attention

1 Introduction

The current state-of-the-art approaches for multimodal machine learning [5, 18, 19, 22, 25, 32, 34] are based on the BERT encoders [6] that use the Transformer architecture [36]. These BERT-based models follow two design paradigms for intra-modal and inter-modal interactions. The first paradigm utilizes a single-stream BERT encoder to jointly encode representations from these modalities, such as those from vision and language [5, 18, 19, 22, 32]. In this case, intra-modal interactions and the implicit association between modalities are jointly modeled with the multi-head attention mechanism [36]. The second paradigm learns modal-specific representations through different BERT encoders, for instance using dual-stream BERT encoders on vision and language [25, 34]. These



(a) Illustration of modality misalignment in VilBERT(fixed structure).

(b) Comparison between popular multi-modal architectures with Switch-BERT.

Fig. 1: (a) The text and visual encoder are separate before their interactions in VilBERT [25]. By varying the depth of text encoder from 4 to 10, the accuracy of VilBERT is changed relative to that from depth 4. The optimal relative improvements in accuracy are different in the four tasks. For VilBERT, the misalignment can degrade performances by approximately 20% relatively. (b) In contrast to fixed structures, Switch-BERT learns to attend outputs from various depth and has learnable layer-wise and cross-layer interactions.

methods achieve inter-modal interactions via specially designed structures such as cross-attention sub-layers [25, 14, 40].

However, misalignment between modal semantics is a challenging problem for these methods. For example, the visual modality observation often is based on region-level semantic feature from detection models such as Faster R-CNN [29], whereas the text modality observation can be simply raw tokens or sub-word tokens such as word-pieces [37]. For single-stream models, these visual features with high-level semantics and text input with low-level semantics are both fed to the BERT encoder simultaneously. Given that these observations are not at the same semantic level, using a common encoding process for different modalities seems to be contradictory. The dual-stream models can ease the misalignment problem with distinct encoding process for each modality. However, the interaction between modalities of dual-stream models is restricted to specific layers that can be inflexible.

Fig. 1(a) illustrates this problem by tuning the depth of a BERT-based language encoder before interaction with the visual stream in ViLBERT [25] on a set of tasks. Though with deeper encoder that usually extracts higher level semantics [23, 35], performances don't reveal monotonous trend with the depth. This indicates that misalignment between modal semantics poses challenges to optimal multimodal performances. Another observation in Fig. 1(a) is that the optimal depths are different for these tasks, indicating that a fixed architecture is hardly optimal for every task. This suggests necessity for more flexible architectures. The modality misalignment problem is however not well studied.

In this paper, we propose Switch-BERT to alleviate the modality misalignment problem. As illustrated in Fig. 1(b), Switch-BERT extends the recently developed multimodal methods but has sample-specific interactions among modalities, instead of fixed architectures adopted in the previous approaches for every sample. Specifically, it introduces two modules, respectively for layer-wise switch operation in Switch-Attention Block (SAB) and cross-layer switch operation in Switch Input Block (SIB). The SAB module learns to attend to, given a sample, particular modality and choose from a set of predefined operations for interactions among modalities. The SIB module introduces sample-specific modeling of cross-layer modal representations and learns to switch inputs among representations at various depths.

We pre-train Switch-BERT on Conceptual Captions [30] to learn task independent visual and text grounding. Proxy pre-training tasks include masked language modeling with visual clues (MLM), masked region classification with KL-divergence (MRC-KL) [5] and Image-Text Matching (ITM). We evaluate Switch-BERT on three downstream tasks including visual question answering, cross-modal retrieval and referring expression comprehension, and perform experiments on VQAv2 [9], Flick30k [27] and RefCOCO+ [14] datasets. Experimental results show Switch-BERT can learn better multimodal representations, compared with previous single- and dual-stream models. We conduct ablation studies and show that Switch-BERT can learn task-specific multimodal interactions end-to-end, including layer-wise interaction selection and cross-layer input selection. This task-specificity is an advantage over other methods with fixed architectures.

2 Methodology

2.1 Preliminaries

Language BERT Encoder. BERT [6] was originally proposed for natural language processing tasks to learn semantic representations for each input token via a stack of transformers [36]. A BERT encoder consists of L transformer layers, in which representation X_l at l-th layer is obtained from the representation X_{l-1} in its lower layer as follows:

$$X_l = LN(\bar{X}_l + GeLU(\bar{X}_l W_1)W_2), \tag{1}$$

$$\bar{X}_l = LN(\hat{X}_l + X_{l-1}), \tag{2}$$

$$\hat{X}_l = MHA(Q_l, K_l, V_l) \tag{3}$$

where $MHA(\cdot)$ implements the multi-head attention mechanism [36], with query, key, and value at layer l each computed as $Q_l = X_{l-1}W^Q$, $K_l = X_{l-1}W^K$, and $V_l = X_{l-1}W^V$. LN is layer normalization [3], $GeLU$ [10] is the activation function of feed forward block. $W^Q, W^K \in R^{d\times d^q}$, $W^V \in R^{d\times d^i}$, and $W_1, W_2^\top \in R^{d\times d^f}$ are learnable matrices. The multi-head attention block and feed forward block form a transformer layer.

Multimodal BERT Encoder. Multimodal BERT [15, 25] extends the language BERT with multimodal input vector sequences. For instance, for tasks that consist of image and text, the model assigns two types of inputs: image can be a sequence of vectors as $X^i = [IMG, i_1, \cdots, i_{N_i-1}] \in R^{N_i \times d_i}$ and text can be $X^t = [CLS, w_1, \cdots, w_{N_t-2}, SEP] \in R^{N_t \times d_t}$, where *IMG*, *CLS* and *SEP* are embeddings of special markers. Usually, we have $d_i = d_t = d$. Typical approaches include UNITER [5], in which X^i and X^t are concatenated, forming a single stream of input $X_0 = [X^i X^t]$ to compute query, key and value matrices. In contrast, ViLBERT [25] computes query from one modality but key and value from other modality, and vice versa, forming dual steams of computations.

2.2 Generalizing BERT Encoder

We would like to generalize the encoder in Eqs. (1-3) beyond the multimodal architectures described above. To this end, we first use $X \in \{X^i, X^t\}$ to denote either the image modality observation X^i or text modality observation X^t. We use $\neg X$ to denote complementary of X; e.g., $\neg X = X^i$ if $X = X^t$. Notice that X^i and X^t are for purpose of notations, and can be generalized beyond image and text modalities.

We further generalize the multi-head attention mechanism in Eq. (3) beyond linear projections on input X_{l-1}, in which query, key, and value are obtained via certain transformations. Formally, we rewrite Eq. (3) as follows:

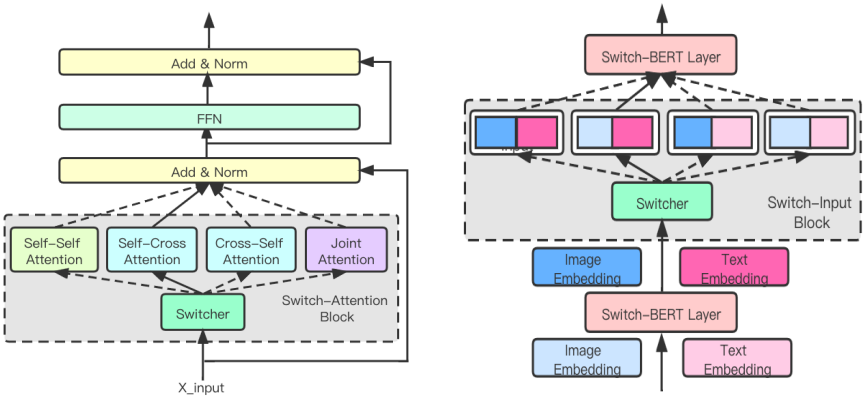
$$\hat{X}_l = MHA(q(X_{input}), k(X_{context}), v(X_{context})), \tag{4}$$

where $q(\cdot)$, $k(\cdot)$ and $v(\cdot)$ extract query, key and value representations, respectively. Notice that key and value operations share the input observation $X_{context}$, whereas query $q(\cdot)$ operates on X_{input}.

Eq. (4) enables us to relate the previously proposed multimodal approaches. For dual-stream models [25, 34], intra-modal and inter-modal interactions are independently modeled explicitly. Using Eq. (4), Self-Attention for intra-modal interaction is modeled with $X_{input} = X_{l-1}$ and $X_{context} = X_{l-1}$. Cross-Attention for inter-modal interaction can be achieved using $X_{input} = X_{l-1}$ and $X_{context} = \neg X_{l-1}$. For single-stream models [18, 5, 19], intra-modal and inter-modal interactions are implicitly modeled with Joint-Attention using $X_{input} = X_{l-1}$ and $X_{context} = [X_{l-1}, \neg X_{l-1}]$, the latter is obtained via concatenation and enables attention to the whole multimodal context.

Table 1: Multimodal interaction mode spaces

Interaction Mode	Attention Mechanisms
M_0: Self-Self Attention	X & $\neg X$: Self-Attention
M_1: Self-Cross Attention	X: Self-Attention, $\neg X$: Cross-Attention
M_2: Cross-Self Attention	X: Cross-Attention, $\neg X$: Self-Attention
M_3: Joint-Attention	X & $\neg X$: Joint-Attention



(a) Illustration of Switch-BERT layer and Switch-Attention Block.

(b) Illustration of Switch-Input Block.

Fig. 2: (a) The Switch-BERT layer extends the Multi-head Joint Attention block in a normal transformer encoder layer with our proposed Switch-Attention Block. (b) The Switch-Input Block brings in modality representations from current and previous layers for its successive Switch-BERT layer. (best viewed in color)

Among the above described attention mechanisms, Joint-Attention uses whole multimodal context, therefore has potential of representation of both Self-Attention and Cross-Attention. However, its multimodal context can face potential semantic misalignment between modalities, as described in Sec. 1. On the other hand, Self-Attention and Cross-Attention restrict the modal context to attend, easing semantic misalignment of modalities, but leads to limited representation to particular modality.

We therefore design a more complete space of multimodal interactions. Table 1 lists four interaction modes between X and its complementary $\urcorner X$. Self-Self Attention invokes self-attention on each modality. Self-Cross Attention has X use Self-Attention and $\urcorner X$ use Cross-Attention, and vice versa for Cross-Self Attention. Joint-Attention has both X and $\urcorner X$ conduct their own Joint Attention operations. Those attention operators share the same attention weights in our setting and can be implemented with different layer-wise attention masks.

2.3 Switch Attention and Input Block

Switch-Attention Block. Unlike conventional multimodal models that limit the modality interaction between specific layers, we employ the Switch-Attention Block (SAB) to achieve learning layer-wise multimodal interaction in an end-to-end manner. As illustrated in Fig. 2 (a), SAB depends on an attention switcher module to search for an appropriate mode from the multimodal interactions de-

scribed in Table 1. The search space can be formally defined as a set of operations $\{M_n\}_{n=1}^{N_a}$, where N_a indicates the number of interaction modes.

We describe in the following a switcher method to search for proper interaction. Given a holistic representation of image and text as X_l^i and X_l^t, we apply average pooling over the image and text tokens to obtain global features of each modality:

$$z_l^i = AvgPool(X_l^i), z_l^t = AvgPool(X_l^t))). \tag{5}$$

Then we define the modality "alignment degree" of the l-th layer as $d_l = z_l^i \odot z_l^t$, and apply a trainable MLP, f_{MLP} with Softmax activation to obtain the probability of the interaction modes π:

$$\pi = Softmax(f_{MLP}(d_l)). \tag{6}$$

We use Gumbel-Softmax reparameterization [12] to sample a particular interaction based on the above probability, in which probability of interaction M_n is

$$p(M_n) = \frac{\exp\left((\log\left(\pi_n\right) + g_n)/\tau\right)}{\sum_{j=1}^{N_a} \exp\left((\log(\pi_j) + g_j)/\tau\right)}, \tag{7}$$

where g_n is sampled Gumbel noise, computed as $g_n = -\log(-\log(u_n))$, with u_n sampled from uniform distribution of $Uniform(0, 1)$. τ is the smooth parameter for Gumbel-Softmax distribution.

Given $X_{input} = X_l^i \cup X_l^t$, SAB performs "soft weighting" or "hard selection" of interaction modes by:

$$\begin{aligned} y_{soft} &= \sum_{i \in N_a} p(M_i) M_i(X_{input}) \\ y_{hard} &= M_{n^\star}(X_{input}), n^\star = \underset{n}{\operatorname{argmax}}\{p(M_n)\} \end{aligned} \tag{8}$$

.

For training, we start at a high temperature in Eq.7 for small gradient variance, then anneal to a small but non-zero temperature to make the output distribution $p(M)$ approximate one-hot. We adopt "soft weighting" of attention modes during training and "hard selection" for inference.

Switch-Input Block. To ease semantic misalignment between modalities, we propose Switch-Input Block (SIB) to bring in cross-layer modal representation. SIB enables Switch-BERT layer, illustrated in Fig. 2 (a), to take input either from the output of its lower layer or from the residual connection in the lower layer, which connects to the output from the layer further below. Concretely, for l-th layer with $l \geq 2$, its input is in a set of $\{X_{l-1}\urcorner X_{l-1}\}$ $\cup\{X_{l-1}\urcorner X_{l-2}, X_{l-2}\urcorner X_{l-1}, X_{l-2}\urcorner X_{l-2}\}$. We then apply switch operation on the set and obtain an element from the set as input X_{input} to layer l. The switcher algorithm follows Eq. 6, Eq.7 and Eq. 8 but is trained specifically for SIB. Fig. 2 (b) illustrates the Switch-Input Block.

2.4 The Switch-BERT Model

The Switch-BERT model's components are described in details below. Further details are in the supplementary material.

Visual and Text Embedding. Following [25], images are represented with detected objects. We extract the bounding box and visual feature of each object from the widely used Faster-RCNN [29] detector trained on Visual Genome [16]. We also add a type field (VisualType/TokenType) to distinguish visual and text input. The region feature, position and type field are fed into a visual embedding layer to obtain the visual embedding for Switch-Encoder. A special IMG token representing the entire image segment is also inserted at the beginning of the visual sequence. The text embedding is generated following BERT [6], in which we tokenize the input sentence and keep orders of tokens as their position ids. The token, position and type field are fed into a text embedding layer to perform embedding lookup.

Switch-Encoder. Given the pair of visual and text embedding, the Switch-Encoder learns to model layer-wise multimodal interactions. The Switch-Encoder consists of a stack of Switch-BERT layers, with Switch-Input Block inserted between consecutive Switch-BERT layers. Switch-BERT layer in Fig. 2 (a) generally follows the architecture of the Transformer encoder layer [36], but distinguishes it with the adaptive multimodal attention mechanism using Switch-Attention Block. It takes the entire representations from visual and text embedding, but selects sample-specific interactions of these representations. The Switch-Input Block routes the modality input for the following Switch-BERT layer to help alleviate semantic misalignment. The rest of the Switch-Encoder proceeds similarly as that in BERT encoder, resulting in a multimodal feature as its output.

Pretraining Tasks. Task-agnostic multimodal pre-training can help learn associations between modalities. Like previous work [25, 19, 21, 26, 33, 41, 22], we first pre-train Switch-BERT on proxy tasks and then adapt it to downstream tasks through finetuning. Three proxy tasks are used for pre-training. (1) Masked language modeling with visual clues (MLM). This task follows the MLM objective in BERT [6] but with the above described contextualized multimodal input. In this task, word tokens are randomly masked but with their positions preserved. The model needs to predict the token from the left visual and textual context. (2) Masked region classification with KL-divergence (MRC-KL) [5]. Similar to MLM, this task masks approximately 15% of the region features. MRC-KL then trains the model to predict the class distribution from the object detector for the region, rather than reconstructing the feature of masked regions. (3) Image-Text matching (ITM). Given paired image-and-text as positives, their negative pairs are generated by randomly replacing texts in the positive pairs with unrelated ones. The ITM task is for the model to distinguish positive pairs from negatives.

Table 2: Statistics of Datasets for the Downstream Tasks

Dataset	Tasks	Train	Test	Metric
Flick30k	Image-Text Retrieval	29k	1k	Recall@k
RefCOCO+	Referring Expression	120k	10.6k	Accuracy
VQAv2	Visual Question Answering	657k	107.3k	VQA-score

3 Experiments

3.1 Datasets and Downstream Tasks

We evaluate Switch-BERT on different types of downstream tasks including image-text retrieval, referring expressions and vocab-based VQA. Their statistics are shown in Table 2.

Image-Text Retrieval. Given images or captions, the image-text retrieval task requires the model to perform cross-modal retrieval. We conduct experiments on Flick30k [27] dataset, which has images paired with five captions. Following [25], we train models on Flick30k in a 4-way multiple-choice setting. For each image-text pair, three negatives are generated by replacing the caption with a random one and replacing the image with a random and a hard one. The model outputs similarity scores of these four image-text pairs as the ITM task. Once softmax is computed on the similarity scores, cross-entropy loss is applied to learn the models. We report Recall@1.

Referring Expressions Comprehension. This task focuses on localizing objects queried by a natural language expression. For the RefCOCO+ [14] dataset, we take the bounding boxes detected by [40] and select the top 36 regions with the highest class scores. Following the conventions in [32, 25], a simple fully-connected layer is added on top to regress the matching degree, defined as the IOU with the ground truth box, with the referring expression for each input region. We train the model with binary cross-entropy loss. To evaluate, regions with matching degree above threshold of 0.5 are considered correct. We apply the accuracy score as the evaluation metric.

Visual Question Answering. Given questions about an image, this task expects the model to give correct answers. Following [1], we consider the VQA [9] task a multi-label classification problem on a closed answer pool and generate the target soft-label based on its relevance to ten human answer responses. We add two fully-connected layers to map the multimodal representation, which is the element-wise product fusion of image and text representation, to the answers' space and apply binary cross-entropy loss for training. Following the same

protocol with SOTA baselines, we train models on train-val split and report VQA-score [2] on the test-dev split.

3.2 Controlled Settings

Shown in [25, 34, 5, 28], the quality and volume of the pre-training data significantly impact the performance of multimodal BERTs. This explain most of the claimed performance differences in downstream tasks [4]. In this paper, we focus our discussion on the independent contribution of architecture design. To exclude performance influences other than architectures and enable fair comparison under limited resources, we adopt the controlled settings introduced by [4]. Specifically, we pre-train multimodal BERTs on the same subset of 2.7M image-text pairs of Conceptual Captions [30] for 10 epochs and employ the same proxy tasks as our Switch-BERT model. We use the VOLTA [1] implementation for all state-of-the-art models for comparison in our experiments, and train these multimodal BERTs with a fixed set of hyperparameters, such as encoder dimensions, methods for modality fusion, number of MLP layers in the finetune head, to exclude possible confounds that may interfere with a fair comparison of these architectures. Models with the best validation set performance are chosen for downstream tasks evaluation [2]. Due to space constraints, more implementation details as well as hyper-parameter settings are split into the supplementary materials.

3.3 Main Results

We compare the proposed Switch-BERT against existing multimodal architectures of both single and dual-stream on three widely-used benchmark datasets. Baselines for comparison include the state-of-the-art multimodal architectures of ViLBERT [25], UNITER [5], VisualBERT [19], VL-BERT [32] and LXMERT [34]. These baselines and Switch-BERT follow the pre-train-then-fine-tune procedure with the controlled settings described above and have the same context for comparison.

Table 3 presents the experimental results of the model, together with results from these baselines. We observe that Switch-BERT has performances that are on par or better than the previous state-of-the-art architectures in these downstream tasks. The absolute improvements of 0.9% on RefCOCO+, 1.8% on VQAv2 and 1.1% on Flick30K Image Retrieval over previous SOTA [3] indicating that Switch-BERT can learn better vision and language representations that generalize better than these alternative methods to the downstream tasks. The controlled settings ensure the improvements are mainly contributed from the

[1] https://github.com/e-bug/volta

[2] We train with three different random seeds and report their average performances

[3] For overall SOTA numbers that can be achieved without the controlled settings, readers can refer to [39] for VQAv2 and Flick30K Retrieval datasets, and [13] for RefCOCO+.

Table 3: Results on downstream tasks. We adopt the re-implementation from the VOLTA[4] framework for baseline models. All models perform the same controlled settings and "*" denotes models without pre-training on Conceptual Captions[30]. We report std of Switch-BERT as well as baseline models on three runs with different random seeds.

Models		Params	VQAv2	Flick30K-Retrieval		RefCOCO+
				Image Retrieval	Text Retrieval	
Single-stream (Fixed)	UNITER[5]	114.9M	68.8±0.4	60.9±0.7	76.4±1.3	71.9±0.67
	VL-BERT[32]	116.1M	68.3±0.31	57.9±1.1	70.9±1.7	71.1±0.23
	VisualBERT[19]	114.9M	68.9±0.27	61.1±1.2	75.5±1.8	69.7±0.31
Dual-stream (Fixed)	LXMERT[34]	211.4M	67.1±0.34	58.6±1.4	74.9±2.7	69.8±0.46
	VilBERT[25]	242.1M	68.7±0.82	59.8±0.8	**78.3**±1.6	70.8±0.58
Dynamic	Switch-BERT*	130.6M	66.7 ±0.97	38.2 ±1.7	57.3 ±2.3	68.9 ±0.82
	Switch-BERT		**70.7**±0.62	**62.2**±0.9	78.2±1.6	**72.8** ±0.45

proposed architectures of Switch-Attention and Switch-Input blocks, which aim at easing the semantic misalignment between modalities and learning image-text modality interactions. Table 3 also includes the results of Switch-BERT without pre-training on Conceptual Captions dataset, i.e., initialized only from BERT in [6]. The degradation in performance demonstrates that the Switch-BERT benefits from pretraining as other multimodal BERTs.

3.4 Ablation Studies

Effectiveness of the Switch-Attention and Switch-Input Blocks. We start by investigating the influences of Switch-Attention and Switch-Input blocks. Following our controlled settings, we compare Switch-BERT with its three variants on downstream tasks. (i) SIB-ONLY: this variant uses normal encoder-style transformer layers instead of the Switch-BERT layer, (ii) SAB-ONLY: in this variant, we fix the input to each Switch-BERT layer to the output from its lower layer as usual. (iii) No-SIB-SAB: this variant is a normal single stream BERT encoder. All variants are evaluated following the pre-train-then-fine-tune procedure, and share the same hyperparameter setting with Switch-BERT.

Results in Table 5 clearly show better performances by Switch-BERT than its variants. Given that SIB brings cross-layer input, we conclude that the semantic-level misalignment exists in single stream models and reducing misalignment between semantics of modalities results in better representations. The improvements of SAB-ONLY over the No-SIB-SAB variant also hint that our switching attention mechanism that learns to model modality associations is superior to the widely used single Joint-Attention mechanism. This bring us to the second question: Is the Joint-Attention necessary for Switch-Attention block?.

Table 4: An ablation study of interaction modes. Cross-Self (Self-Cross) and Joint stand for interaction modes. Pretraining indicates whether the models are pre-trained on Conceptual Captions before adapt to RefCOCO+. The default interaction mode is Self-Self Attention for all tested models, which means no interactions between modalities.

Model			RefCOCO+
Pretraining	Cross-Self & Self-Cross	Joint	Accuracy
✓	✓		71.5
✓		✓	71.2
✓	✓	✓	**72.8**
	✓		68.3
		✓	67.5
	✓	✓	**68.9**

Table 5: An ablation study of Switch-Attention and Switch-Input blocks.

Model		Flick30K		VQAv2	RefCOCO+
SIB	SAB	IR(r@1)	TR(r@1)	VQA-score	Accuracy
		60.7	76.2	67.8	69.5
	✓	61.7	76.9	68.9	72.4
✓		60.8	77.7	68.5	71.7
✓	✓	**62.2**	**78.2**	**70.7**	**72.8**

Table 6: An ablation study on effect of models' depth.

Model	VQAv2	RefCOCO+
#layers	6 → 12	6 → 12
UNITER	64.2 → 68.8	69.7 → 71.9
Switch-BERT	**65.4 → 70.7**	**70.2 → 72.8**
SAB-ONLY	65.0 → 68.9	69.4 → 72.4

Necessity of Joint-Attention. We perform experiments on the RefCOCO+ dataset to verify the necessity for Joint-Attention. Table 4 shows the results of Switch-BERT and its variants of the attention mode space with different initialization in the upper and lower panel. Models with the Cross-Self and Self-Cross Attention show similar results (71.5 vs 71.2) to those with Joint-Attention when pre-trained. However, even with Cross-Self & Self-Cross, using Joint-Attention with negligible additional parameters consistently outperforms those without using it. Therefore, results support the necessity of Joint-Attention.

Effect of Model's Depth. We also compare transferred results from models of varying depths including Switch-BERT and UNITER. Since Switch-BERT's SIB block introduces cross-layer connections given to more sensitivity to the model's depth, we also add the SAB-ONLY variant to the comparison. As shown in Table 6, Switch-BERT of various depth show superior performance compared to its counterparts UNITER baseline. In addition, we observe meaningful improvements of Switch-BERT on the SAB-ONLY variant of fewer layers across multiple tasks evaluated, proving SIB help adapt to different tasks regardless of the model's depth.

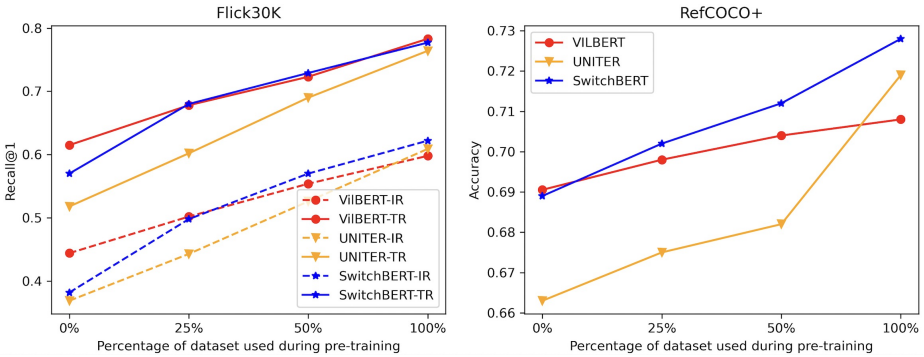


Fig. 3: Effects on scale of pre-training sets. *-IR and *-TR represents the image-to-text retrieval and text-to-image retrieval tasks, respectively. We find large performance drop with less pre-training data for UNITER – implying single-stream models with only Joint-Attention "eagers" for larger pre-training data before fully-trained.

Table 7: Computation overhead(FLOPs) and performances. Top-K routes are activated in Switch-Attention Blocks during fine-tuning.

Models	VQAv2		RefCOCO+	
	FLOPs	VQA-score	FLOPs	Accuracy
UNITER	$2.31*1e^{16}$	68.8	$3.68*1e^{15}$	71.9
VilBERT	$2.72*1e^{16}$	68.7	$4.29*1e^{15}$	70.8
Switch-BERT(K=4)	$8.02*1e^{16}$	70.7	$10.57*1e^{15}$	72.8
Switch-BERT(K=2)	$3.07*1e^{16}$	70.2	$5.27*1e^{15}$	72.1
Switch-BERT(K=1)	$1.97*1e^{16}$	68.2	$3.12*1e^{15}$	70.8

Impact on scale of pre-training sets. We now turn our attention to the effect of pre-training dataset's scale on Switch-BERT's performance. For this experiment, we take random subsets of 25% and 50% from our conceptual caption dataset to pre-train models and then adapt them to various downstream tasks under our predefined controlled settings. Shown in Fig. 3, we can see that Switch-BERT benefits from increasing amounts of data as well as UNITER and VilBERT. Another observation is that larger performance gaps emerge between UNITER and VilBERT with less pre-training data on both evaluated tasks, we conjecture that UNITER(single-stream models) with only Joint-Attention eagers for larger pre-training data volumes to get fully-trained, and Switch-BERT alleviates this problem with complete interaction mode space.

Computation overhead of Switch-BERT. We estimate the number of floating point operations of training a model on each downstream task for static approaches. For Switch-BERT, we track its routing path and accumulate the operation count during training due to its dynamic characteristics. Results are

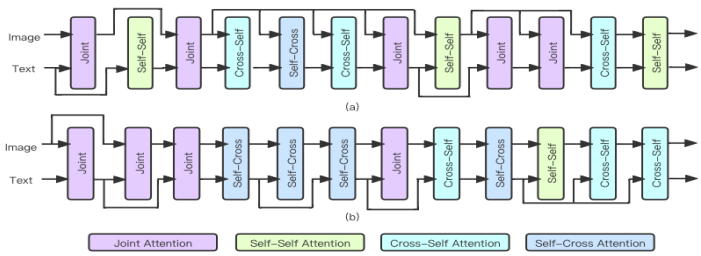


Fig. 4: Architectures Learned by Switch-BERT. (a) and (b) shows the learned architectures for referring expression comprehension and cross-modal retrieval tasks, respectively.

shown in Table 7. Switch-BERT indeed requires extra computation to converge compared to traditional static models, the overhead is mainly caused by SABs that activate all paths at the beginning of training. We further investigate this with only top-K paths in SAB activated, and observe an acceptable overhead and performance balance when K=2.

3.5 Qualitative Studies.

With SAB and SIB modules, Switch-BERT should be able to adapt its architecture to different multimodal tasks. To confirm this, we analyze utility of SAB and SIB on Switch-BERT fine-tuned on referring expression comprehension and cross-modal retrieval tasks. We sort the learned architectures according to their occurrence frequency on each task. Fig. 4 illustrates the most frequently used architectures by Switch-BERT on the two tasks. For the referring expression comprehension task, Switch-BERT learns to use cross-layer representations more frequently for the visual modality than the text modality. On the cross-modal retrieval task, Switch-BERT uses Self-Self attention once but more frequently with other attention modes that involve interactions between modalities. The frequencies of selecting these most-frequent architectures are dominantly 48.79% and 31.09%, respectively, on the two tasks. These results indicate that Switch-BERT is able to extract task-specific architecture.

4 Related Work

Multimodal BERTs. BERT-style representations [6, 24, 38, 17, 7] have been advancing the state-of-the-art performances in natural language processing in recent years. Its success has encouraged researchers to apply them more widely to tasks including multimodality. Methods based on BERT architectures have been proposed recently and have become the dominant approaches in applications such as video captioning [33]. The works of VisualBERT [19], UNITER [5], VL-BERT [32], and PixelBERT [11] employ a single-stream BERT encoder for joint modeling of interactions between modalities. The other dual-stream approach

including ViLBERT [25] and LXMERT [34] has representations separately for each modality and uses cross-attention mechanism to model their interactions. The proposed method distinguishes from the above methods in using flexible Switch Attention-and-Input mechanism to select proper interaction modes and cross-layer input. It aims at alleviating the not-well-studied semantic misalignment problem. Empirically, we have confirmed its superior performances over the other methods.

Conditional Computation Models. The proposed Switch-BERT dynamically adjusts its architecture according to inputs. It is therefore in line with Mixture of Experts (MoE) methods in [31, 8]. The method in [31] uses a gating function to select experts to perform computations. The method in [8] introduces the MoE layer into the Transformer architecture and applies a routing algorithm that sends tokens to their token-specific experts. Switch-BERT differs from these works in two aspects: i) instead of using MoE as a substitute of the FFN layers in [31], it selects sample-specific attention and input with the novel Switch Attention-and-Input blocks; ii) whereas MoE is conducted at token-level in [8], Switch-BERT conducts switch operations at modality-level and cross-layer. Besides, our switch input mechanism learns to "select" or "skip" a transformer layer, which shares the same spirit with variable depth in Transformer[20]. Work in [20] explores using a shared deep Transformer for multiple tasks with the learned distribution of layer selection, the learned distribution is restricted on task-level. While for Switch-BERT, the layer selection distribution is conditioned on modality inputs, such that it performs sample-specific switch operations. To our best knowledge, Switch-BERT is the first attempt to have conditional computation for multimodal learning.

5 Conclusion

In this paper, we proposed Switch-BERT to effectively alleviate the modality misalignment problem for multimodal representation learning. Switch-BERT learns to model intra- and inter-modal interactions and select interaction mode for each layer individually. It also learns to select, for each layer, the inputs that are not restricted to the current layer and therefore learns selecting inputs cross layers. We verified its effectiveness through controlled settings on multimodal tasks including visual question answering, cross-modal retrieval, and referring expression comprehension. We also carried out ablation studies to confirm that Switch-BERT is capable of learning task-specific architectures. Experimental results show that Switch-BERT dynamically adapts its structure and consistently achieve better or comparable performances than other state-of-the-art fixed architectures on a variety of multimodal tasks. In future work, we plan to explore the efficiency of variant mechanisms and reveal the internal alignment with more details.

Acknowledgments. The authors would like to thank the anonymous reviewers for their helpful feedback that improved this work.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
4. Bugliarello, E., Cotterell, R., Okazaki, N., Elliott, D.: Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. Transactions of the Association for Computational Linguistics **9**, 978–994 (2021)
5. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European Conference on Computer Vision. pp. 104–120. Springer (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: The North American Chapter of the Association for Computational Linguistics (2019)
7. Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., Hon, H.W.: Unified language model pre-training for natural language understanding and generation. Advances in Neural Information Processing Systems **32** (2019)
8. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022), http://jmlr.org/papers/v23/21-0998.html
9. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)
10. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
11. Huang, Z., Zeng, Z., Liu, B., Fu, D., Fu, J.: Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849 (2020)
12. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: International Conference on Learning Representations (2017), https://arxiv.org/abs/1611.01144
13. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021)
14. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014)
15. Kiela, D., Bhooshan, S., Firooz, H., Testuggine, D.: Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950 (2019)
16. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)

17. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
18. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11336–11344 (2020)
19. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
20. Li, X., Stickland, A.C., Tang, Y., Kong, X.: Deep transformers with latent depth. Conference and Workshop on Neural Information Processing Systems, NeurIPS (2020)
21. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision. pp. 121–137. Springer (2020)
22. Lin, J., Yang, A., Zhang, Y., Liu, J., Zhou, J., Yang, H.: Interbert: Vision-and-language interaction for multi-modal pretraining. arXiv preprint arXiv:2003.13198 (2020)
23. Lin, Y., Tan, Y.C., Frank, R.: Open sesame: Getting inside bert's linguistic knowledge. Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (2019)
24. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
25. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
26. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889 (2020)
27. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
28. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966 (2020)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
30. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
31. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
32. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)

33. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7464–7473 (2019)
34. Tan, H., Bansal, M.: LXMERT: Learning cross-modality encoder representations from transformers. Empirical Methods in Natural Language Processing pp. 5100–5111 (2019)
35. Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4593–4601. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1452, https://aclanthology.org/P19-1452
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
37. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., *et al*: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144 (2016)
38. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019)
39. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
40. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1307–1315 (2018)
41. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13041–13049 (2020)