Transformer Models for Image, Video and Text Processing

Prof. C. Chandra Sekhar

Department of Computer Science and Engineering
Indian Institute of Technology Madras

Sequence-to-Sequence Mapping Models

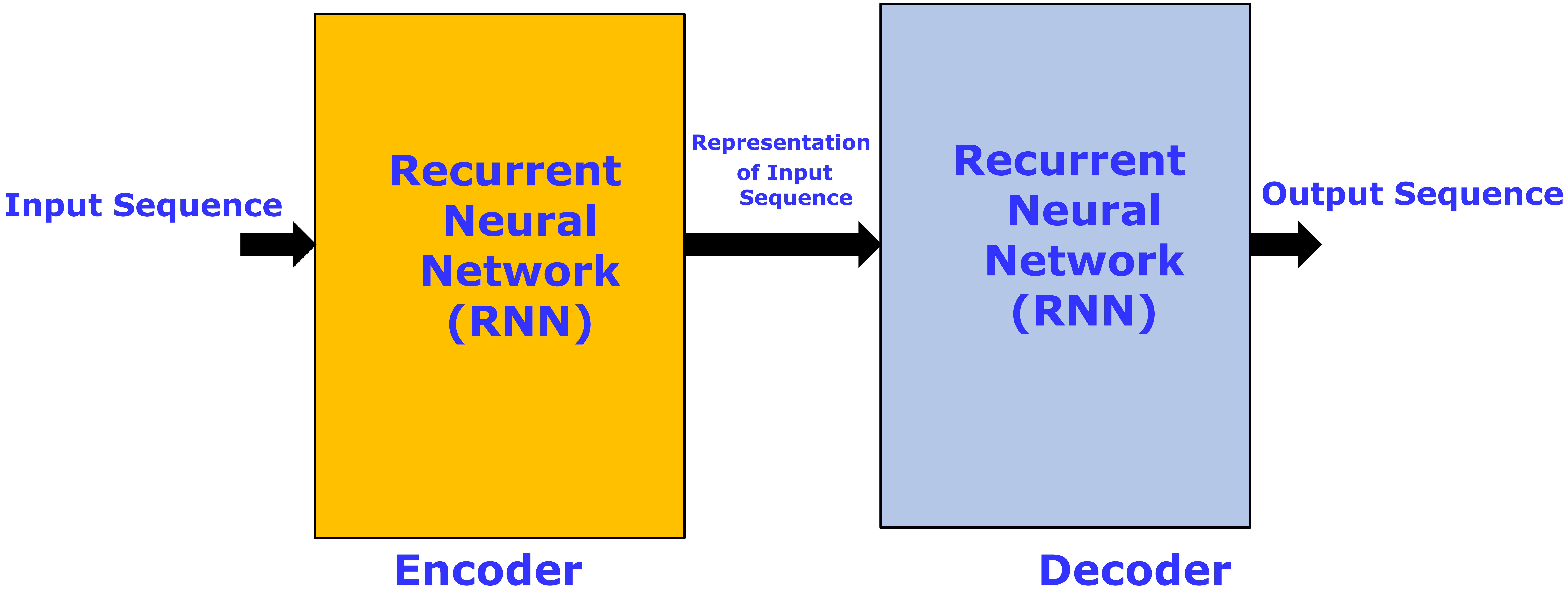
Text Processing Tasks

- **Sentence classification**
- **Parts-of-speech tagging**
- **Named entity recognition**
- **Machine translation**
- **Text summarization**
- **Textual question answering**

Sequence-to-Sequence Mapping Tasks

- **Neural Machine Translation: Translation of a sentence in the source language to a sentence in the target language**
  - **Input: A sequence of words**
  - **Output: A sequence of words**
- **Video Captioning: Generation of a sentence as the caption for a video represented as a sequence of frames**
  - **Input: A sequence of feature vectors extracted from the frames of a video**
  - **Output: A sequence of words**
- **Each of the above tasks involves mapping an input sequence to an output sequence**

Encoder-Decoder Paradigm for Sequence-to-Sequence Mapping



Encoder-Decoder Paradigm for Sequence-to-Sequence Mapping

- **Sequence-to-Sequence Mapping using Encoder-Decoder Paradigm**
  - **Encoder: Generate a representation of the input sequence**
  - **Representation generated by Encoder is given as input to Decoder**
  - **Decoder: Generate the output sequence (A sequence of words)**
- **Relationship among the elements of a sequence:**
  - **Typically, an element in the input sequence is related to a few other elements in the input sequence**
  - **Typically, a word in the output sequence to be generated is related to a few elements in the input sequence**
- **LSTM based approach to Sequence-to-Sequence Mapping**
  - **Bidirectional LSTM based Encoder captures dependencies among elements in the input sequence**
  - **Bidirectional LSTM based Decoder captures dependencies among elements in the output sequence**
  - **Attention mechanism is introduced to capture dependencies of elements in the output sequence on elements in the input sequence**
- **Training the LSTM based Sequence-to-Sequence mapping systems is computationally intensive, and there is not much scope for parallelization of operations in the training process**

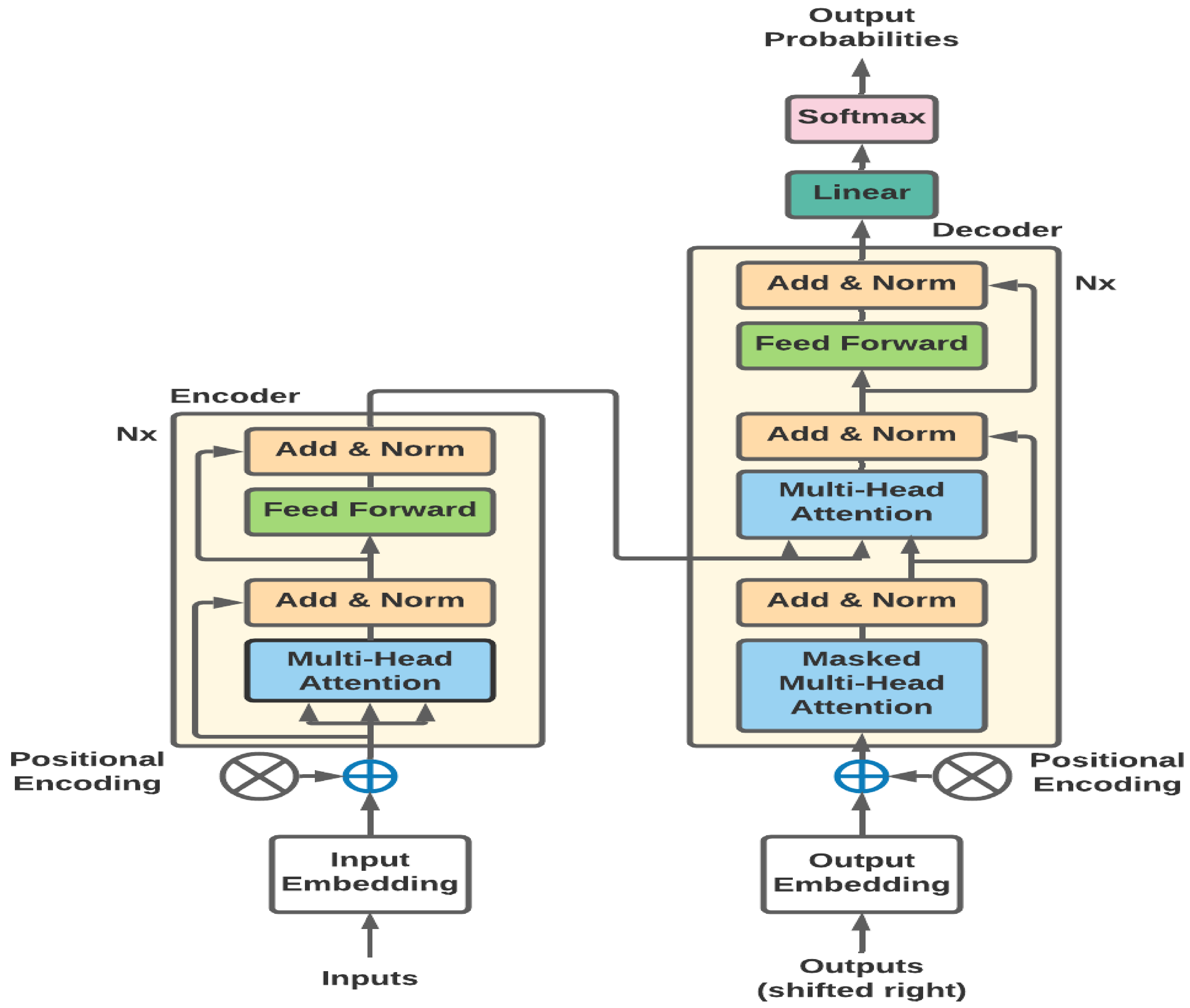
Transformer Models

Attention based Models for Sequence-to-Sequence Mapping

- **Attention based models try to capture and use**
  - **Relations among elements in the input sequence (Self-Attention)**
  - **Relations among elements in the output sequence (Self-Attention)**
  - **Relations between elements in the input sequence and elements in the output sequence (Cross-Attention)**

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," NIPS, 2017.

Attention-based Model: Transformer



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," NIPS, 2017.

Scaled Dot-Product Attention

- **Terminology borrowed from linear regression method for function approximation:**
- **Approximation of a function *f* of *d* variables:** $x_1, x_2, \ldots, x_d$
- **Let** $\boldsymbol{x} = [x_1, x_2, \ldots, x_d]^t$
- **Function approximation task: Given a set of *N* examples in the training dataset ,** $\boldsymbol{D} = \{\boldsymbol{x_n}, \ y_n\}, n = 1, 2, \ldots, N$**, predict the approximate estimate** $\hat{y}$ **of** $y = f(\boldsymbol{x})$**.**
- **Linear regression method for function approximation:**

$$\hat{y} = \sum_{n=1}^{N} g(\overset{\text{Query}}{\boldsymbol{x}}, \overset{\text{Key}}{\boldsymbol{x}_n})\, \overset{\text{Value}}{y_n}$$

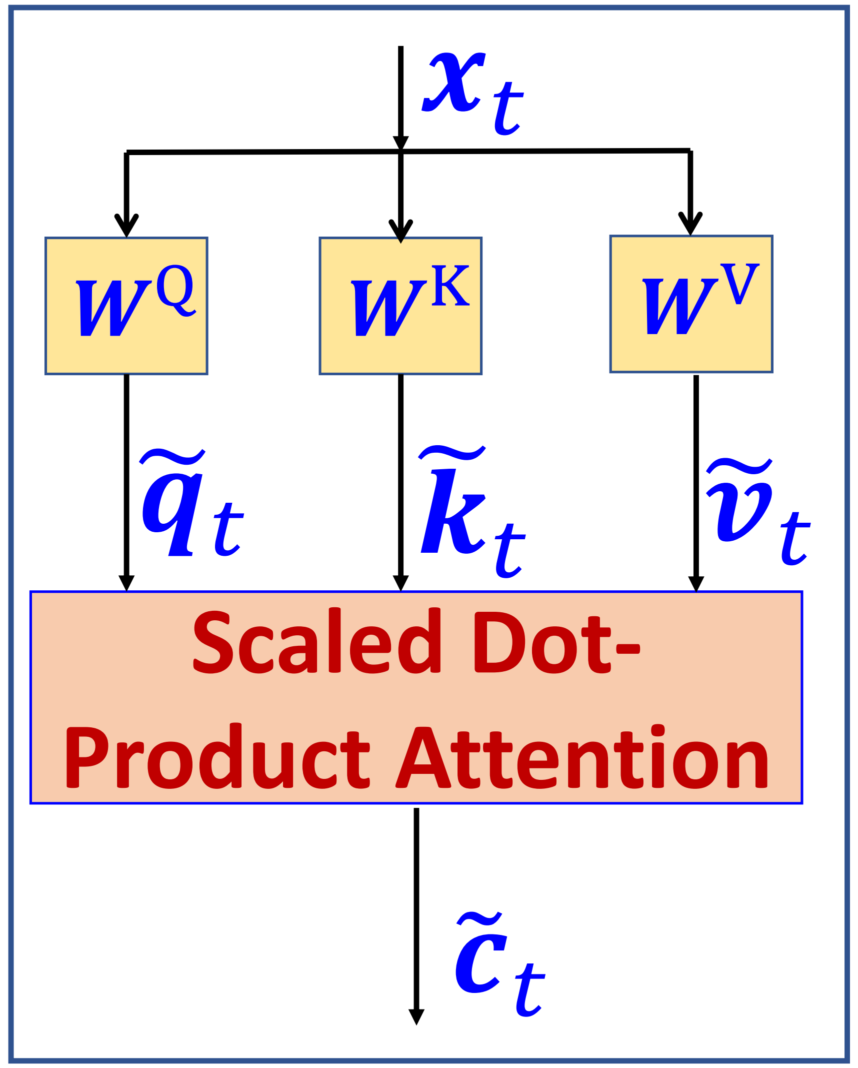
- **Here** $g(\boldsymbol{x}, \boldsymbol{x}_n)$ **is a basis function such as Gaussian.**

Scaled Dot-Product Attention (SDPA)

- **Consider the following d-dimensional vectors, with $t = 1, 2, \dots, T$:**
  - **Query vectors: $\boldsymbol{q}_t$**
  - **Key vectors: $\boldsymbol{k}_t$**
  - **Value vectors: $\boldsymbol{v}_t$**
- **Scaled dot-product between $\boldsymbol{q}_t$ and $\boldsymbol{k}_m$ is given by** $\alpha_{tm} = \frac{\langle \boldsymbol{q}_t, \boldsymbol{k}_m \rangle}{\sqrt{d}}$
- **Attention score:** $a_{tm} = \text{softmax}\,(\alpha_{tm}) = \frac{e^{\alpha_{tm}}}{\sum_{j=1}^{T} e^{\alpha_{tj}}}$
- **Context vector associated with Query vector $\boldsymbol{q}_t$: $\boldsymbol{c}_t = \sum_{m=1}^{T} a_{tm} \boldsymbol{v}_m$**
- $\boldsymbol{c}_t = \text{SDPA}\,(\boldsymbol{q}_t, \boldsymbol{K}, \boldsymbol{V})$ **where *K* and *V* are the matrices with Key vectors and Value vectors as their columns.**
- **The context vector captures the relation of $\boldsymbol{q}_t$ with the Key vectors and is obtained as a weighted combination of the corresponding Value vectors.**

Self-Attention and Single-Head Attention

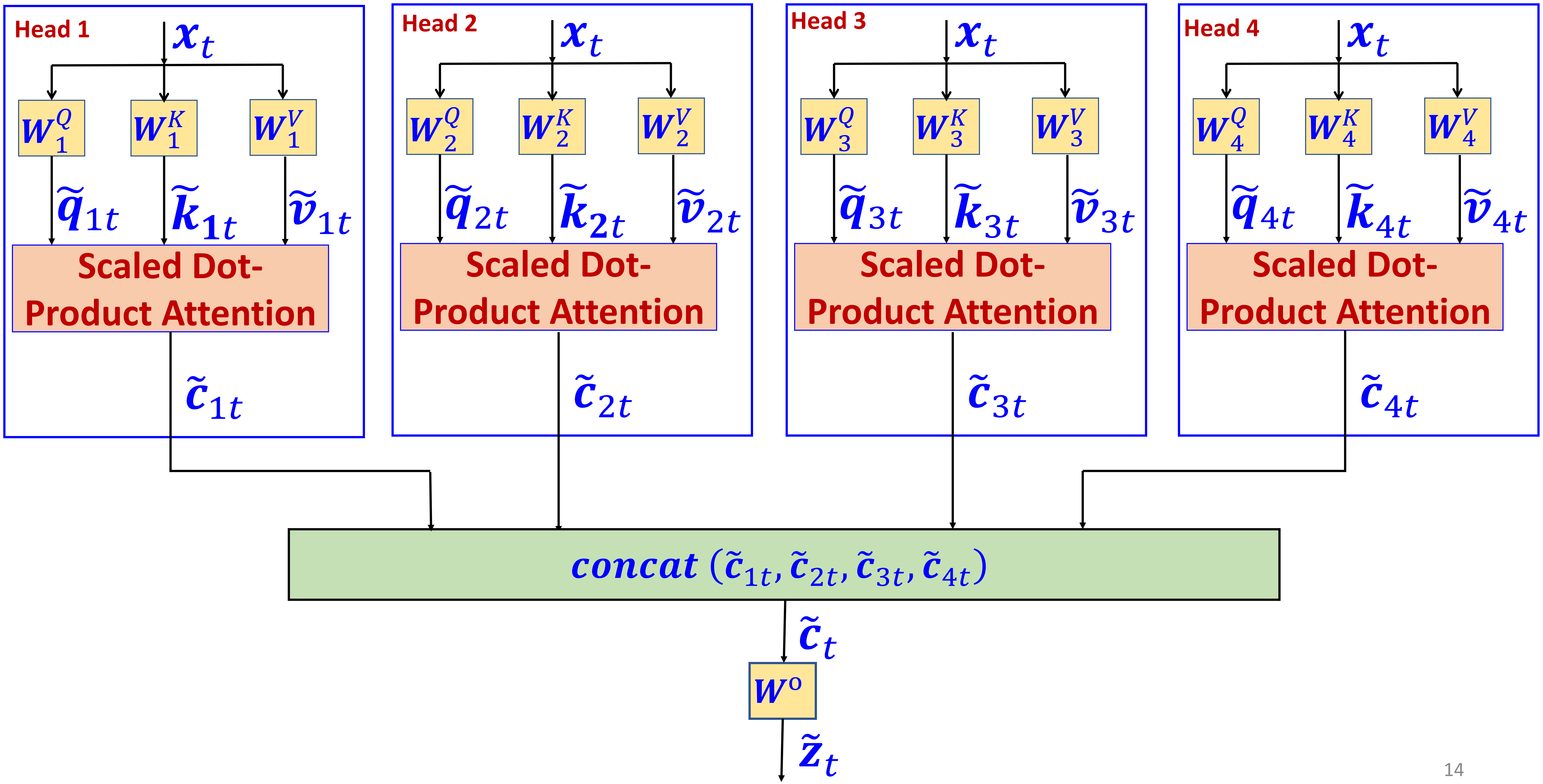
- **Self-attention captures the relations among the elements of a sequence**
- **Consider a sequence of d-dimensional T feature vectors: $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_T)$**
- **Query vectors, Key vectors and Value vectors are generated using the elements in the sequence *X*.**
- **Different weight matrices are used to transform an element $\boldsymbol{x}_t$ in the sequence *X*, to generate the corresponding query, key and value vectors as follows:**
  - **Query vector:** $\widetilde{\boldsymbol{q}}_t = \boldsymbol{W}^{\mathrm{Q}} \boldsymbol{x}_t$
  - **Key vector:** $\widetilde{\boldsymbol{k}}_t = \boldsymbol{W}^{\mathrm{K}} \boldsymbol{x}_t$
  - **Value vector:** $\widetilde{\boldsymbol{v}}_t = \boldsymbol{W}^{\mathrm{V}} \boldsymbol{x}_t$
- $\tilde{\boldsymbol{c}}_t = \mathbf{SDPA}\left(\widetilde{\boldsymbol{q}}_t, \widetilde{\boldsymbol{K}}, \widetilde{\boldsymbol{V}}\right)$
- **Transformation (weight) matrices $\boldsymbol{W}^{\mathrm{Q}}, \boldsymbol{W}^{\mathrm{K}}$ and $\boldsymbol{W}^{V}$, are of size lxd with $l < d$**
- **The context vector matrix generated from the sequence X is an lxT matrix, $\widetilde{C} = [\tilde{c}_t]_{t=1}^{T}$**
- **Single-Head Attention: Generation of context vector matrix from the sequence *X* using one set of transformation matrices**



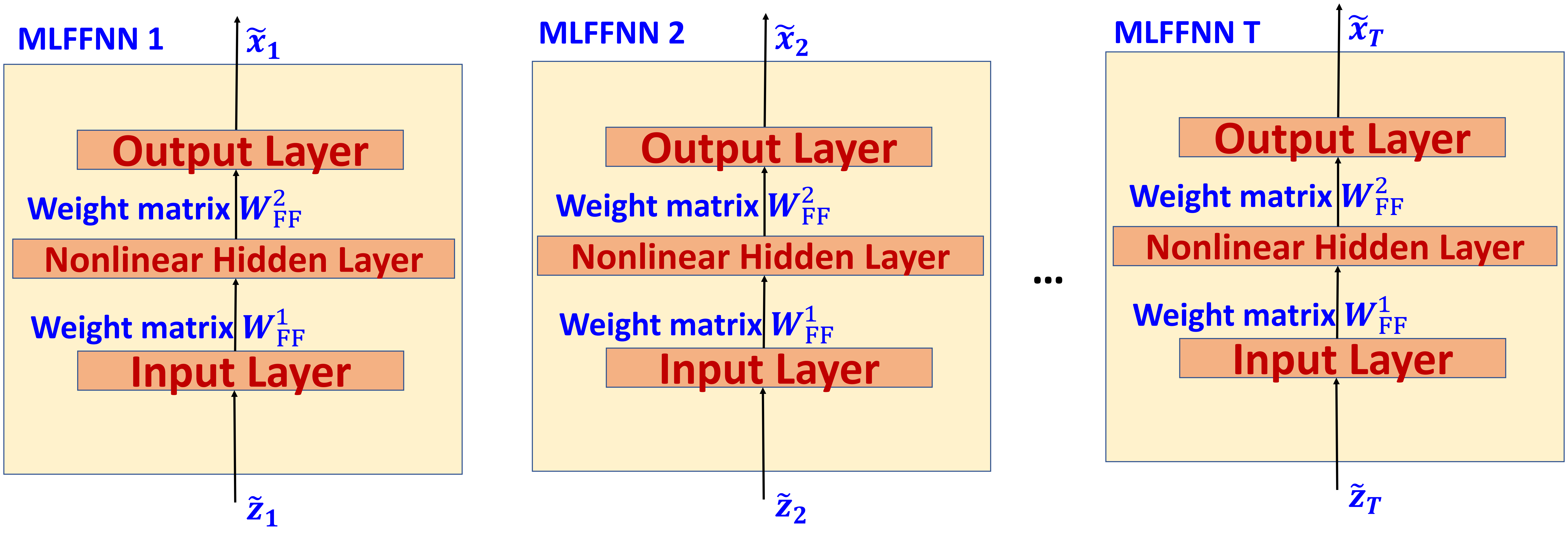
Self-Attention and Multi-Head Attention

- **Multi-Head Attention (MHA): Multiple sets of transformation matrices are used to generate multiple context vectors.**
- **Transformation matrices associated with the ith head:** W_i^Q, W_i^K, W_i^V
- **Query, Key and Value vectors generated using the ith head:**
  - **Query vector:** $\widetilde{\boldsymbol{q}}_{it} = \boldsymbol{W}_i^Q \boldsymbol{x}_t$
  - **Key vector:** $\widetilde{\boldsymbol{k}}_{it} = \boldsymbol{W}_i^K \boldsymbol{x}_t$
  - **Value vector:** $\widetilde{\boldsymbol{v}}_{it} = \boldsymbol{W}_i^V \boldsymbol{x}_t$
- **Context vector generated using the ith head:** $\tilde{\boldsymbol{c}}_{it} = \mathrm{SDPA}\left(\widetilde{\boldsymbol{q}}_{it}, \widetilde{\boldsymbol{K}}_i, \widetilde{\boldsymbol{V}}_i\right)$
- **Number of heads:** h
- **Dimension of query, key and value vectors:** $l = \frac{d}{h}$
- **Context vector in MHA is a d-dimensional vector:** $\tilde{\boldsymbol{c}}_t = concat\,(\tilde{\boldsymbol{c}}_{1t}, \tilde{\boldsymbol{c}}_{2t}, \ldots, \tilde{\boldsymbol{c}}_{ht})$
- **Context vector is transformed using the dxd matrix, W^{o}, to generate the output vector:** $\tilde{\boldsymbol{z}}_t = W^{\mathrm{o}} \tilde{\boldsymbol{c}}_t$
- **Output of MHA is the sequence:** $Z = (\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \ldots, \tilde{\boldsymbol{z}}_T)$
- **Self-attention MHA is a sub-layer in the encoder layer of Transformer model**

Self-Attention and Multi-Head Attention

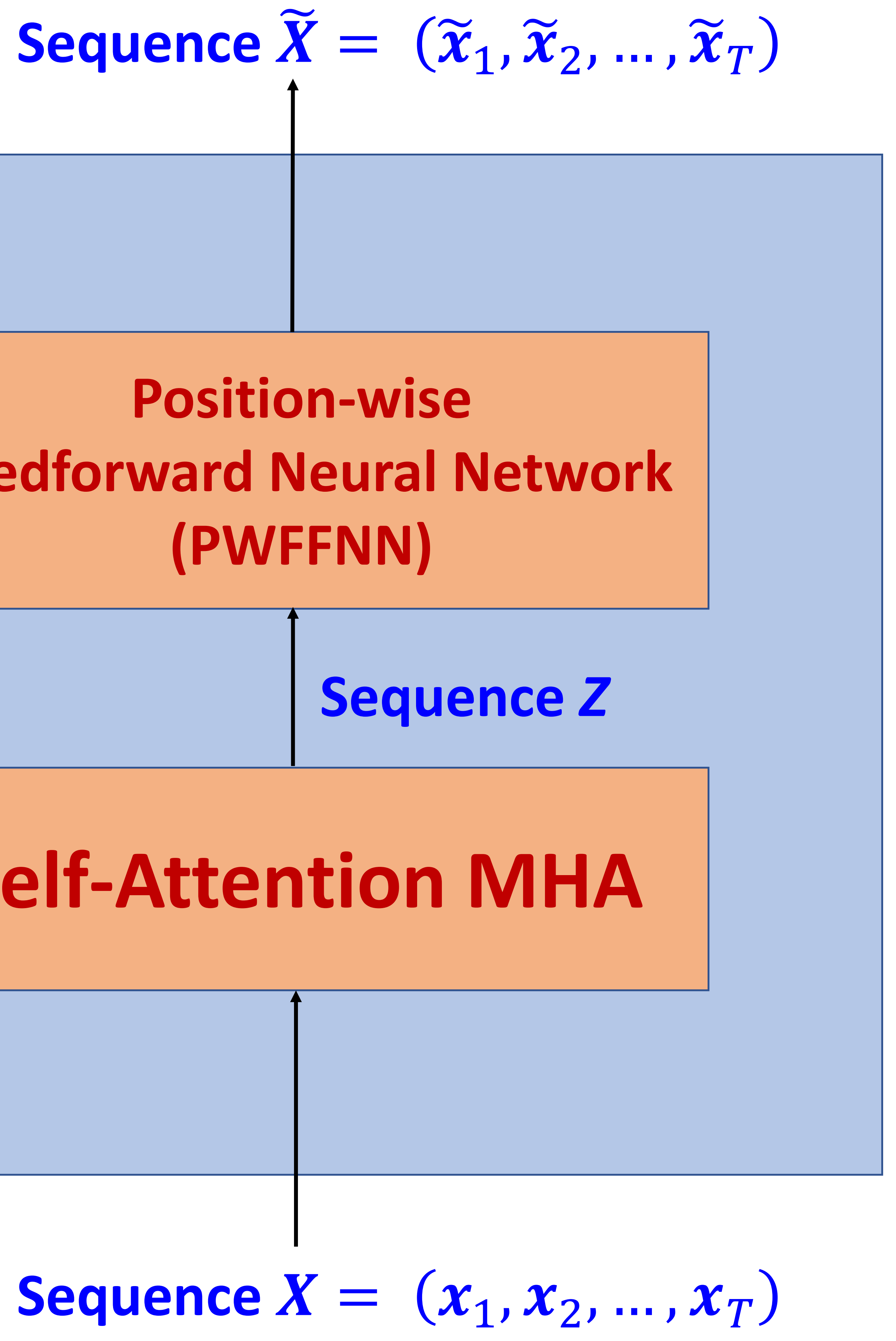


Position-Wise Feedforward Neural Network (PWFFNN)



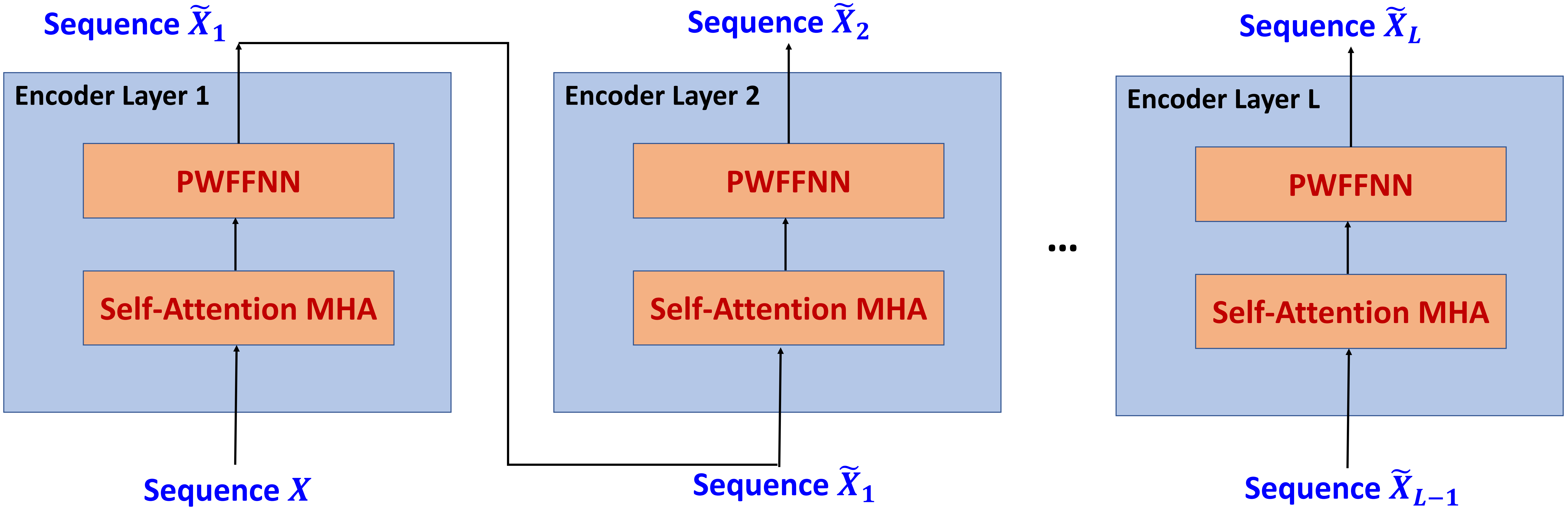
One Multilayer Feedforward Neural Network (MLFFNN) is used for every position t in the sequence. There are T MLFFNNs in the PWFFNN. The weight matrices are shared across the MLFFNNs in the PWFFNN.

Encoder Layer in Transformer Model

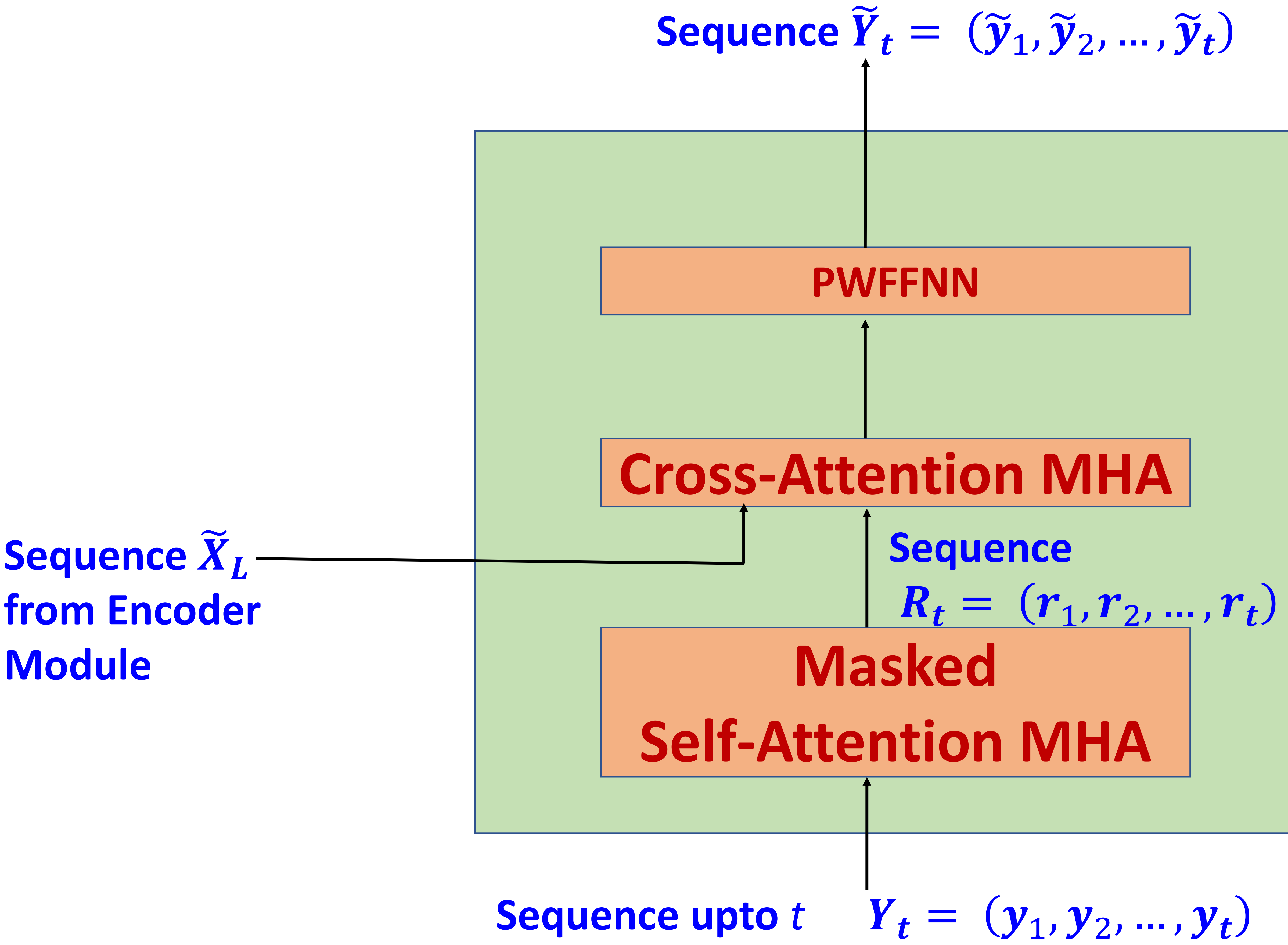


Encoder Module in Transformer Model

- **The Encoder module in the Transformer model includes several Encoder layers**
- **The encoder output is the sequence $\widetilde{X}_L$ obtained after several transformations on the sequence X**

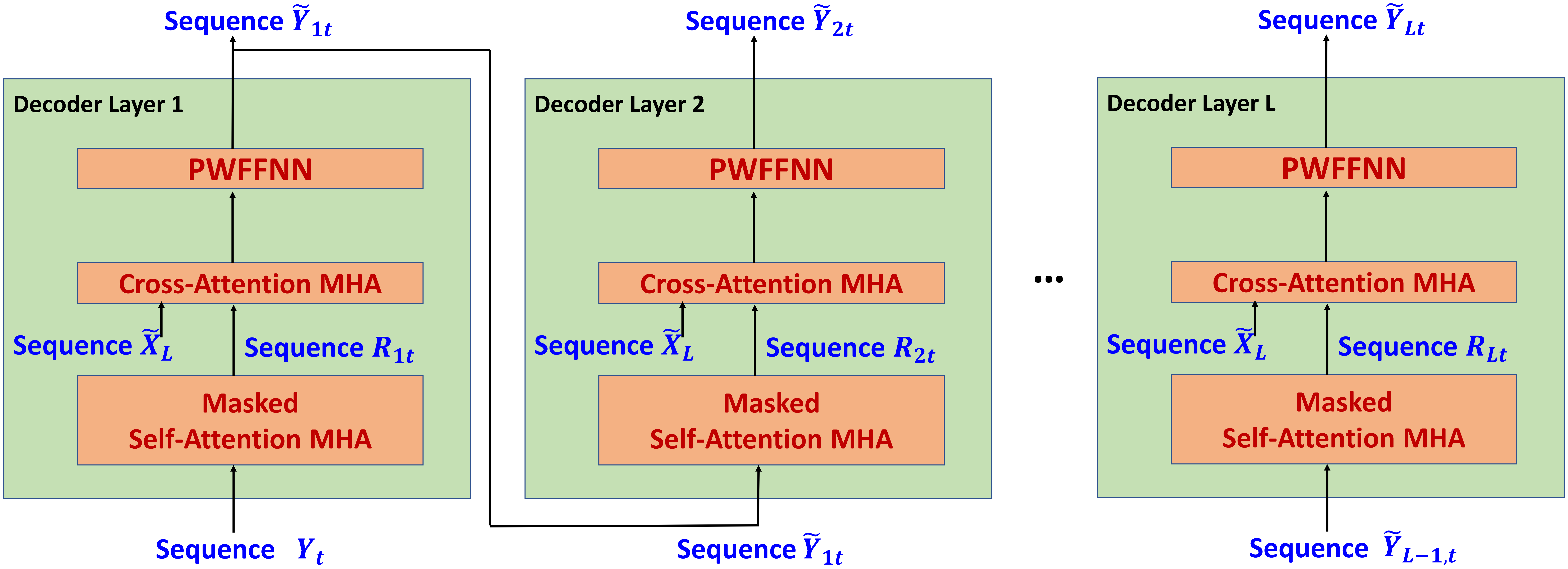


Decoder Layer in Transformer Model

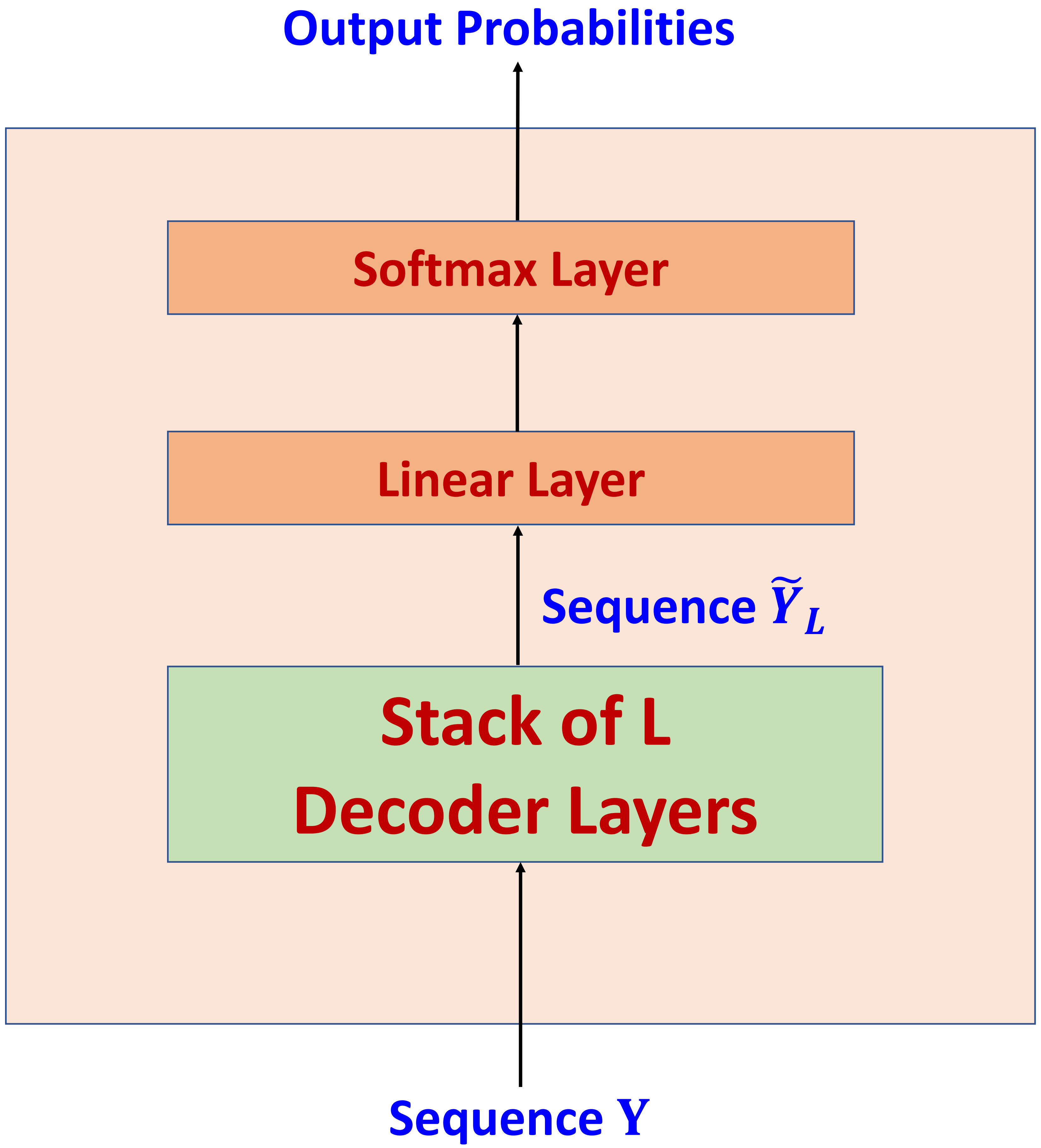


Cross-attention MHA is performed with the Key and Value vectors generated using Sequence $\widetilde{X}_L$, and the Query Vectors generated using Sequence Y_t

Stack of Decoder Layers in Transformer Model



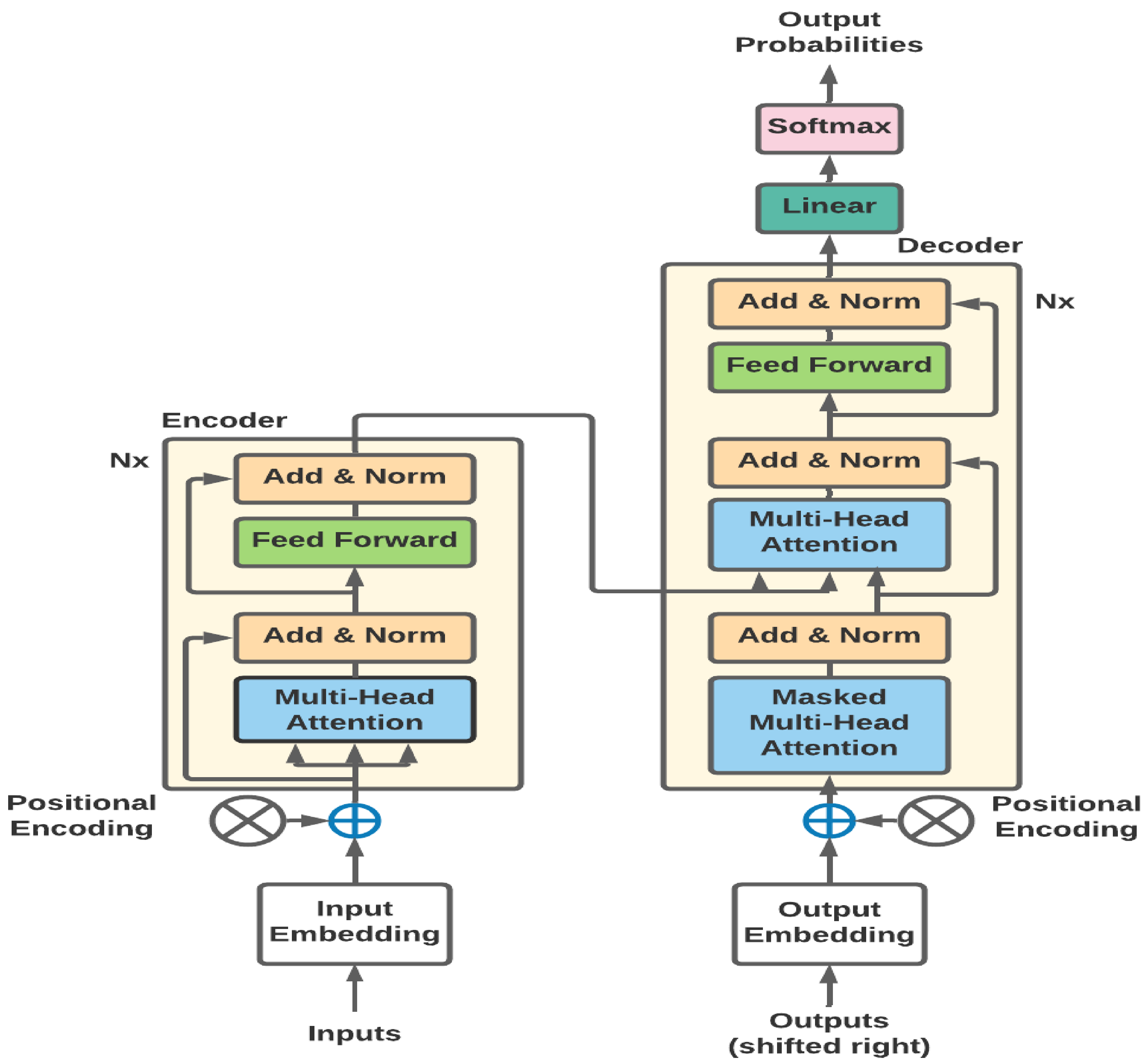
Decoder Module in Transformer Model



Training phase: Desired output sequence is given as the input to Decoder module

Testing phase: Output sequence generated up to time t is given as the input to Decoder module to predict the next element in the output sequence

Attention-based Model: Transformer



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," NIPS, 2017.

Sequence-to-Sequence Mapping Tasks

- **Neural Machine Translation: Translation of a sentence in the source language to a sentence in the target language**
  - **Input: A sequence of words**
  - **Output: A sequence of words**
- **Video Captioning: Generation of a sentence as the caption for a video represented as a sequence of frames**
  - **Input: A sequence of feature vectors extracted from the frames of a video**
  - **Output: A sequence of words**
- **Each of the above tasks involves mapping an input sequence to an output sequence**

Vision Transformer (ViT) for Image Classification

Representation of an image using transformer encoder in ViT:

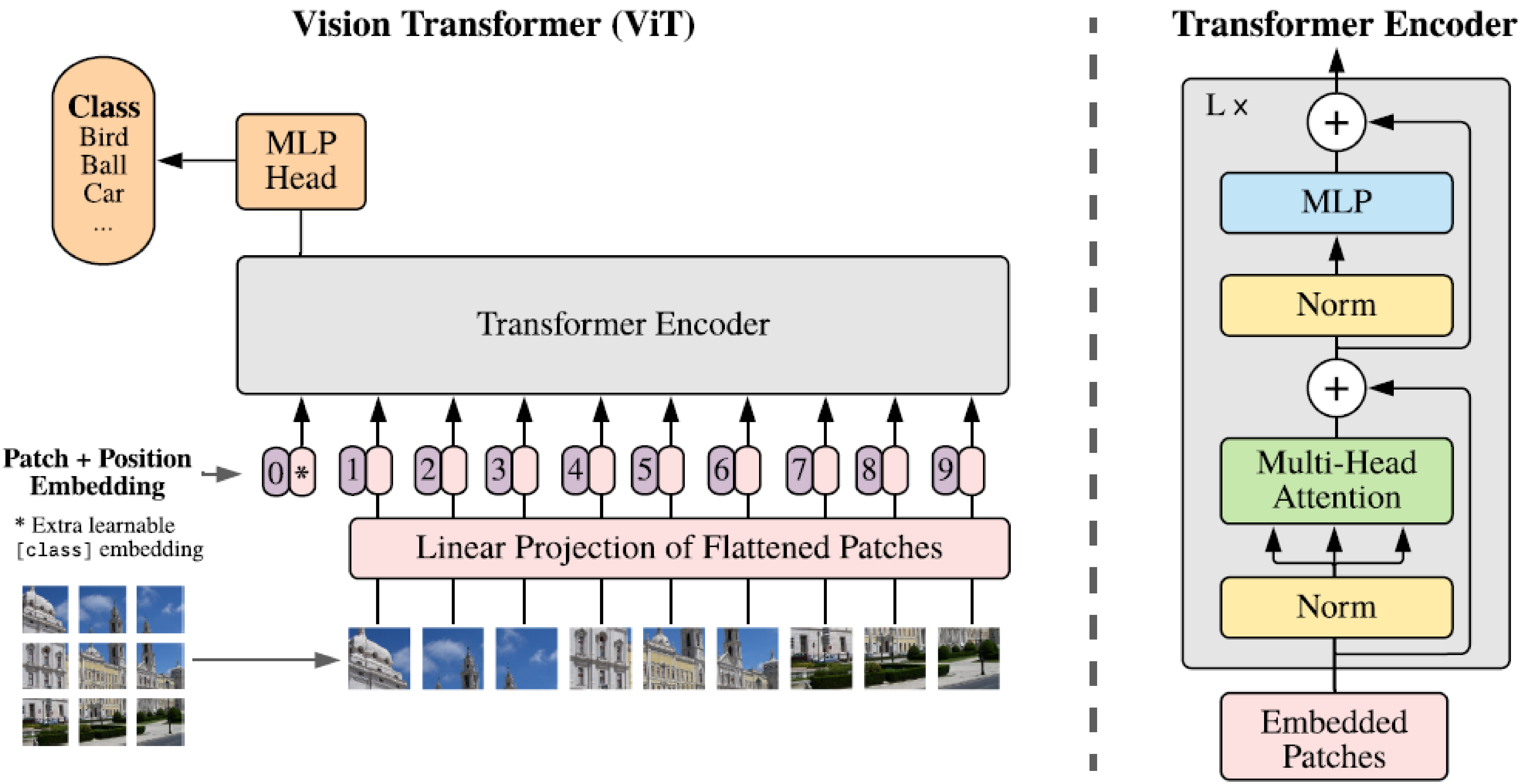
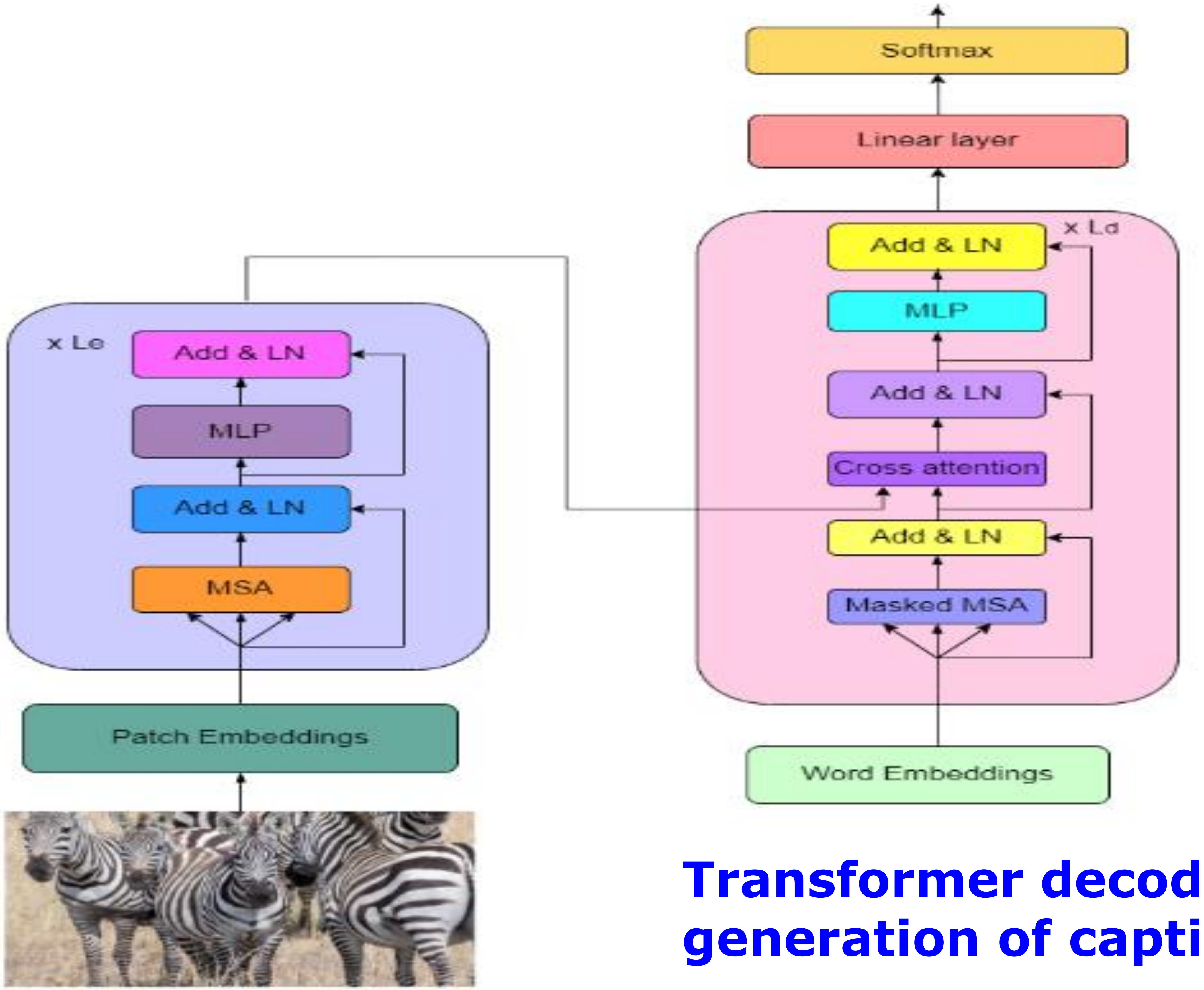


Image Captioning using Vision Transformer



Transformer decoder for generation of caption

Transformer encoder in ViT for representation of image

Pre-training of Transformer

Encoder and/or decoder of transformer can be pre-trained using huge amount of unlabeled data, and then fine-tuned using small amount of labeled data for a downstream task.

- **Encoder pre-training for text data**
  - **Bidirectional Encoder Representation from Transformer (BERT)**
- **Encoder pre-training for visual-linguistic data**
  - **Vision-and-Language BERT (ViLBERT)**
- **Decoder pre-training for text data**
  - **Generative Pre-trained Transformer (GPT)**

Bidirectional Encoder Representation from Transformer (BERT)

- **Pre-train the generic representation for several Natural Language Processing (NLP) tasks**

- **Pre-training Methods:**
  - **Masked Language Modelling (Mask LM)**
  - **Next Sentence Prediction (NSP)**

- **Fine-tuned for tasks such as**
  - **Sentence classification**
  - **Sentence relationship**
  - **Textual question answering**

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.

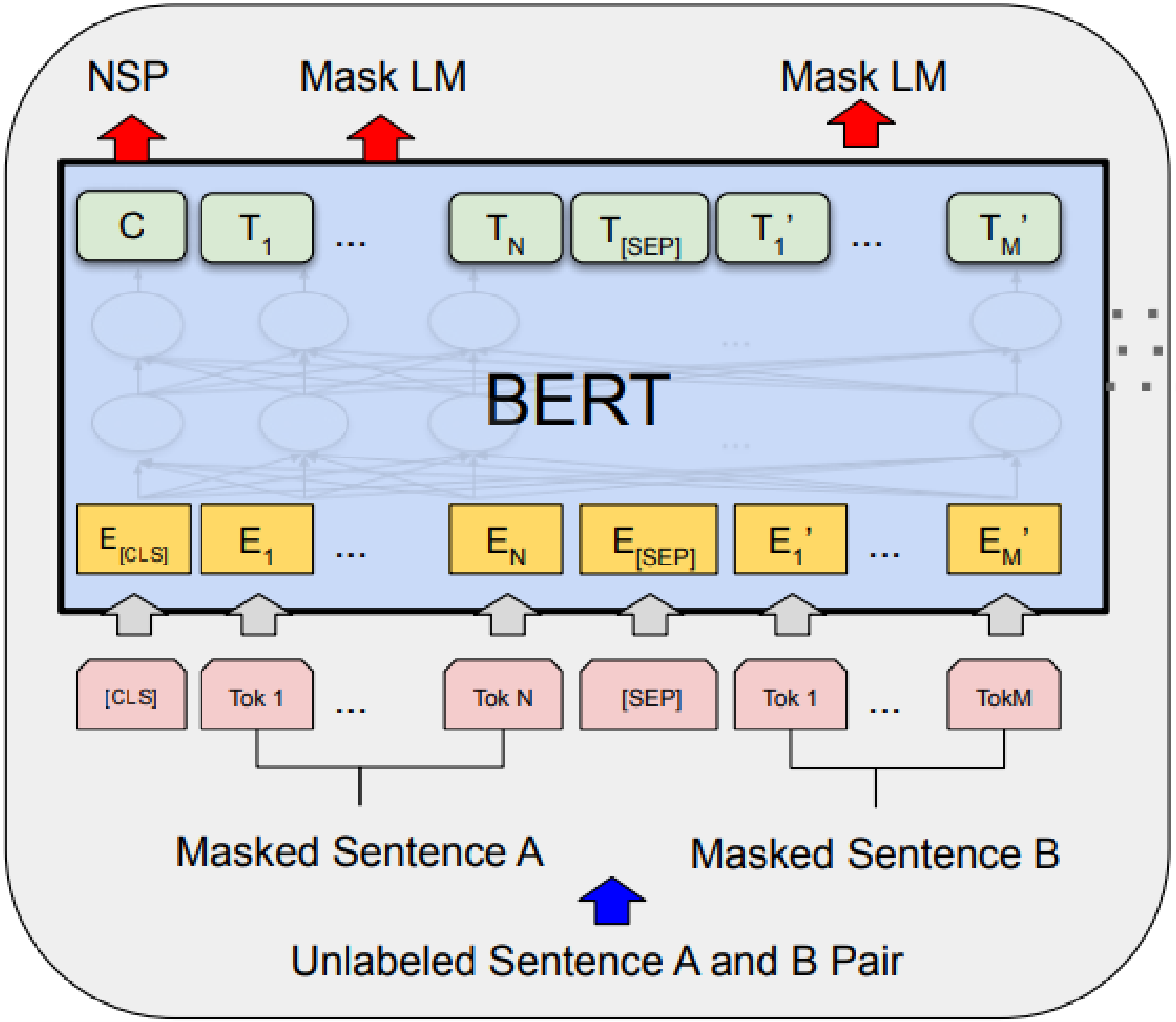
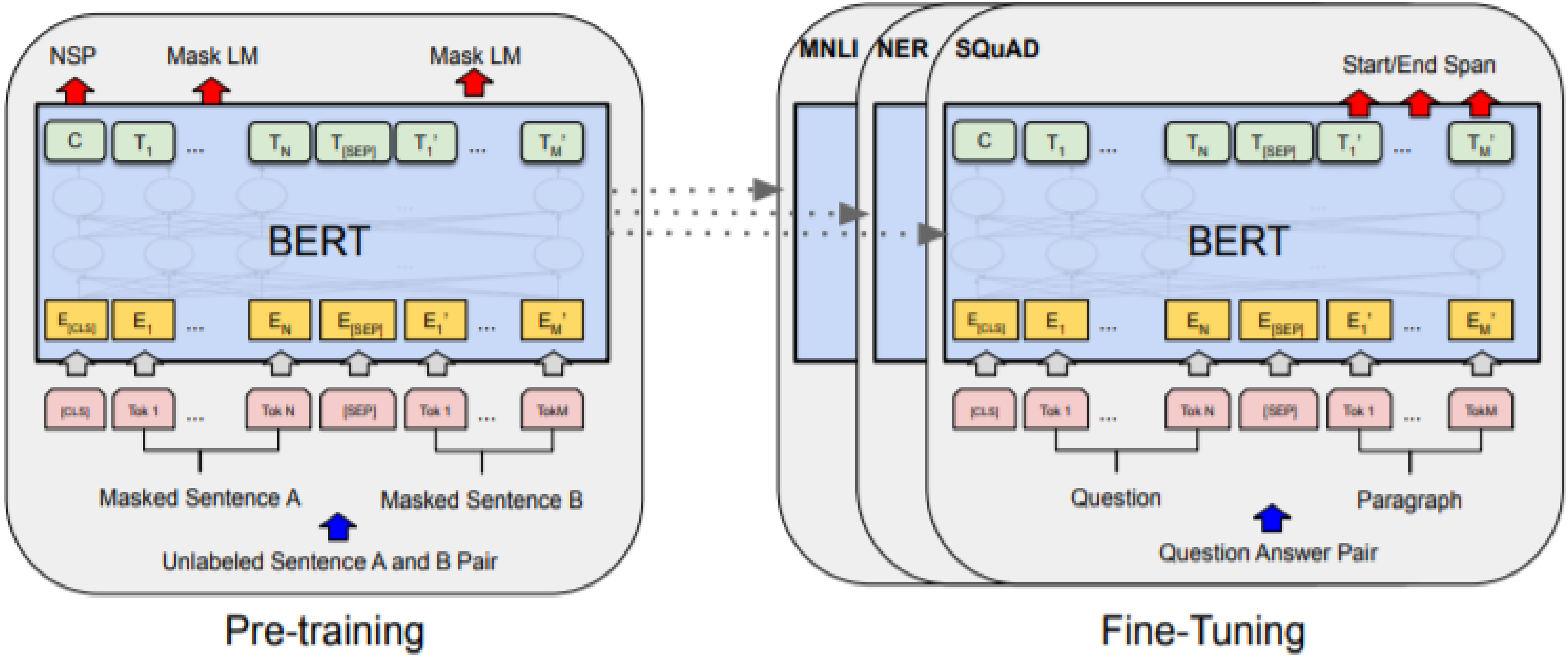


Image source : BERT(Devlin et al., 2019)

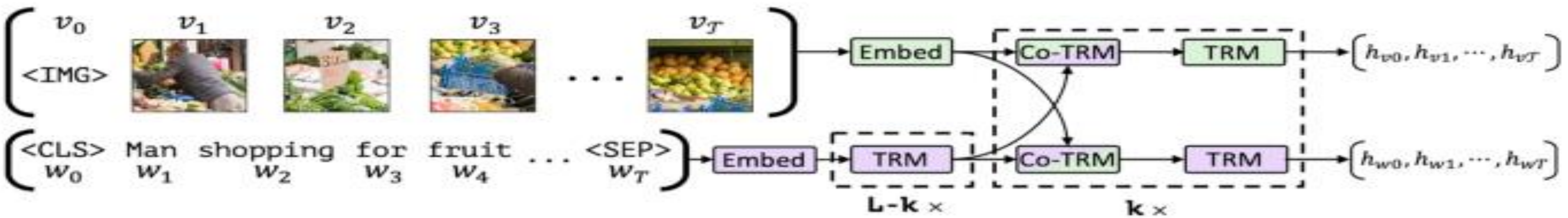
Pre-Training and Fine-Tuning using BERT



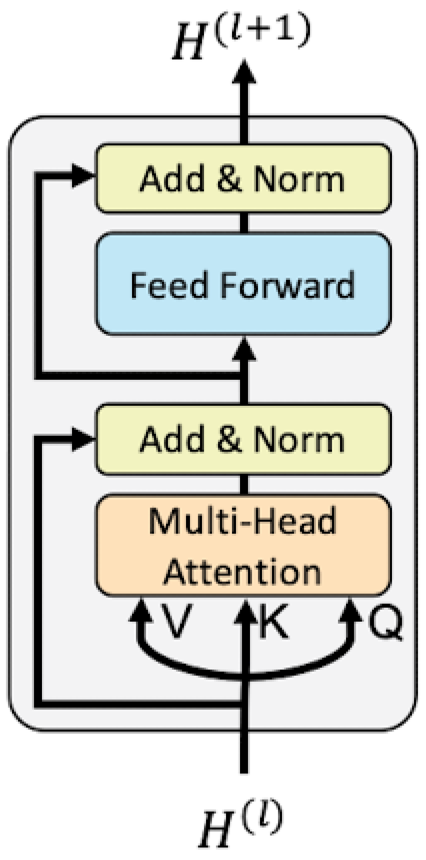
- **Fine-tuning for downstream tasks:**
  - **Textual question answering on the Stanford Question Answering Dataset (SQuAD)**
  - **Named Entity Recognition (NER)**
  - **Multi-Genre Natural Language Inference (MNLI)**

Image source : BERT(Devlin et al., 2019)

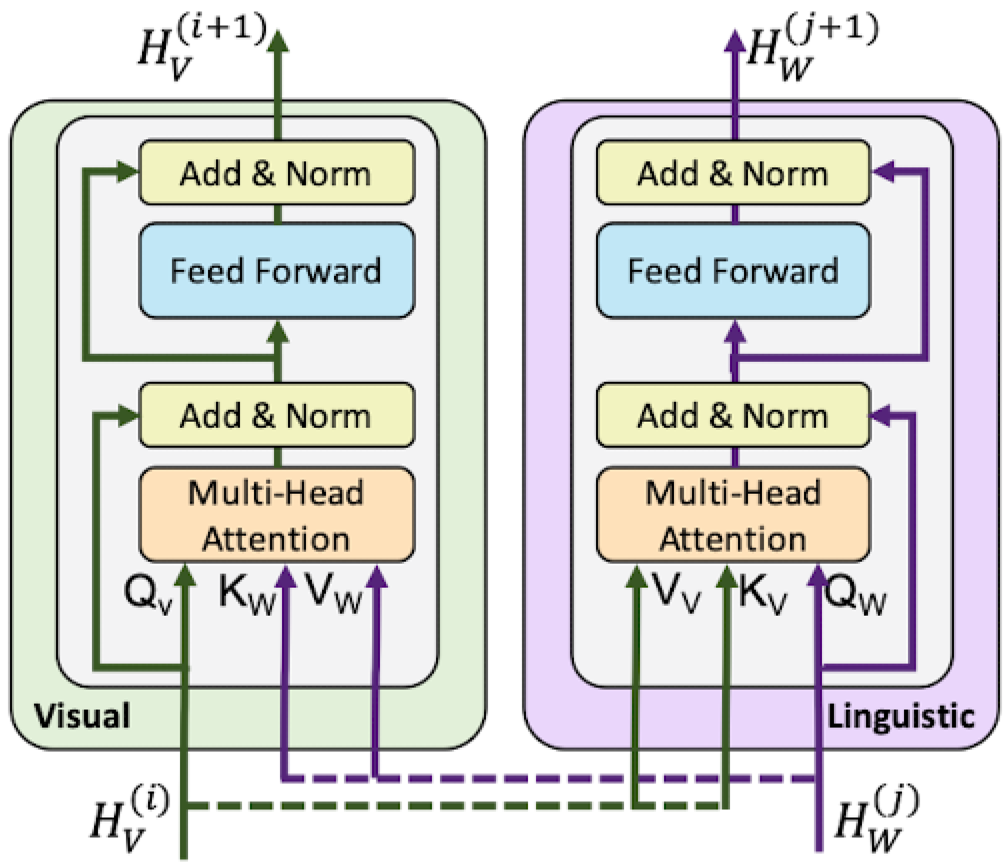
Vision-and-Language BERT (ViLBERT)



- **TRM: Transformer encoder**
- **Co-TRM: Co-attention transformer layer**



Transformer encoder



Co-attention transformer layer

J.Lu, D.Batra, D.Parikh and S.Lee, , "ViLBERT: Pre-training Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, ," NeurIPS, 2019.

Generative Pre-trained Transformer (GPT)

- **Transformer decoder is pre-trained using unlabeled text data**
- **GPT can be fine-tuned for downstream tasks that involve text data**
- **Auto-regressive model: A word in a sentence is predicted using all the words preceding that word in the sentence**
- **Masked multi-head self-attention (MSA) in each layer of transformer decoder takes the sequence of words preceding a word in a sentence.**
- **The decoder is trained to predict the next word in the sentence.**
- **GPT-1, GPT-2 and GPT-3: Pre-trained models with different number of layers trained with different corpora for different pre-training tasks**

A.Redford, K.Narasimhan, T.Salimans and I.Sutskever , "Improving Language Understanding by Generative Pre-training," 2018

A.Redford, J.Wu, R.Child, D.Luan, D.Amodei and I.Sutskever, "Language Models are Unsupervised Multitask Learners," 2019

T.Brown et al., "Language Models are Few-Shot Learners," arXiv:2005.14165v4, 22nd July, 2020

Visual Question Answering (VQA) for Images

Is there something to cut the vegetables with?

Yes

No

Who is wearing glasses?

Man

Woman

How many children are in the bed?

Two

One

Image VQA Frameworks

Image

Image Encoder

Representation of Image

Question

How many children are in the bed?

Question (Text) Encoder

Representation Question

Fusion of Represen-tations

Answer Generator

Answer

Two

Image Encoder: CNN, ViT Encoder, Swin Tranformer

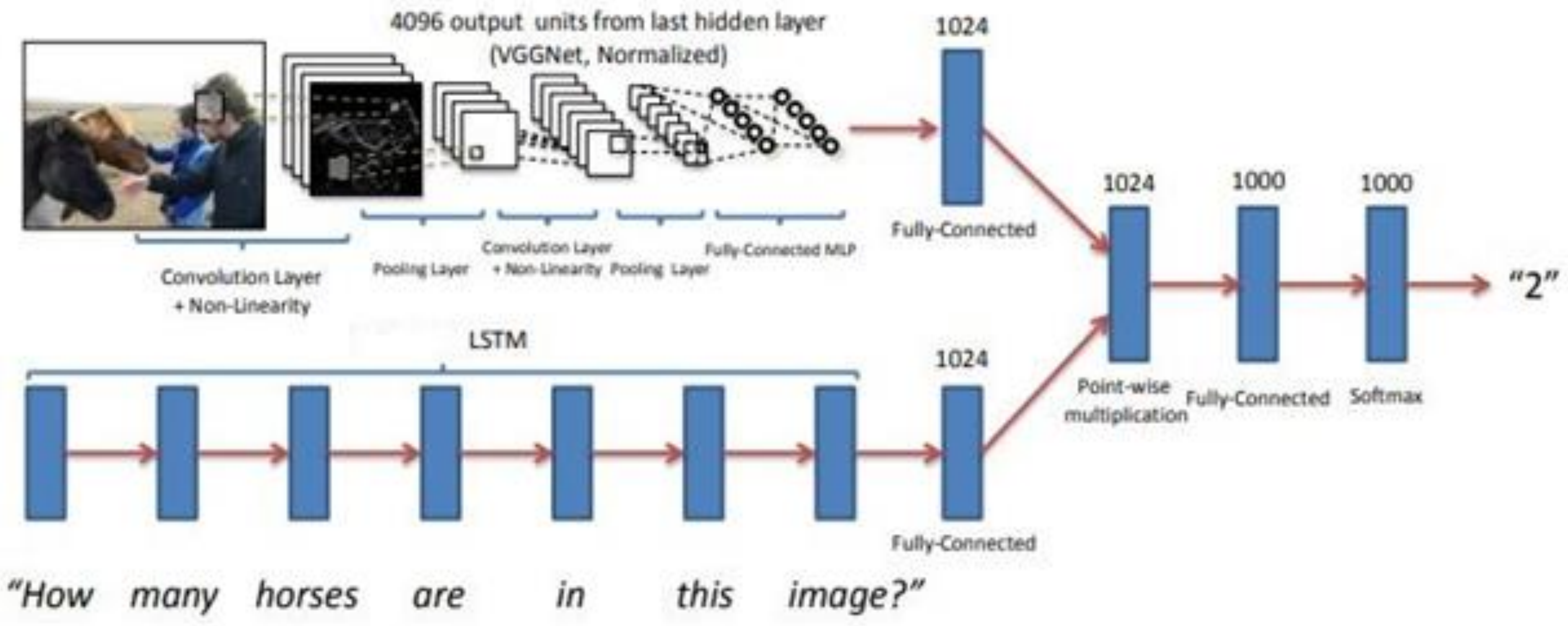
Question Encoder: LSTM, Transformer encoder, BERT fine-tuned with questions in VQA dataset

Fusion of Representations: Concatenation, Co-attention transformer

Answer Generator: Classifier, Text generator such as GPT fine-tuned with answers in VQA dataset

VQA for Images

- **CNN based encoder for image**
- **LSTM based encoder for question**
- **Answering: Multi-class classification with k classes for k possible answers**



A.Agrawal, J.Lu, S.Antol, M.Michell, C.L.Zitnick, D.Batra and D.Parikh, "VQA: Visual Question Answering," arXiv:1505.00468v7, 27th October, 2016.

Image VQA Frameworks

Image

Question

How many children are in the bed?

ViLBERT based Encoder of Image and Question

Answer Generator

Answer

Two

Encoder: ViLBERT fine-tuned with image and questions in VQA dataset

Answer Generator: Classifier, Text generator such as GPT fine-tuned with answers in VQA dataset

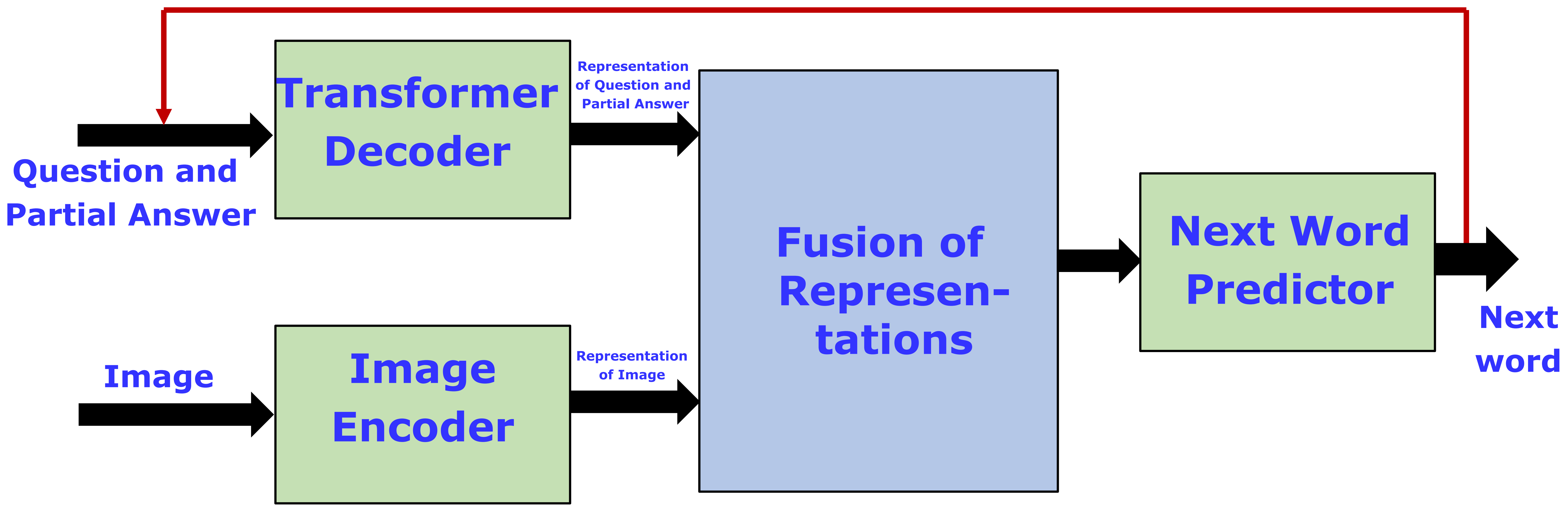
J.Lu, D.Batra, D.Parikh and S.Lee, , "ViLBERT: Pre-training Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, ," NeurIPS, 2019.

Open Ended VQA

Question - What is the Zebra doing?
Traditional VQA - Eating, Grazing
Open Ended VQA - The Zebra is grazing in grasslands

Question - What is in the dog's mouth?
Traditional VQA - Toy, Purple toy
Open Ended VQA - The dog is playing with a toy in its mouth.

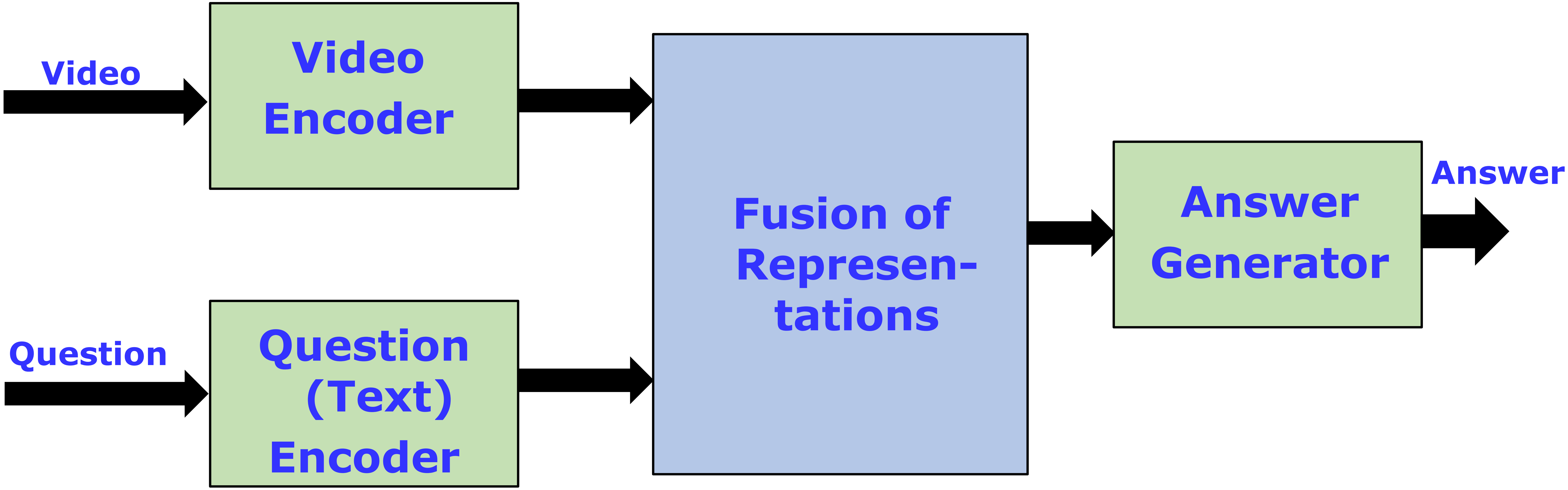
Open Ended VQA



In open ended VQA, the answer is a sequence of words. The system generates one word of the answer at a time. The next word in the answer is predicted using the representations of image, question, and the partial answer corresponding to the sequence of words generated so far.

A.M.Bellini, N.Parde, M.Matteucci and M.J.Carman, "Towards Open-Ended VQA Models using Transformers, " EMNLP, 2020.

Video VQA Frameworks



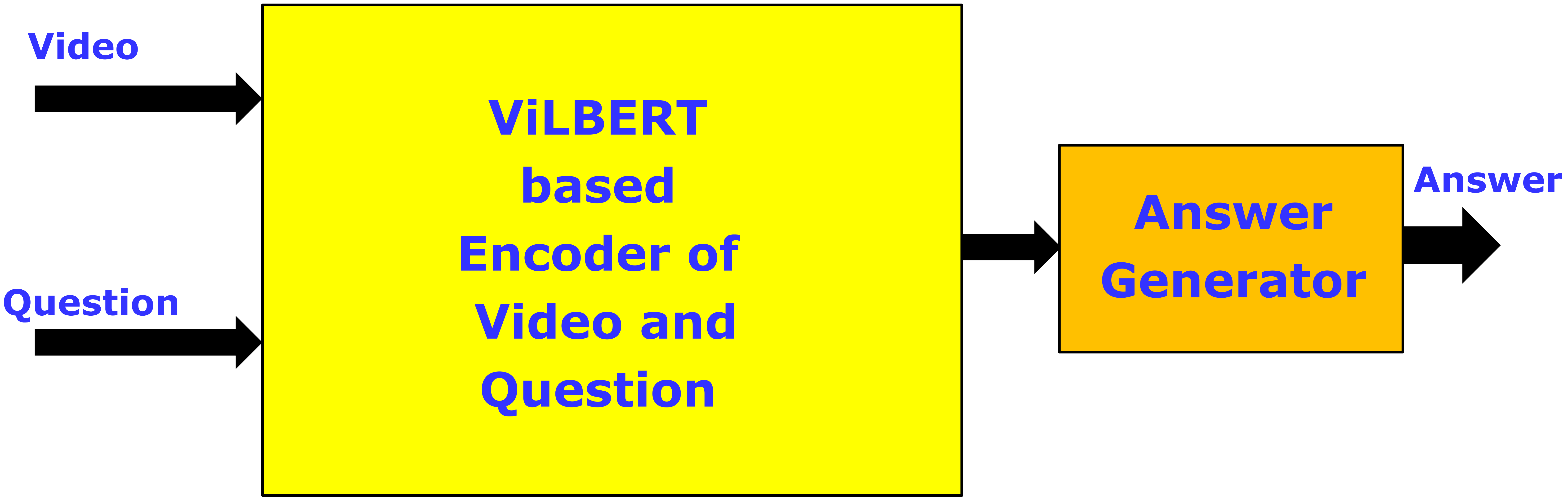
Video Encoder: RCNN, LSTM, Tranformer

Question Encoder: LSTM, Transformer encoder, BERT fine-tuned with questions in VQA dataset

Fusion of Representations: Concatenation, Co-attention transformer

Answer Generator: Classifier, Text generator such as GPT fine-tuned with answers in VQA dataset

Video VQA Frameworks



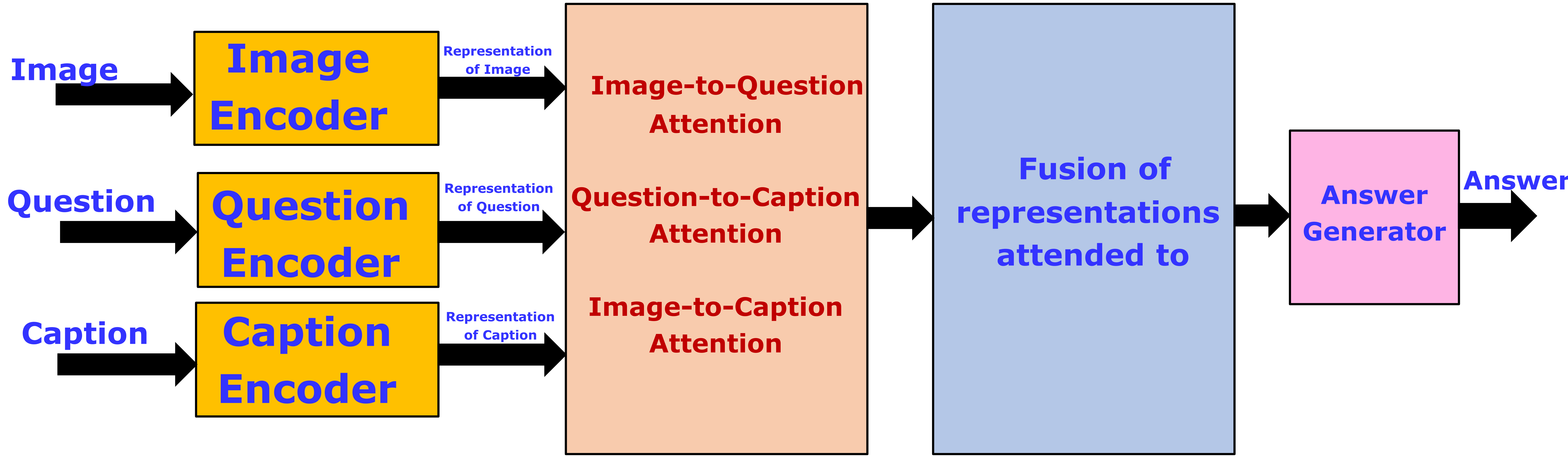
Encoder: ViLBERT fine-tuned with videos and questions in VQA dataset

Answer Generator: Text generator such as GPT fine-tuned with answers in VQA dataset

Z.Yang, N.Garcia, C.Chu, M.Otani, Y.Nakashima and H.Takemura , "BERT Representations for Video Question Answering," WACV, 2020.

Z.Yang, N.Garcia, C.Chu, M.Otani, Y.Nakashima and H.Takemura , "A Comparative Study of Language Transformers for Video Question Answering," Neurocomputing, vol.445, pp.121-133, 2021.

Image VQA using Caption



Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, Jiebo Luo, "VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions, " ECCV, 2018.

Visual Commonsense Reasoning

1. Where is this happening?

a) There's a conference in this room.	27.2%
b) This is happening in a fancy restaurant.	18.8%
c) This is a wedding.	52.8%
d) This is happening in an industrial zone.	1.3%

Options for answer

I think so because...

a) You can see they are in a restaurant by the other tables, and you can tell it is a fancy restaurant by the wine in a bucket on his table. [person1] is obviously happy and is focused across his table.	8.9%
b) This is a formal setting and everyone is dressed nicely.	1.6%
c) [person1] is dressed fancily and the background is fancy.	88.7%
d) Drinking while in a car is illegal, and some restaurants have strange seating to draw in customers.	0.8%

Options for rationale

R.Zellers, Y.Bisk, A.Farhadi and Y.Choi,, "From Recognition to Cognition: Visual Common Sense Reasoning," CVPR , 2018.

Z.Li, Y.Guo, K.Wang, Y.Wei, L.Nie and M.Kankanhalli, "Joint Answering and Explanation for Visual Commonsense Reasoning," arXiv: 2202.12626v1, 25 February, 2022.

J.Y.Lee and I.Kim, "Vision-Language-Knowledge Co-Embedding for Visual Commonsense Reasoning," Sensors, 2021.

Sub-Tasks in Visual Commonsense Reasoning

- **Visual input to VCR System: Image or Video**
- **Question: Q**
- **Answer: A**
- **Rationale (Reason): R**

Sub-tasks

- **Answering: Q $\longrightarrow$ A**
- **Justification: QA $\longrightarrow$ R**
- **Answering and Justification: Q $\longrightarrow$ AR**

Types of Sub-tasks

- **Multiple choices for Answer and Rationale: Answering and Justification are considered as classification tasks**
- **Generation of Answer and Rationale: Open ended answering and justification requires the ability of natural language generation**

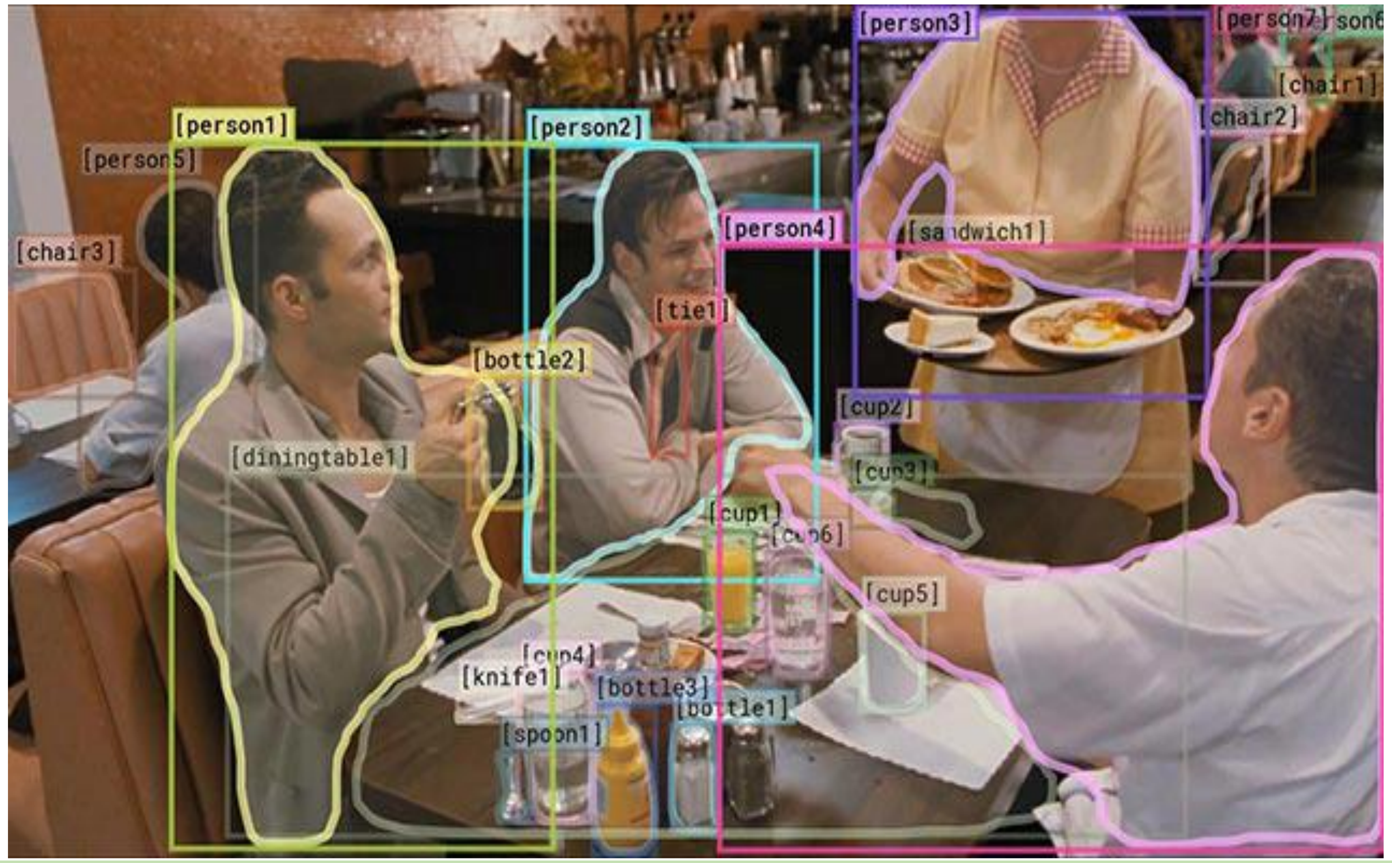
From Recognition to Cognition: Visual Commonsense Reasoning

Image and Regions of Interest → **CNN** → **Grounding Using LSTMs**

Question → **BERT** → **Grounding Using LSTMs**

Answer → **BERT**

Grounding Using LSTMs → **Contextualization of Question, Response and Image Regions** → **Reasoning using LSTM**



Question:
Why is Person 4 pointing at Person 1?

Answer:
He is telling Person3 that Person 1 ordered pancakes

R.Zellers, Y.Bisk, A.Farhadi and Y.Choi,, "From Recognition to Cognition: Visual Commonsense Reasoning," CVPR , 2018.

From Recognition to Cognition: Visual Commonsense Reasoning

- **Image processing to detect objects of interest, identify regions of objects (bounding boxes) and assign tags to objects**
- **Question or Query: A mix of natural language and pointing. Each word in the query is a either a word in a vocabulary or a tag referring to an object**
- **Answer: A mix of natural language and pointing.**
- **Rationale (Reason): A mix of natural language and pointing.**
- **Grounding query and answer: Learning a joint image-language representation for each word in the query/answer with relevant objects in the image.**
- **Contextualization: Use attention mechanisms to contextualize the query and answer with respect to each other, and the image. For each word in query, compute the attention score with respect to each word in answer. Perform attention between query and objects in the image, and between answer and objects in the image.**
- **Reasoning: The contextualized representation of objects in the image, query and answer is given as input to an LSTM trained to choose a rationale option.**

R.Zellers, Y.Bisk, A.Farhadi and Y.Choi,, "From Recognition to Cognition: Visual Commonsense Reasoning," CVPR , 2018.

Generative Models

- Models capable of generation of data (Text, Image, Video, Music)
- Restricted Boltzmann machine (RBM)
- Variational autoencoder
- Generative pre-trained transformer (GPT)
    - Large Language Models (LLMs)
- Generative adversarial network (GAN)
- Diffusion models
    - Text-to-image
    - Text-to-video
    - Text-to-audio
    - Text-to-music

LLMs: Evolution of GPT Models

1 GPT-1

- Released: 2018
- 117 million parameters
- Trained using unsupervised learning
- Gave state of the art performance for LAMBADA.
- Gave Competitive performance for GLUE and SQuAD

2 GPT-2

- Released: 2019
- Improved version of GPT-1.(1.5 billion Params)
- Trained on larger dataset.
- Additional improvements like including a modified training objective, and a more efficient sampling algorithm for generating text.

3 GPT-3

- Released: 2020
- Improved version of GPT-2 (175 billion Params)
- Trained on larger dataset.
- Additional improvements like better training methodology (Gshard) and Few shot learning Capability.

4 GPT-3.5

- Released:2022
- Improved version of GPT-3
- Three variants , 1.3 B, 6B and 175 B params
- The main element is to eliminate toxic outputs to certain extent using RLHF (reinforcement learning with human feedback).

5 GPT-4

- Released: 2023
- Not sure if its a fine tuned version of previous model.
- 1 trillion Parameters
- It can take text and image as inputs and produces text outputs.
- RLHF is used in the training.

NLP Benchmarks:

LAMBADA: LAnguage Modeling Broadened to Account for Discourse Aspects

GLUE: General Language Understanding Evaluation

SQUaD: Stanford Question Answering Dataset