BERT原理
以及在多模态中的应用

赵佳宝

May 25, 2020

華東師範大學
EAST CHINA NORMAL UNIVERSITY

目录

- Bert

Bert结构；预训练方法；应用；

- Multimodal Bert

➢ 单流模型：VisualBERT，Unicoder-VL，VL-BERT(ICLR2020)

➢ 双流模型：ViLBERT(NIPS2019)，LXMERT(EMNLP2019)

BERT Bidirectional Encoder Representation from Transformers

BERT是双向Transformer的Encoder。模型的主要创新点在pre-train方法上，即用了Masked LM和Next Sentence Prediction两种方法分别捕捉词语和句子级别的representation。

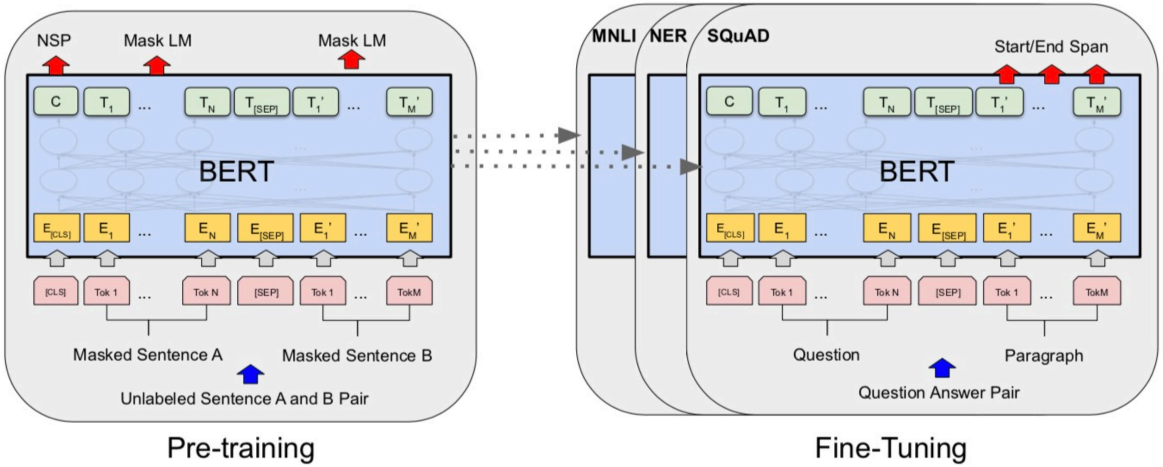


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

Transformer Attention is all you need

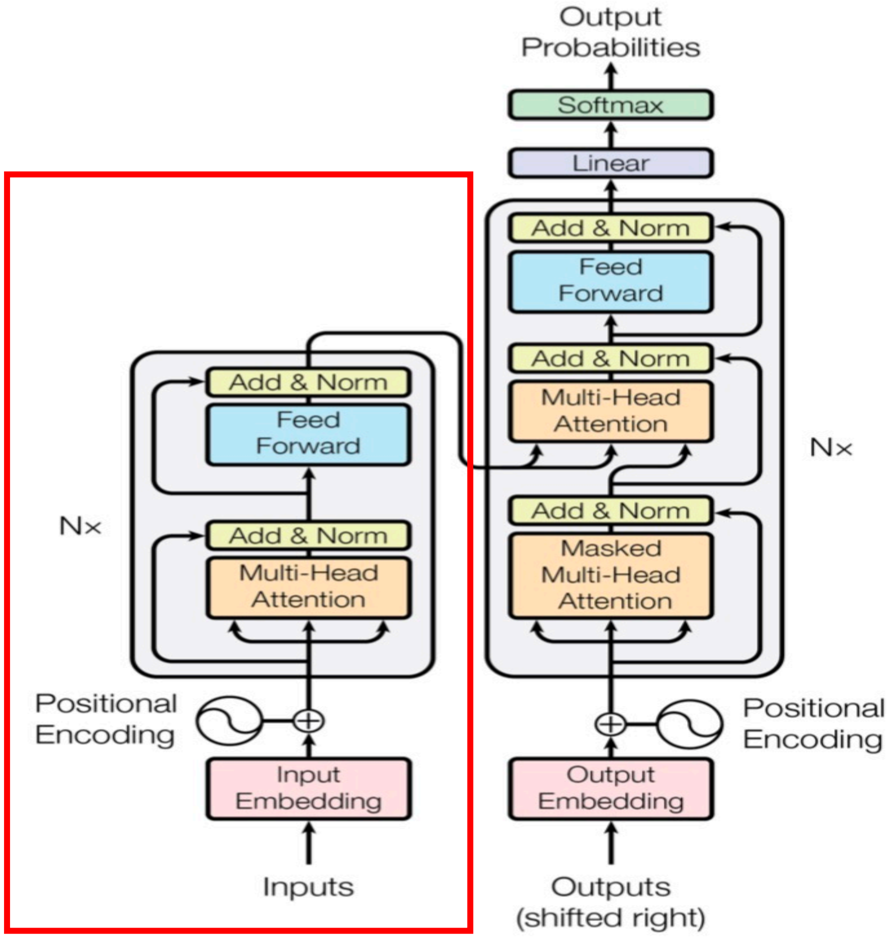


Figure 1: The Transformer - model architecture.

- Self-attention
- Multi-Head Self-attention
- Positional Encoding
- Layer Normalization

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

Self-Attention

Attention is all you need

q: query

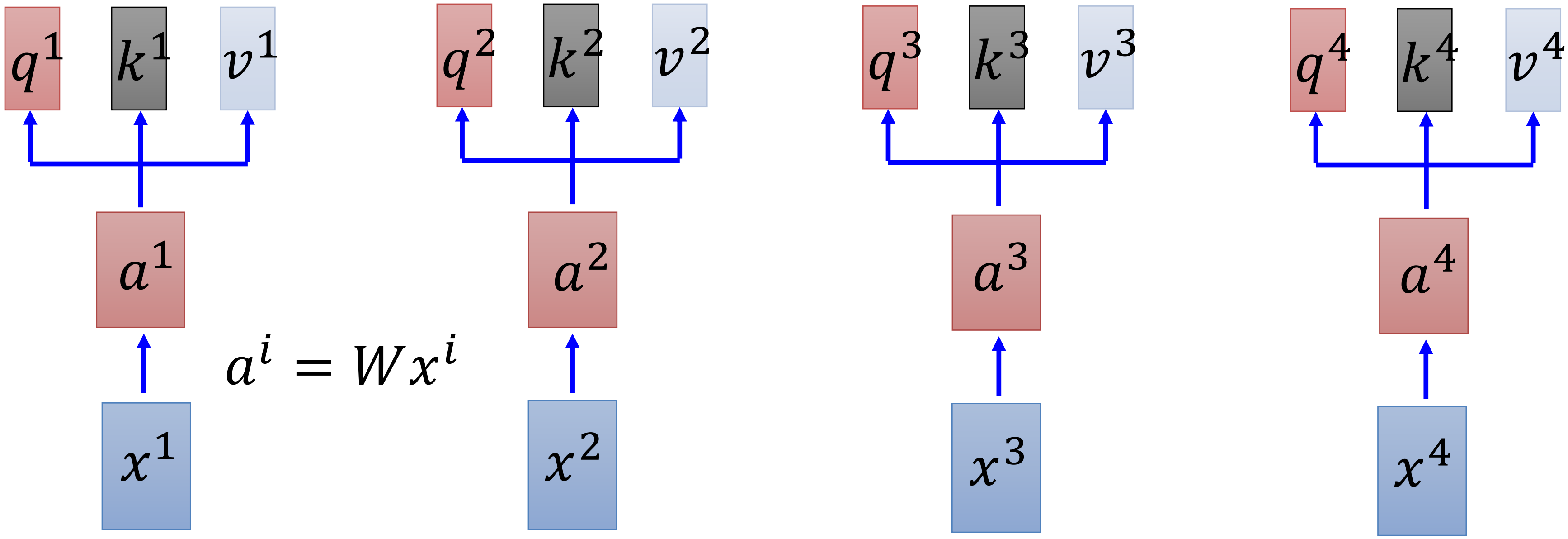
$q^i = W^q a^i$

k: key

$k^i = W^k a^i$

v: information to be extracted

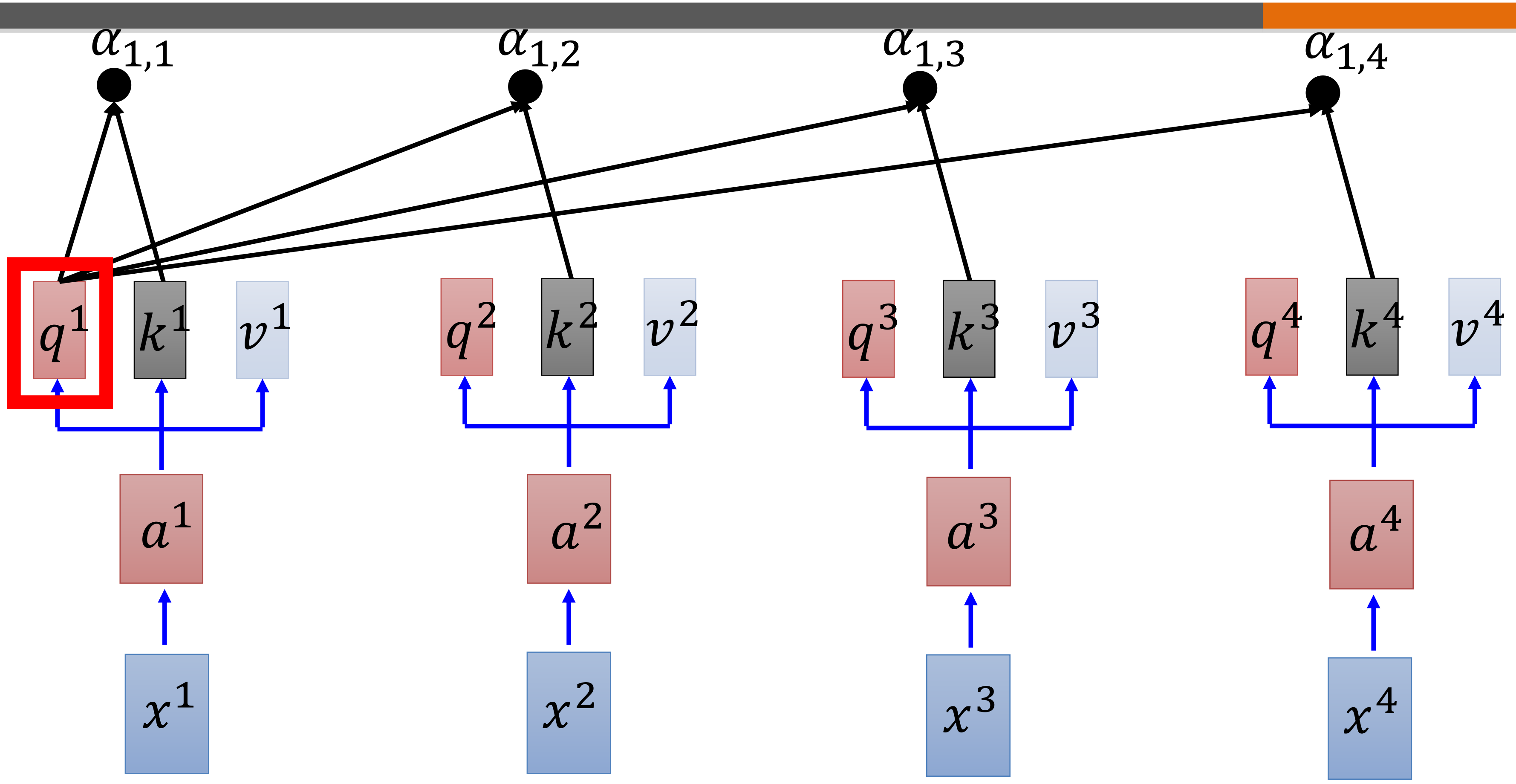
$v^i = W^v a^i$



$a^i = Wx^i$

Self-Attention

Attention is all you need



用每个query q对每个key k做attention

$$\alpha_{1,i} = q^1 \cdot k^i / \sqrt{d}$$

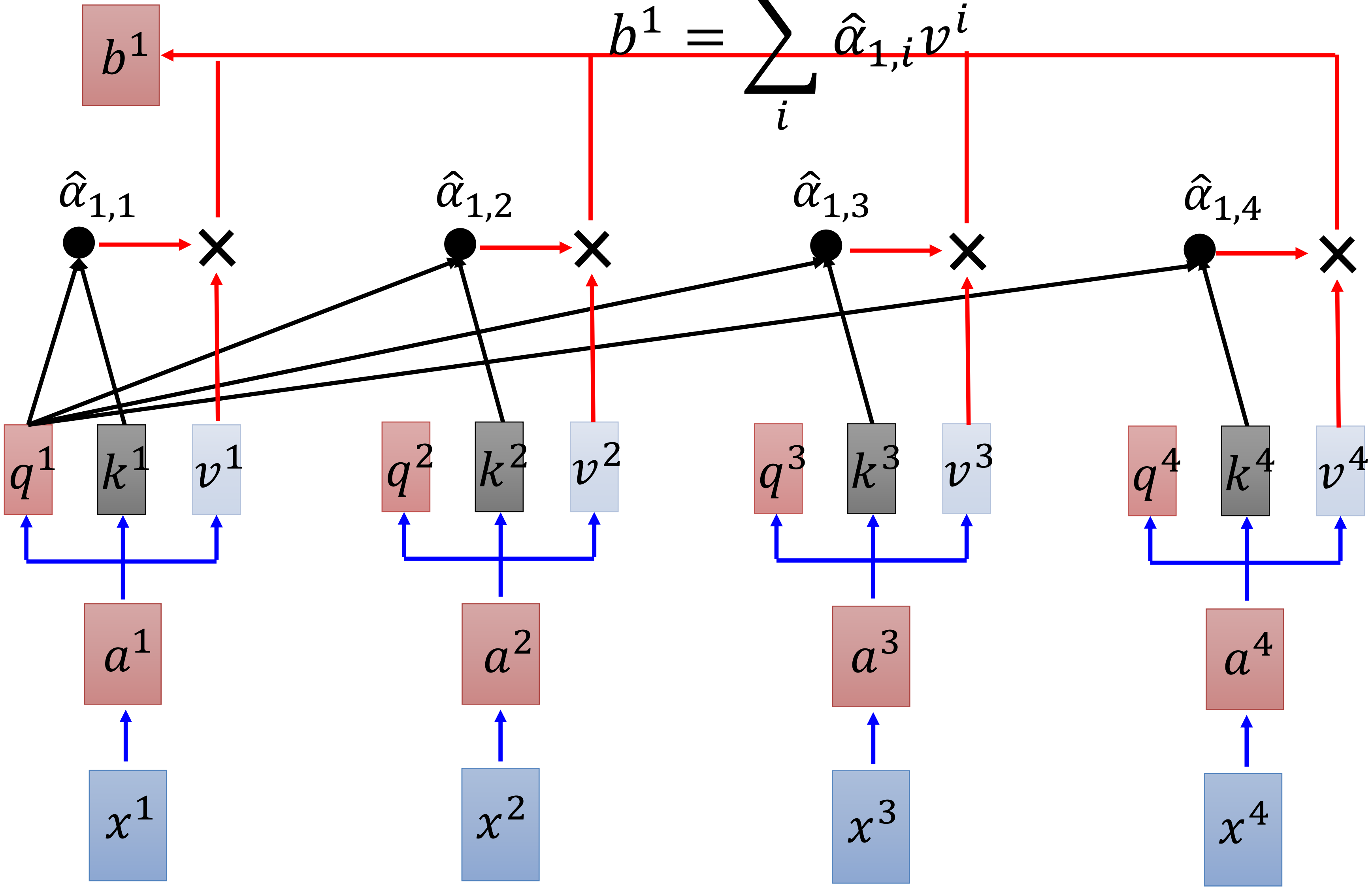
Scaled Dot-Product Attention

$$\hat{\alpha}_{1,i} = exp(\alpha_{1,i}) / \sum_j exp(\alpha_{1,j})$$

Self-Attention

Attention is all you need

$$b^1 = \sum_i \hat{\alpha}_{1,i} v^i$$

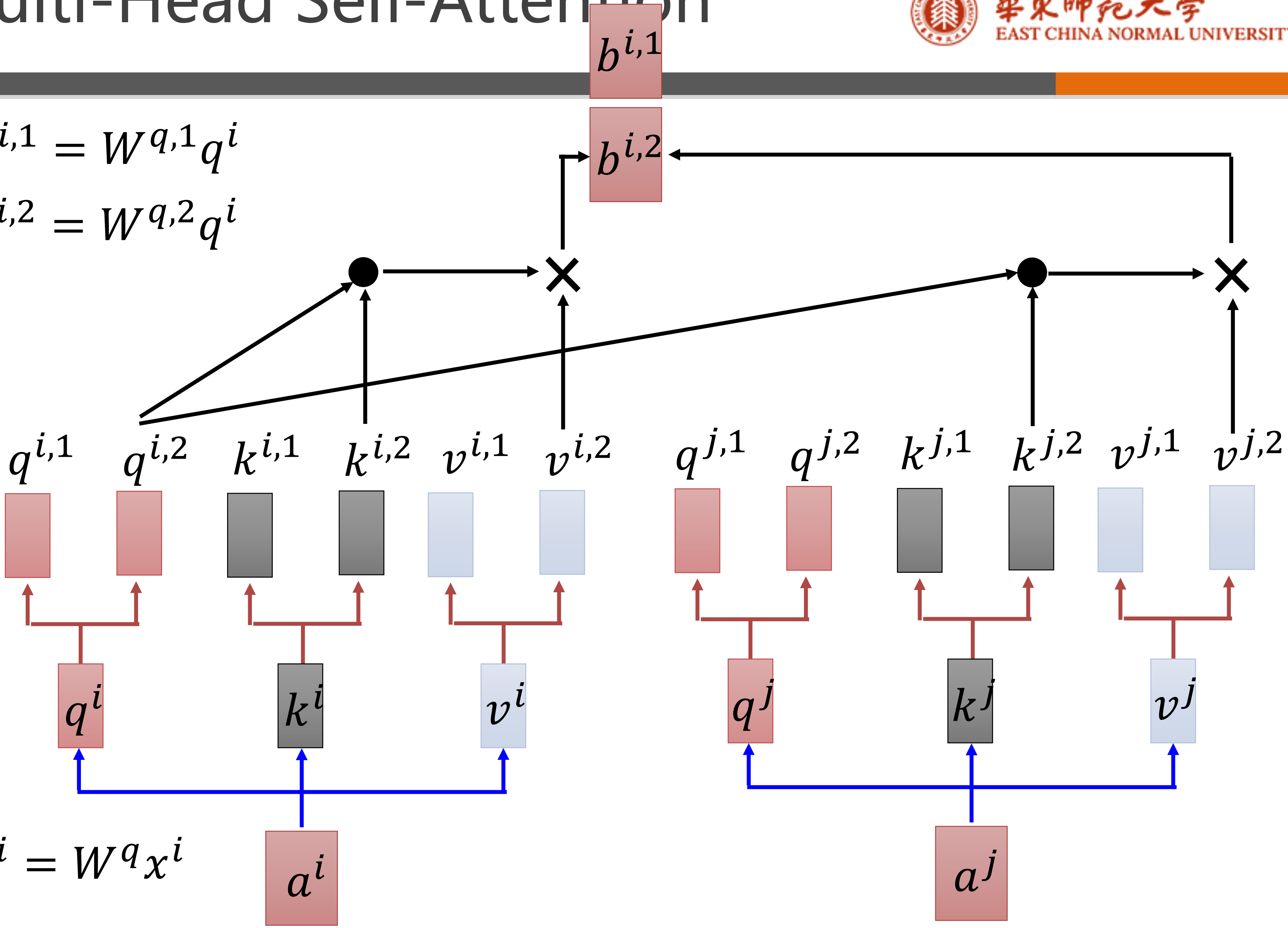


Multi-Head Self-Attention

$$q^{i,1} = W^{q,1} q^i$$

$$q^{i,2} = W^{q,2} q^i$$

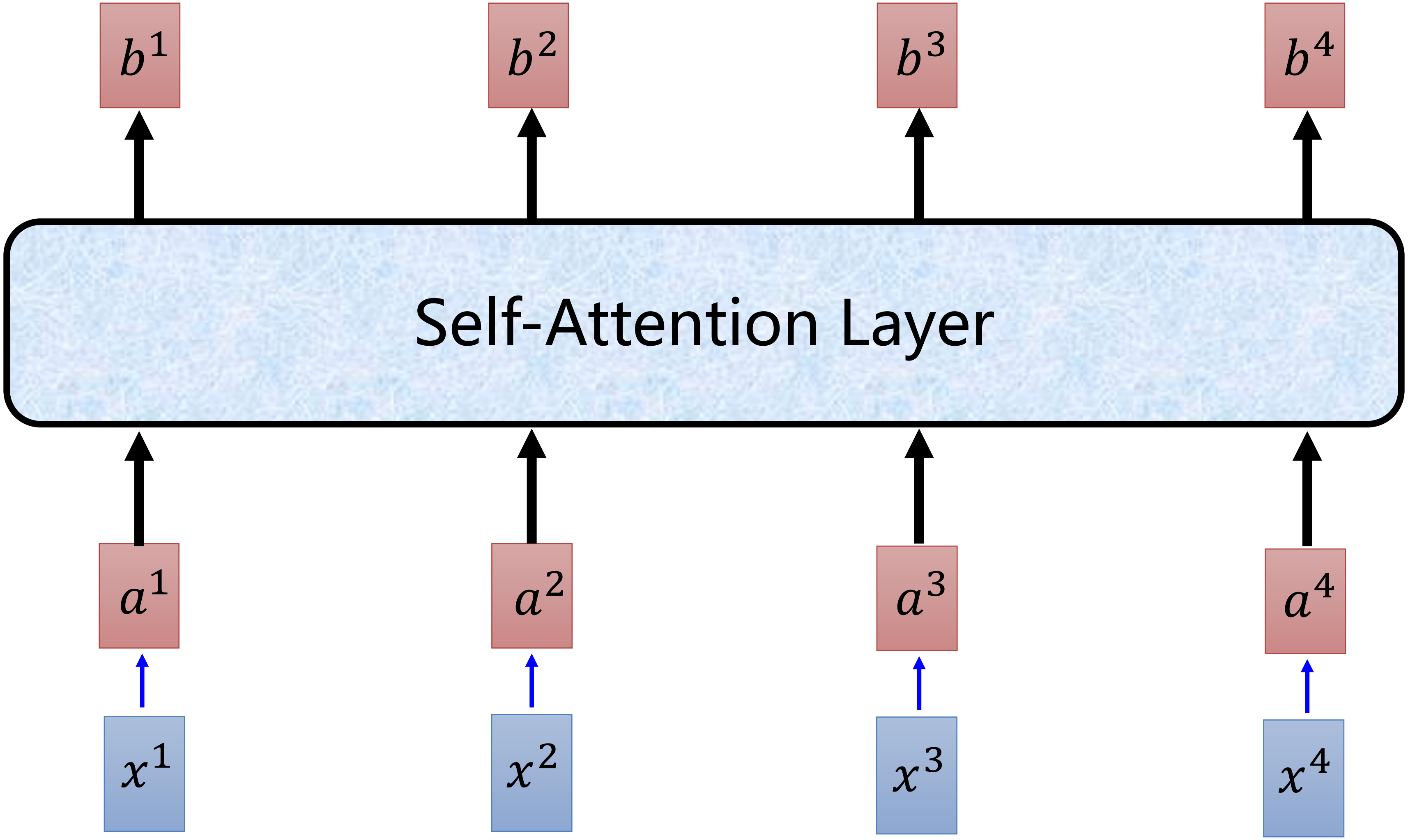
$$q^i = W^q x^i$$



Self-Attention

Attention is all you need

b^1, b^2, b^3, b^4 can be parallelly computed.



Transformer Attention is all you need

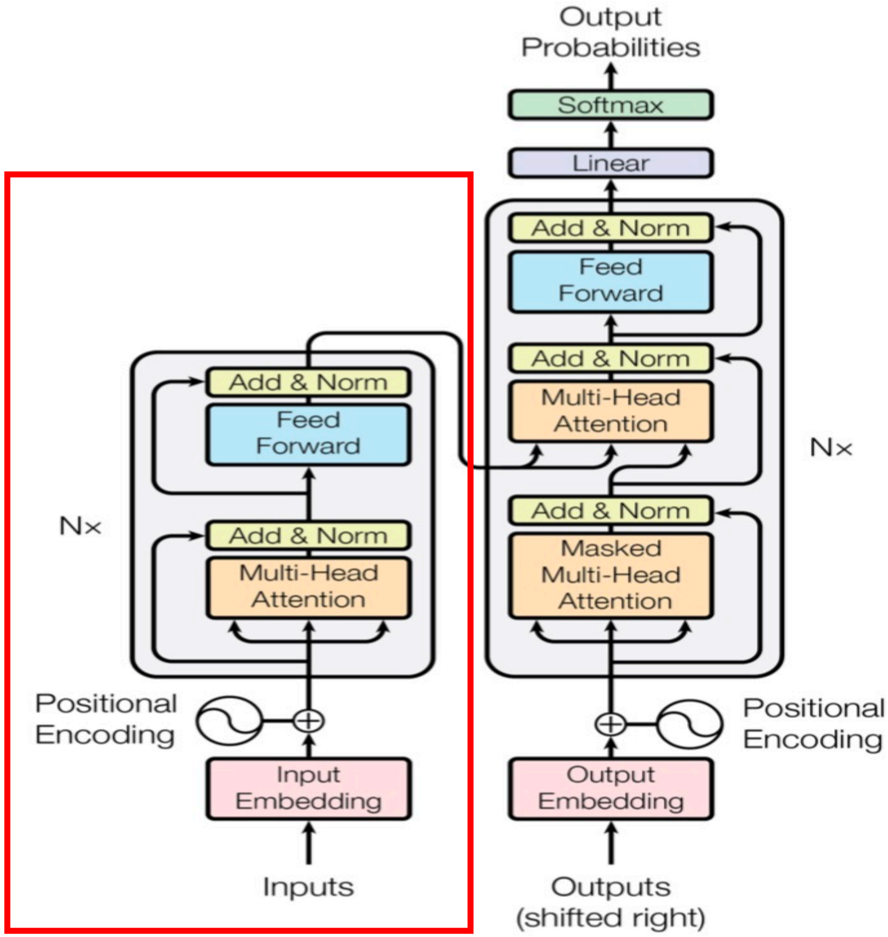


Figure 1: The Transformer - model architecture.

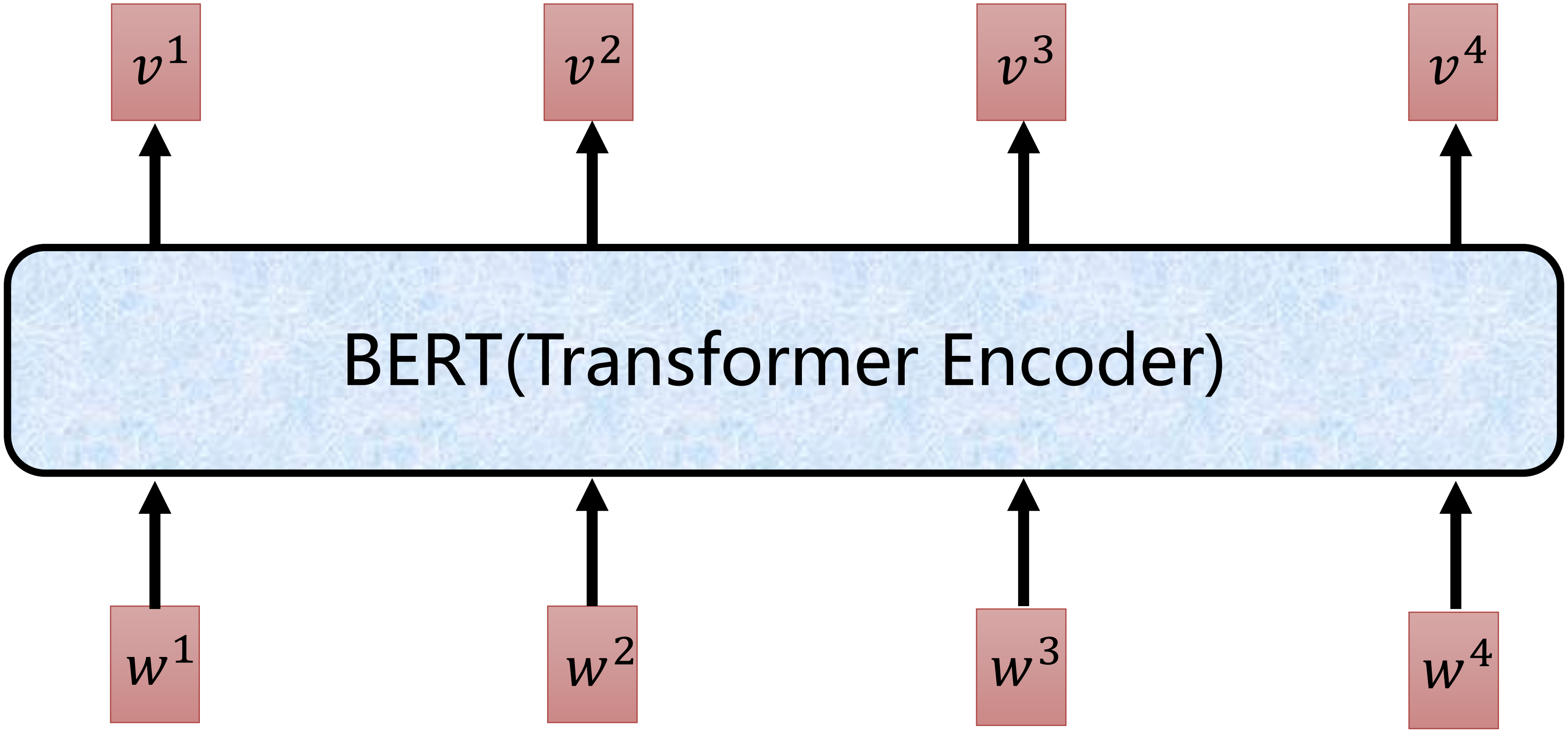
位置信息：由于模型中不包含递归和卷积，为了使模型能够利用序列的顺序，作者注入一些关于序列中标记的相对或绝对位置的信息。所以将“位置编码”添加到编码器和解码器堆栈底部的输入嵌入中。

Layer Normalization：缓解Internal Covariate Shift问题，可以将数据分布拉到激活函数的非饱和区，具有权重/数据伸缩不变性的特点。起到缓解梯度消失/爆炸、加速训练、正则化的效果。具体可以阅读他的原文。

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

BERT Bidirectional Encoder Representation from Transformers

BERT是通过大量句子来训练的，这些句子不需要用标注。



Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

训练方法1:

Masked LM

把输入的句子中，挖空住一定比例的词汇，让BERT进行补全，预测被遮挡的词汇。补全相当于一个Linear Multi-class Classifier。

如果两个词汇填在同一个位置意思相同，他们就会类似的embedding。

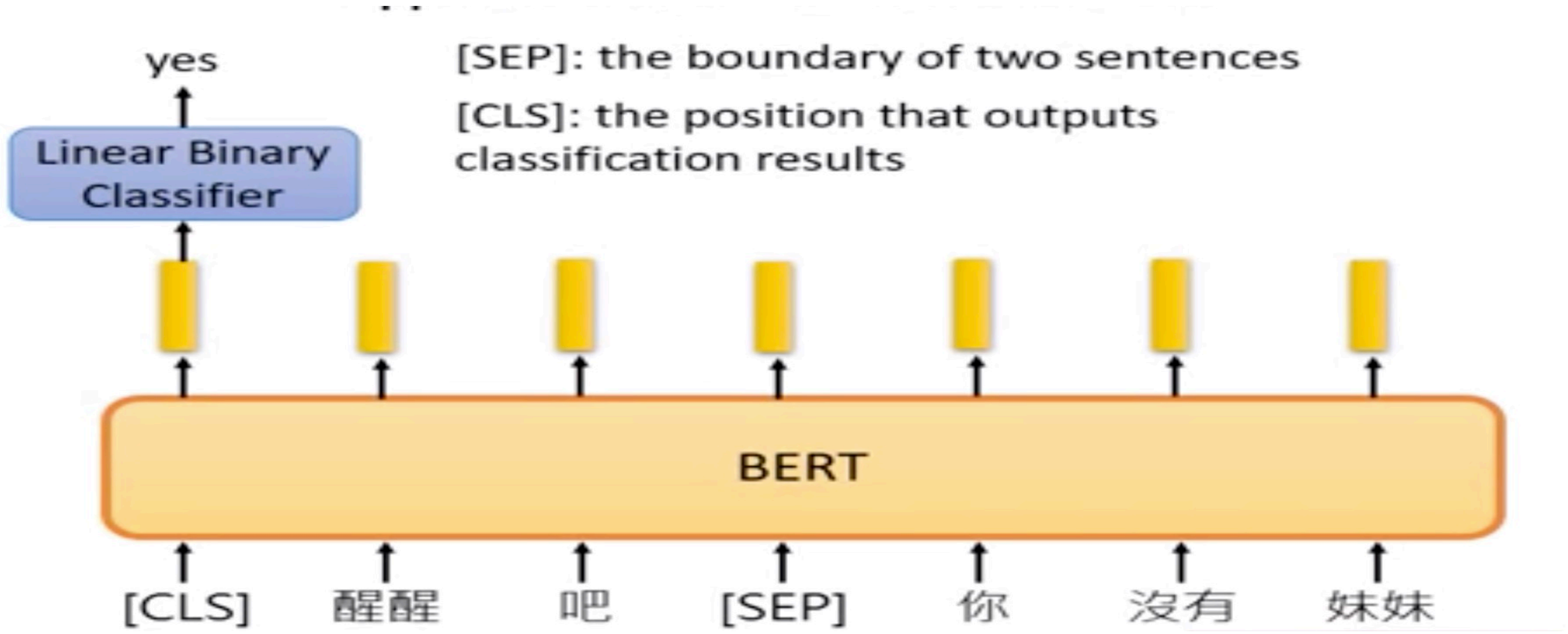
Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

BERT Bidirectional Encoder Representation from Transformers

训练方法2:

Next Sentence Prediction

给模型两个句子，让模型预测这两个句子是否应该接在一起。

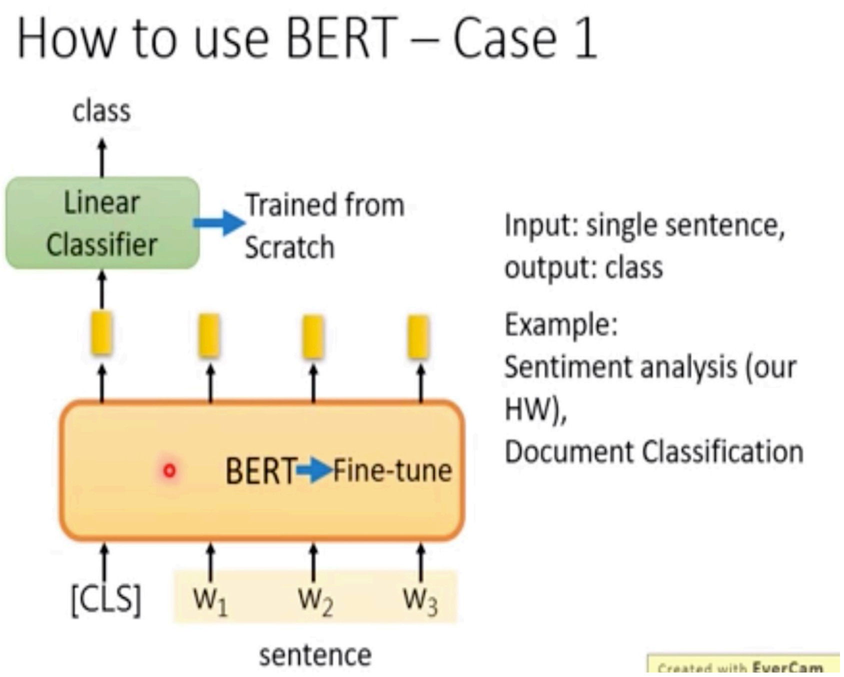


[SEP]: the boundary of two sentences

[CLS]: the position that outputs classification results

Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

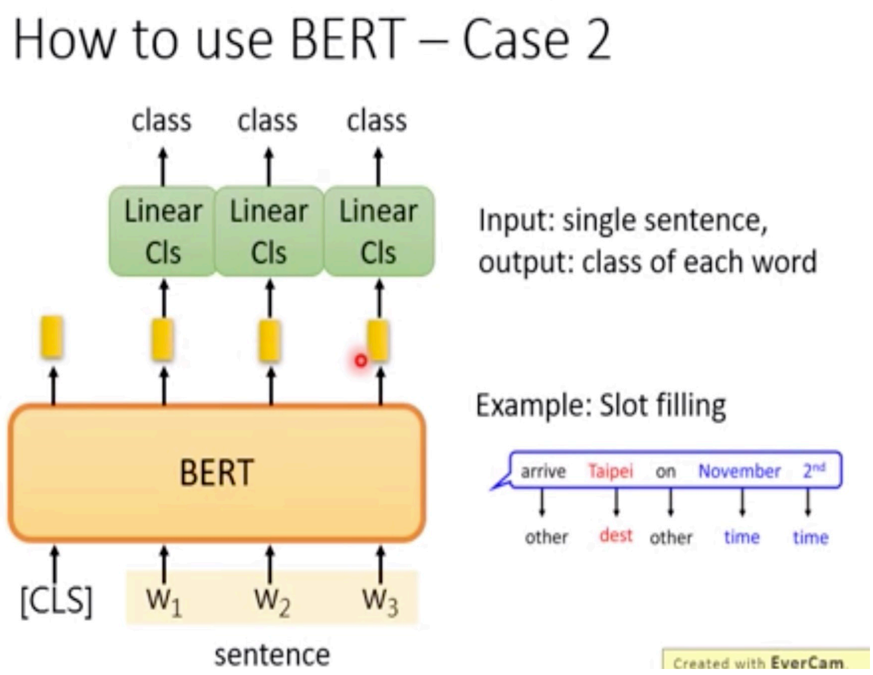
BERT Bidirectional Encoder Representation from Transformers



How to use BERT – Case 1

Input: single sentence, output: class

Example: Sentiment analysis (our HW), Document Classification



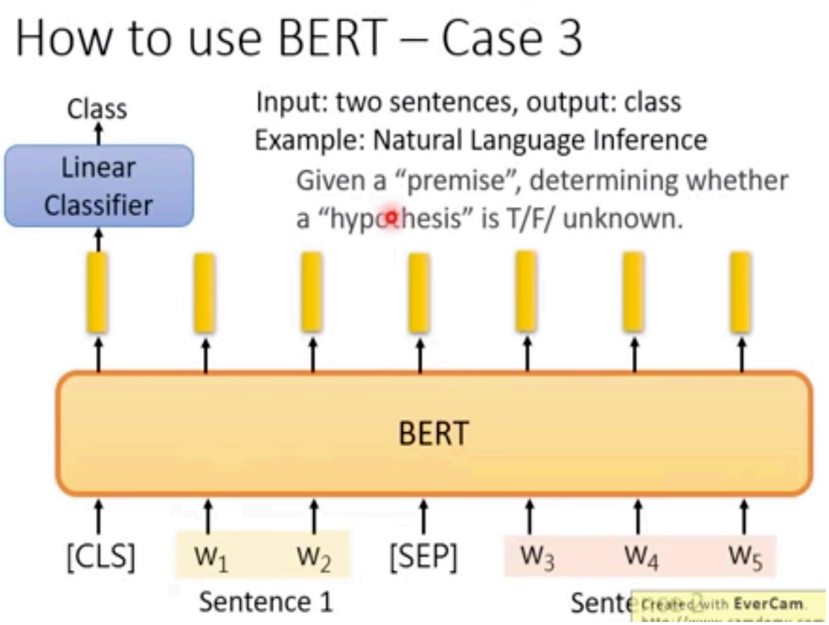
How to use BERT – Case 2

Input: single sentence, output: class of each word

Example: Slot filling

Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

BERT Bidirectional Encoder Representation from Transformers

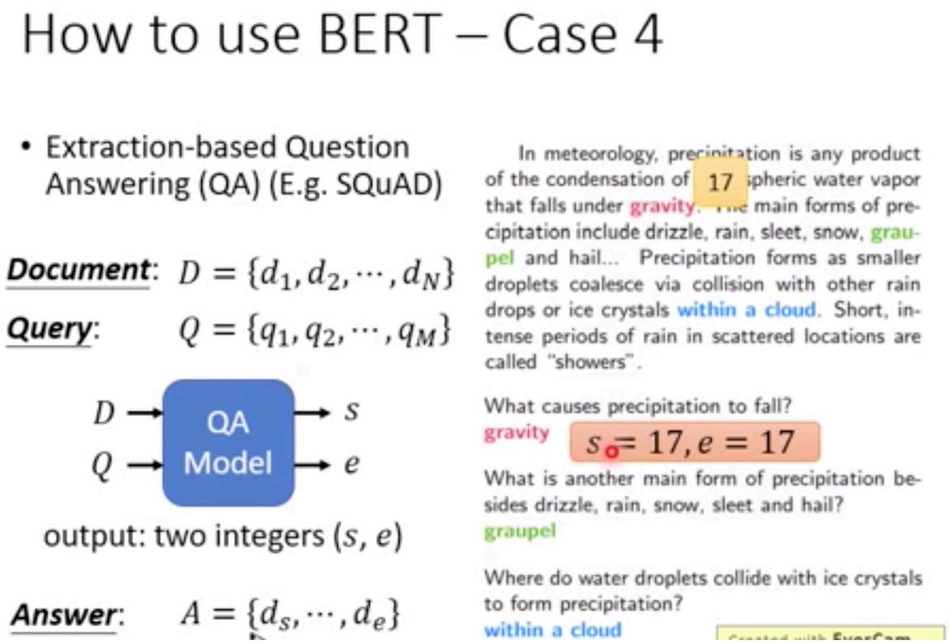


How to use BERT – Case 3

Input: two sentences, output: class

Example: Natural Language Inference

Given a "premise", determining whether a "hypothesis" is T/F/ unknown.



How to use BERT – Case 4

- Extraction-based Question Answering (QA) (E.g. SQuAD)

Document: $D = \{d_1, d_2, \cdots, d_N\}$

Query: $Q = \{q_1, q_2, \cdots, q_M\}$

output: two integers (s, e)

Answer: $A = \{d_s, \cdots, d_e\}$

In meteorology, precipitation is any product of the condensation of 17 spheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity $s = 17, e = 17$

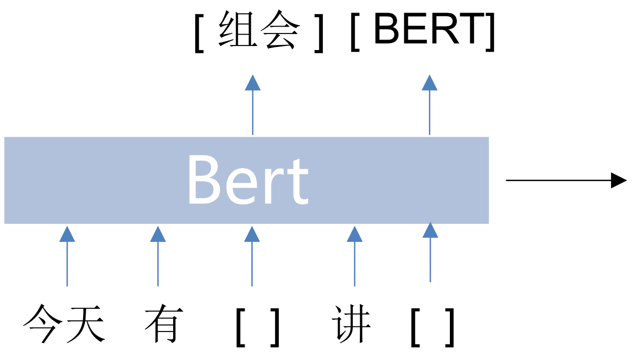
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

关于BERT的思考

1. Bert得到的动态Embedding，是上下文相关的。而传统的词嵌入技术（CBOW，Skip-Gram，Glove）是静态的Embedding，不能解决一次多义的情况，且与上下文无关。
2. Bert是基于token独立性假设的。3. 不适合做生成任务



今天 / 有 / 讲 ----- 组会
今天 / 有 / 讲 ----- BERT

XLNet
考虑上下文，同时可以生成任务。考虑句子中单词的所有顺序。

RoBERTa

MASS（encoder+decoder）

思考：**self-attention这样的机制，是不是弱化了语法作用？**

BERT在多模态领域中的应用

主要可以分类两类：
1.单流模型：在模型中将文本信息和视觉信息在一开始就进行融合。

2.双流模型：文本信息和视觉信息先分别经过两个独立的编码模块，然后通过互相的注意力机制来实现不同模态信息的融合。

单流模型 VisualBERT

和 BERT 类似，VisualBERT 在结构上采用了堆叠的 Transformer。在一开始就将文字和图片信息通过 Transformer 的自注意力机制进行对齐融合。

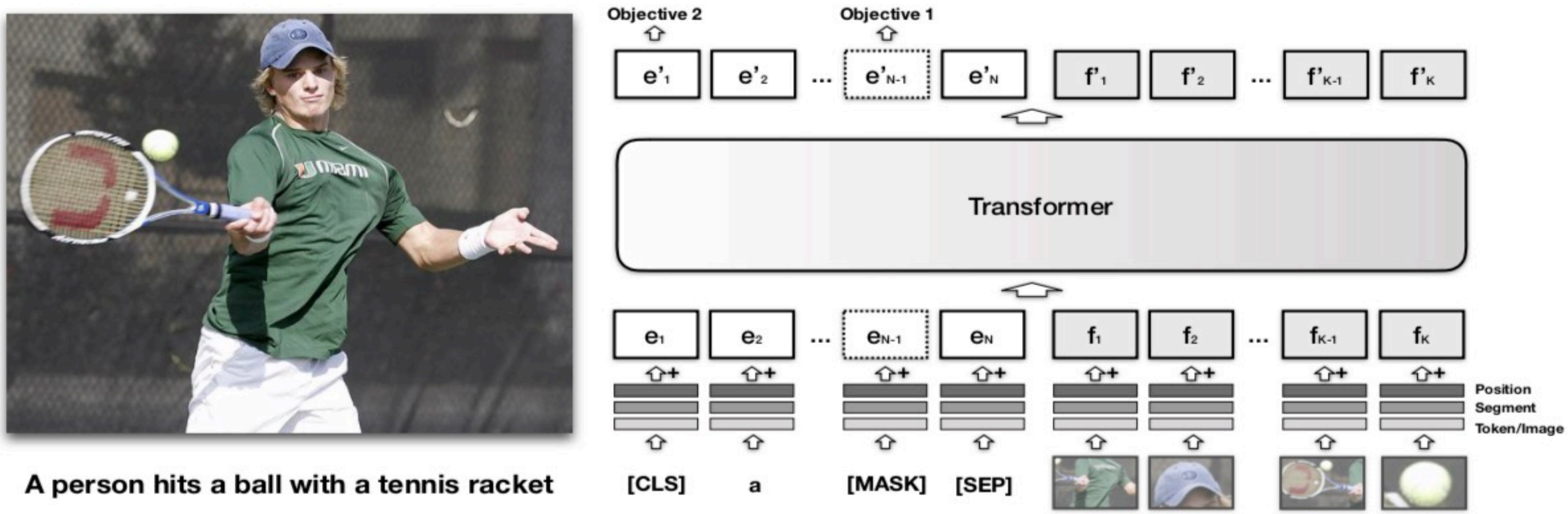


Figure 2: The architecture of VisualBERT. Image regions and language are combined with a Transformer to allow the self-attention to discover implicit alignments between language and vision. It is pre-trained with a masked language modeling (Objective 1), and sentence-image prediction task (Objective 2), on caption data and then fine-tuned for different tasks. See §3.3 for more details.

Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." arXiv preprint arXiv:1908.03557 (2019).

单流模型

VisualBERT

VisualBERT 遵循 BERT 一样的流程，先进行预训练然后在相应的任务上进行微调。

两个预训练任务：

1. BERT 一样的LM Masked

2.句子-图像预测 （判断输入的句子是否为相应图片的描述）。

作者在 VQA，VCR，NLVR2 和 Flickr30k 四个视觉语言任务上进行了测试。进一步的消融实验表明 VisualBERT 可以有效地学习到语言和相应图像区域的联系，同时也具有一定的句法敏感性。

Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." arXiv preprint arXiv:1908.03557 (2019).

单流模型
Unicoder-VL

与VisualBERT 相似，在结构上同样采用堆叠的 Transformer，并且在一开始就对图像和语言信息进行对齐和融合。

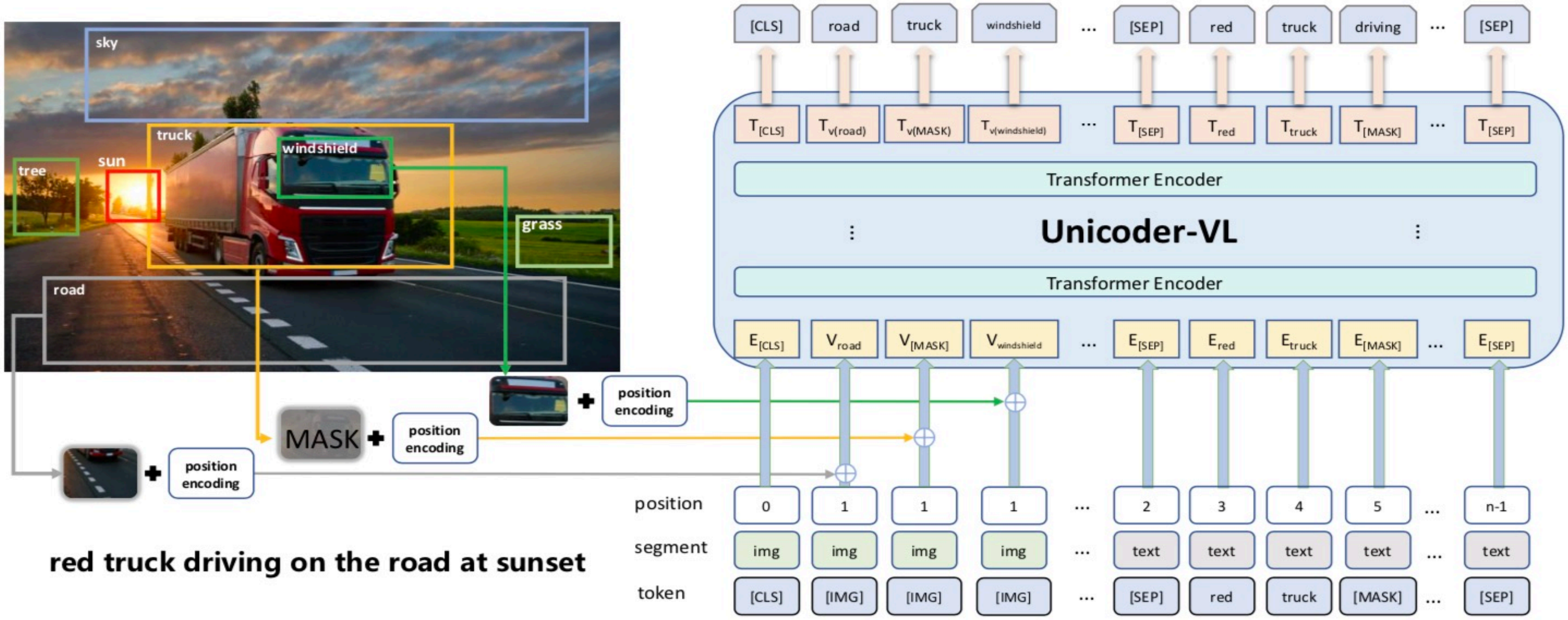


Figure 1: Illustration of Unicoder-VL in the context of an object and text masked token prediction, or *cloze*, task. Unicoder-VL contains multiple Transformer encoders which are used to learn visual and linguistic representation jointly.

Li, Gen, et al. "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training." arXiv preprint arXiv:1908.06066 (2019).

单流模型
Unicoder-VL

与 VisualBERT 最大的不同在于改模型在输入端对图像的处理。其文字部分的输入与（1）中相似。在图像的输入上，其首先通过 Faster-RCNN 提取区域图像特征，然后将该特征与区域图像在图像中的位置编码进行拼接再经过一个连接层投影到与语言输入维度相同的空间。

遵循先预训练后微调的模式。该模型在三个任务中进行预训练，

- 语言掩码
- 图像语言匹配任务
- 图像标签预测，即预测区域图像所物体类别

Li, Gen, et al. "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training." arXiv preprint arXiv:1908.06066 (2019).

单流模型
VL-BERT

在输入端与上述两个模型略有不同。不同在于在文本输入的每个token位置加入了图像的特征embedding（完整图像的特征），而在区域特征这一块直接接上了Fast-RCNN联合训练。

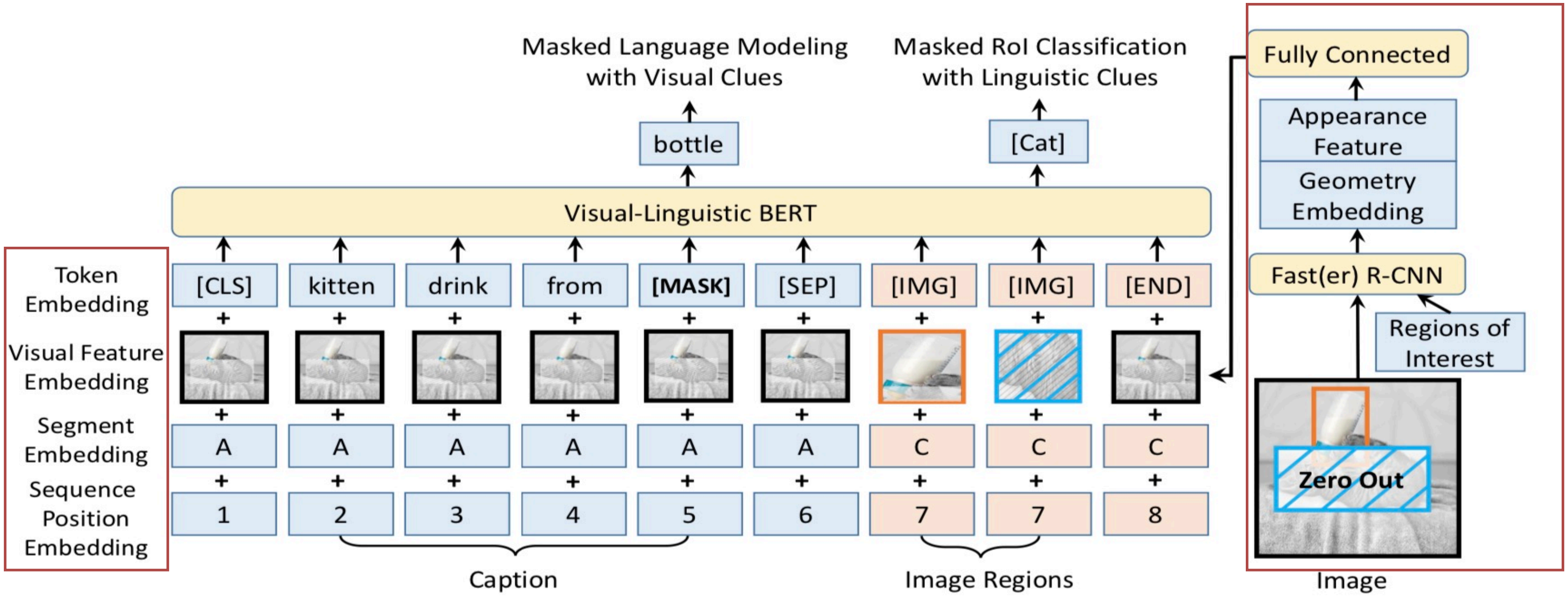


Figure 1: Architecture for pre-training VL-BERT. All the parameters in this architecture including VL-BERT and Fast R-CNN are jointly trained in both pre-training and fine-tuning phases.

Su, Weijie, et al. "Vl-bert: Pre-training of generic visual-linguistic representations." ICLR2020

单流模型总结

基本思想：将文本和图像同时输入到模型中。通过self-attention机制进行对齐，交互等。

图像信息是通过目标检测模型输出的ROI特征作为区域特征输入。和文本不同的是，这是一个无序的排列。

各个模型的差异在于输入端的内容。输入端大概包括：图像区域embedding，token embedding ，segment embedding ，position embedding 等。

存在问题：ROI特征是否和token embedding在同一语义空间，直接求self-attention会不会降低原本性能较优的单模态BERT。

双流模型
ViL-BERT

基于双流的 ViLBERT，在一开始并未直接对语言信息和图片信息进行融合，而是先各自经过 Transformer 的编码器进行编码。分流设计是基于一个假设：语言的理解本身比图像复杂，而且图像的输入本身就是经过 Faster-RCNN 提取的较高层次的特征，因此两者所需要的编码深度应该是不一样的。

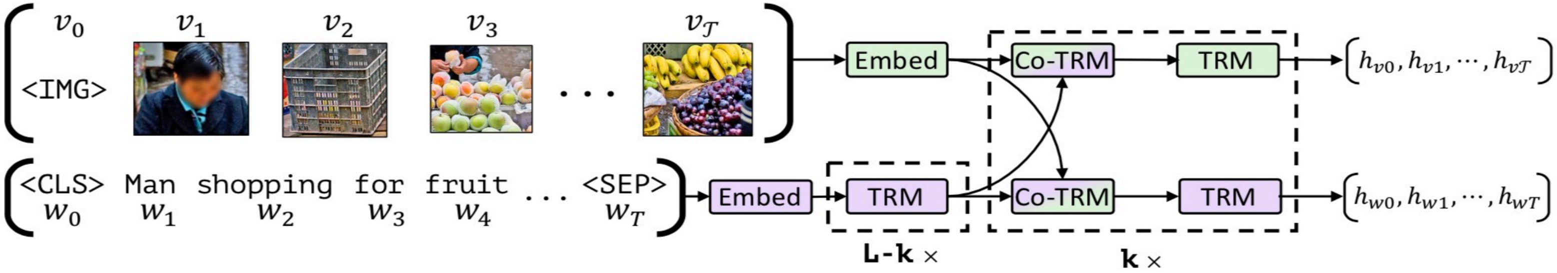
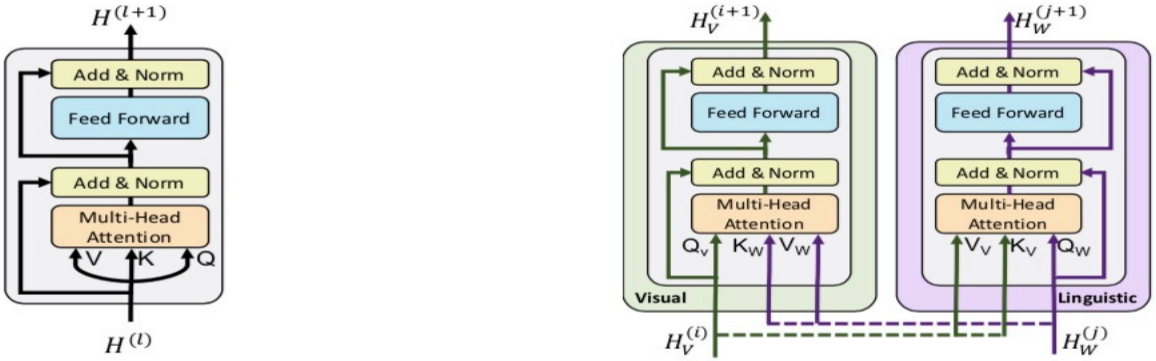


Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.

Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." Advances in Neural Information Processing Systems. 2019.

双流模型
ViL-BERT

当两种模态各自进行编码后，其输出会经过一个共注意力机制模块（如下图右侧所示）。该模块也是基于 Transformer 的结构，只是在自注意力机制中每个模块都用自己的 Query 去和另一模块的 Value 和 Key 计算注意力，由此来融合不同模块间的信息。



(a) Standard encoder transformer block (b) Our co-attention transformer layer

Figure 2: We introduce a novel co-attention mechanism based on the transformer architecture. By exchanging key-value pairs in multi-headed attention, this structure enables vision-attended language features to be incorporated into visual representations (and vice versa).

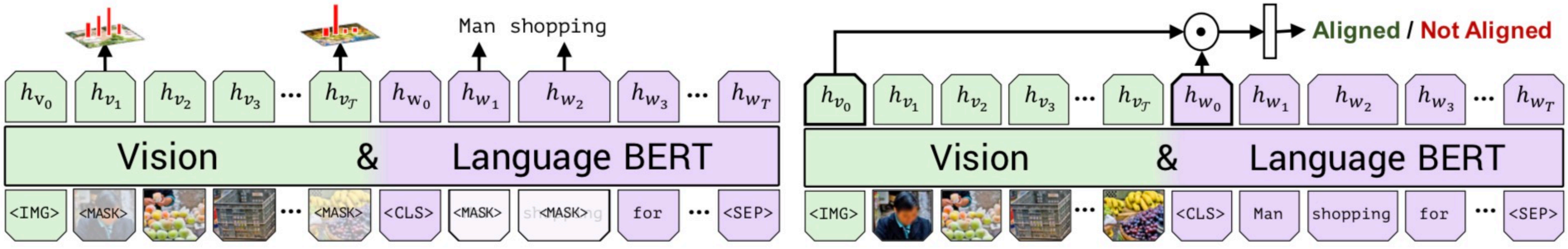
Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." Advances in Neural Information Processing Systems. 2019.

双流模型
ViL-BERT

预训练：

1. Masked multimodal modeling，在图像端任务的目标则是当区域图像被掩盖后模型对其输出的分类分布能够尽可能与用来提取区域特征的模型（这里是 Faster-RCNN）的输出分布一致，因此这里作者使用 KL 散度作为目标函数；

2.Multimodal alignment prediction: 语言图像匹配任务。


Man shopping
h_{v_0} h_{v_1} h_{v_2} h_{v_3} ··· $h_{v_\mathcal{T}}$ h_{w_0} h_{w_1} h_{w_2} h_{w_3} ··· h_{w_T}
Vision & Language BERT
<IMG> <MASK> ··· <MASK> <CLS> <MASK> <MASK> for ··· <SEP>
Aligned / Not Aligned
h_{v_0} h_{v_1} h_{v_2} h_{v_3} ··· $h_{v_\mathcal{T}}$ h_{w_0} h_{w_1} h_{w_2} h_{w_3} ··· h_{w_T}
Vision & Language BERT
<IMG> ··· <CLS> Man shopping for ··· <SEP>

(a) Masked multi-modal learning

(b) Multi-modal alignment prediction

Figure 3: We train ViLBERT on the Conceptual Captions [24] dataset under two training tasks to learn visual grounding. In masked multi-modal learning, the model must reconstruct image region categories or words for masked inputs given the observed inputs. In multi-modal alignment prediction, the model must predict whether or not the caption describes the image content.

双流模型
ViL-BERT

作者分别在 VQA, VCR, GRE, IR, ZSIR 等五个任务中最模型进行测试。实验结果表明ViLBERT在各个任务上都提升了2~10个百分点的精度。此外，ViLBERT针对这些任务的修改很简单，所以该模型可以作为跨多个视觉和语言任务的视觉基础。其后作者又对预训练过程进行分析发现与训练过程中模型已经能够学习到语言与图像在语义上的对齐关系。

Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." Advances in Neural Information Processing Systems. 2019.

双流模型
LXMERT

该模型与 ViLBERT 一样采用了双流模型。语言与图像在一开始先各自经过独立的编码层进行编码，然后再经过一个模态交互编码层进行语言与图像在语义上的对齐和融合。

在交互编码层中，该模型同样的也是使用共注意力机制，即自注意力中的 query 来自一个模态，而 key 和 value 来自另一个模态。该编码层过后，图像与语言各自又经过一层自注意力层进一步提取高层特征。

LXMERT: Learning cross-modality encoder representations from transformers. Hao Tan, Mohit Bansal,EMNLP2019

双流模型
LXMERT

该模型的输出有三个部分，一个语言端的输出，一个图像端的输出，一个多模态的输出。该模型在与训练时使用了四个任务：语言掩码任务，图像掩码任务（该任务有两部分，第一部分为预测被掩图像物体类别，第二部分为 ROI 特征回归任务该任务使用 L2 损失函数，语言图像匹配任务和图像问答任务。

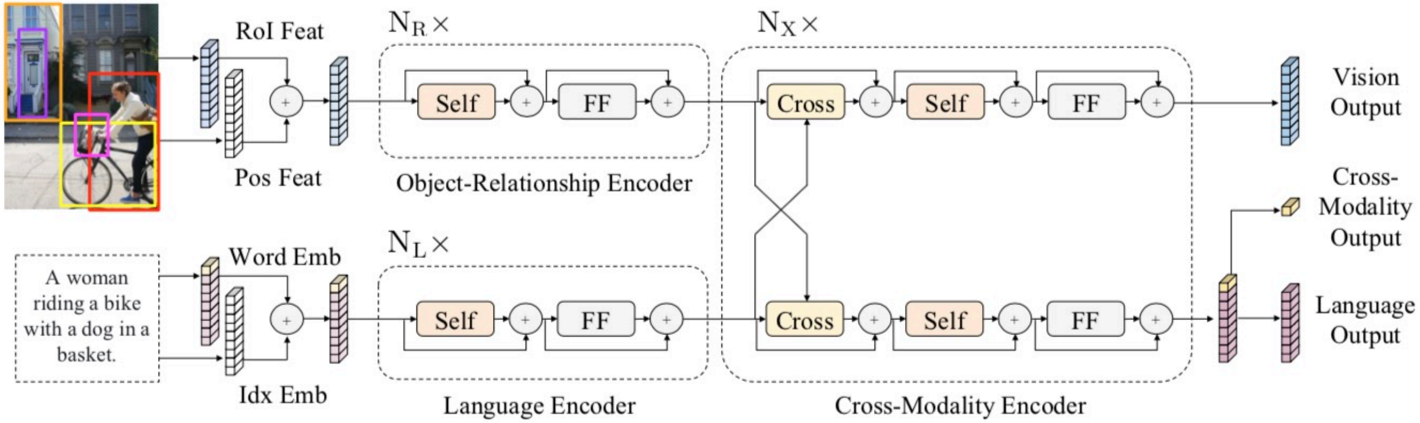
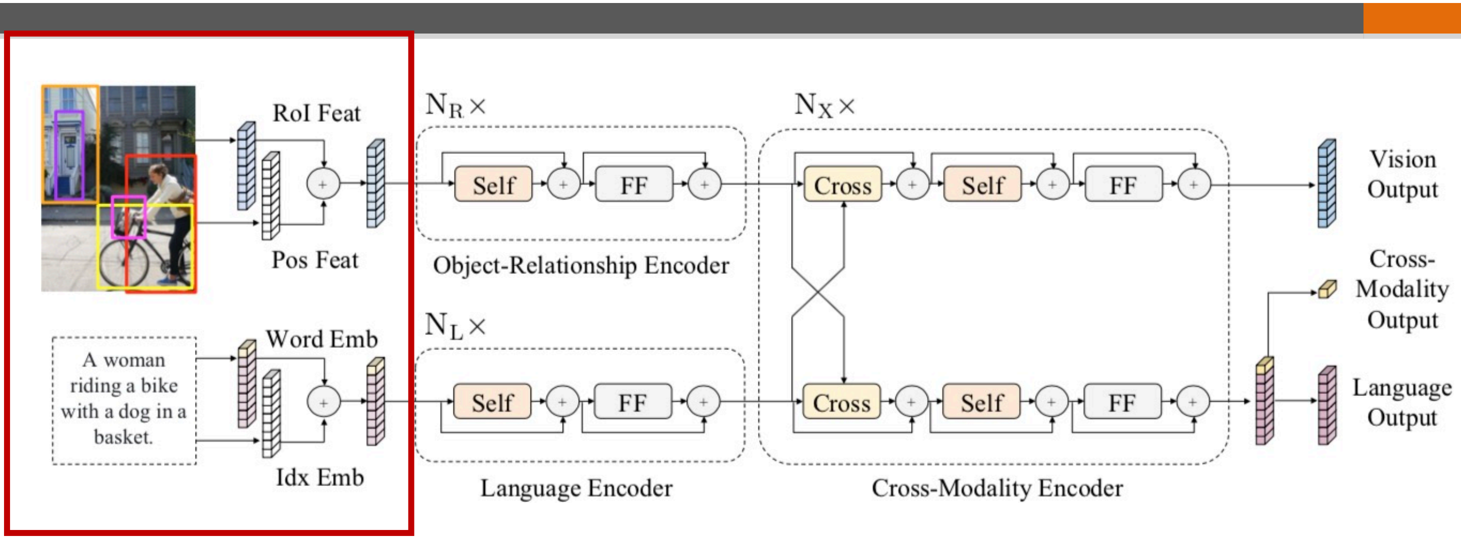


Figure 1: The LXMERT model for learning vision-and-language cross-modality representations. ‘Self’ and ‘Cross’ are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. ‘FF’ denotes a feed-forward sub-layer.

LXMERT: Learning cross-modality encoder representations from transformers. Hao Tan, Mohit Bansal,EMNLP2019

双流模型
LXMERT



Word-Level Sentence Embeddings

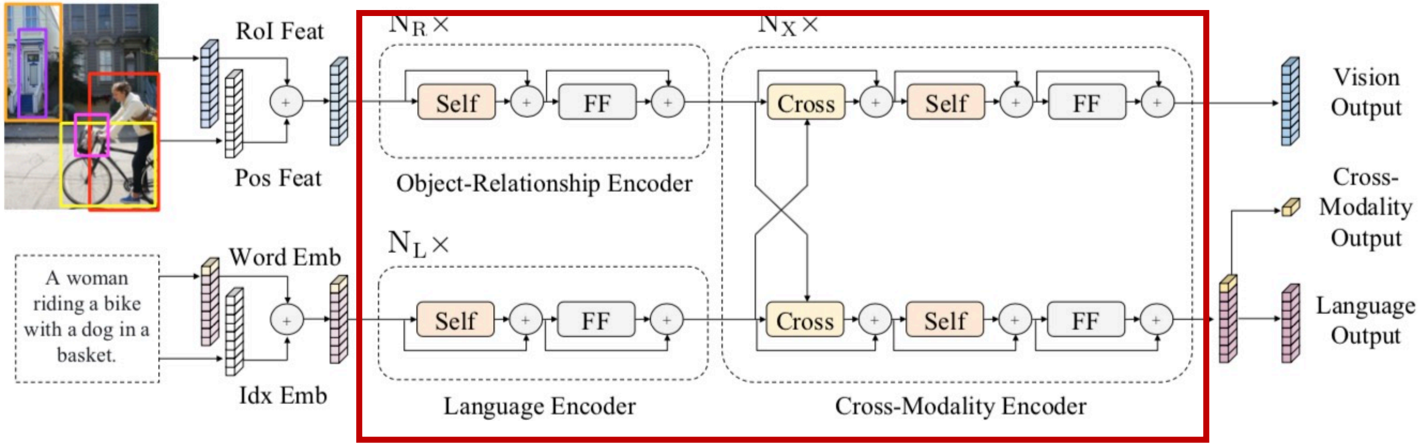
$$\hat{w}_i = \mathrm{WordEmbed}\left(w_i\right)$$
$$\hat{u}_i = \mathrm{IdxEmbed}\left(i\right)$$
$$h_i = \mathrm{LayerNorm}\left(\hat{w}_i + \hat{u}_i\right)$$

Object-Level Image Embeddings

$$\hat{f}_j = \mathrm{LayerNorm}\left(W_{\mathrm{F}} f_j + b_{\mathrm{F}}\right)$$
$$\hat{p}_j = \mathrm{LayerNorm}\left(W_{\mathrm{P}} p_j + b_{\mathrm{P}}\right)$$
$$v_j = \left(\hat{f}_j + \hat{p}_j\right)/2$$

LXMERT: Learning cross-modality encoder representations from transformers. Hao Tan, Mohit Bansal,EMNLP2019

双流模型
LXMERT



Single-Modality Encoders

- ➢ self-attention (‘Self’) sub-layer
- ➢ feed-forward (‘FF’) sub-layer

(two fully-connected sub-layers)

Cross-Modality Encoder

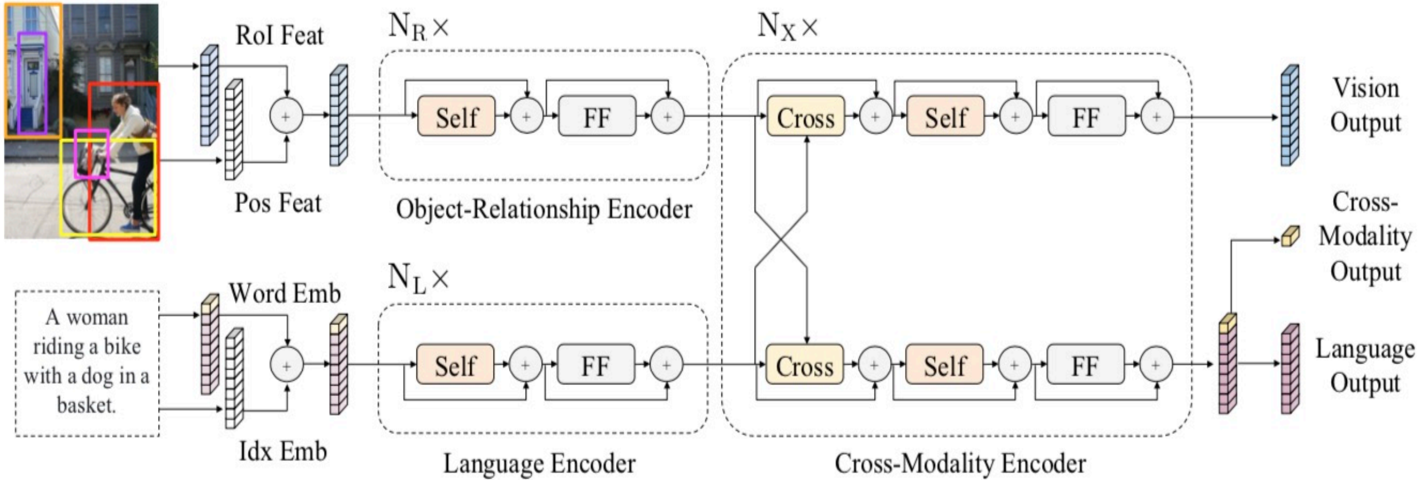
$$\hat{h}_i^k = \mathrm{CrossAtt}_{\mathrm{L}\rightarrow\mathrm{R}}\left(h_i^{k-1},\{v_1^{k-1},\dots,v_m^{k-1}\}\right)$$
$$\hat{v}_j^k = \mathrm{CrossAtt}_{\mathrm{R}\rightarrow\mathrm{L}}\left(v_j^{k-1},\{h_1^{k-1},\dots,h_n^{k-1}\}\right)$$

Cross-modality representations

$$\tilde{h}_i^k = \mathrm{SelfAtt}_{\mathrm{L}\rightarrow\mathrm{L}}\left(\hat{h}_i^k,\{\hat{h}_1^k,\dots,\hat{h}_n^k\}\right)$$
$$\tilde{v}_j^k = \mathrm{SelfAtt}_{\mathrm{R}\rightarrow\mathrm{R}}\left(\hat{v}_j^k,\{\hat{v}_1^k,\dots,\hat{v}_m^k\}\right)$$

LXMERT: Learning cross-modality encoder representations from transformers. Hao Tan, Mohit Bansal,EMNLP2019

双流模型
LXMERT



Pre-Training Tasks

- ➢ Masked cross-modality language modeling
- ➢ Masked object prediction(ROI-feature regression & detected-label classification)
- ➢ Cross-modality matching
- ➢ Image question answering

LXMERT: Learning cross-modality encoder representations from transformers. Hao Tan, Mohit Bansal,EMNLP2019

双流模型
ImageBERT

本文作者提出了一种新的视觉语言预训练模型ImageBERT，该模型基于Transformer架构，并对视觉-语言联合嵌入进行建模。

作者从网络上收集了一千万规模的弱监督图像-文本数据集LAIT，这也是当前所有视觉-语言数据集中较大的数据集。在这个数据集的帮助下，ImageBERT 模型在MSCOCO和Flickr30k的图像-文本检索任务上获得不错的结果。

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, Arun Sacheti: ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. CoRR abs/2001.07966 (2020)

总结

BERT是面向输入为一个或两个句子的 NLP 任务的预训练模型。为了将 BERT 架构应用于跨模态任务中，现在已有多种不同的方法。
ViLBERT和LXMERT 先分别应用一个单模态Transformer到图像和句子上，之后再采用跨模态Transformer来结合这两种模态。其他工作如VisualBERT， B2T2，Unicoder-VL， VL-BERT， Unified VLP，UNITER等等，则都是将图像和句子串联为Transformer的单个输入。很难说哪个模型架构更好，因为模型的性能非常依赖于指定的场景。

除了文本和图像结合的，还有文本和视频结合的VideoBERT，和语音结合的SpeechBERT。

存在问题：模型很大，很难训练。

扩展

基于BERT的模型压缩

1.Sparse Prioris（i.e. Bayesian Compression）

2. Sparse Matrix Factorization（i.e. Huge sparse parameter matrix）, ALBERT

3.Knowledge Distillation, TinyBERT，DistillBERT

Thank you!