UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Multimodal Fusion in Transformer Neural Networks

By: Alan Devkota
Advisor: Dr. Xin Fu

ECE 6011_23545

Sept 29, 2023

Department of ECE, University of Houston
Houston, TX, USA

Outline

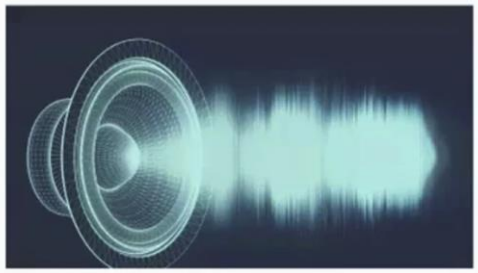
- Overview of Multimodality.
- Overview of Transformer Neural Networks.
- Attention Mechanism in Transformers.
- From Self-Attention to Cross-Attention
- Background of Vision Transformer
- Examples of Cross-modal Interactions
- Applications of Multi-modal Transformers
- Extension to higher Modalities
- Conclusion

Motivation: Our experience of the world is multimodal

- Humans perceive the world through multiple modalities, enabling a holistic understanding of their environment.
- Modality refers to the way in which information is captured or experienced.
- Modalities can include various sensory inputs like audio, vision (images and videos), text, and even less common ones like odors or touch.
- Multimodal fusion involves integrating information from different modalities to create a unified representation.
- For AI to match human intelligence, it's imperative that it learns to interpret, reason, and fuse multimodal information.

Text

Vision

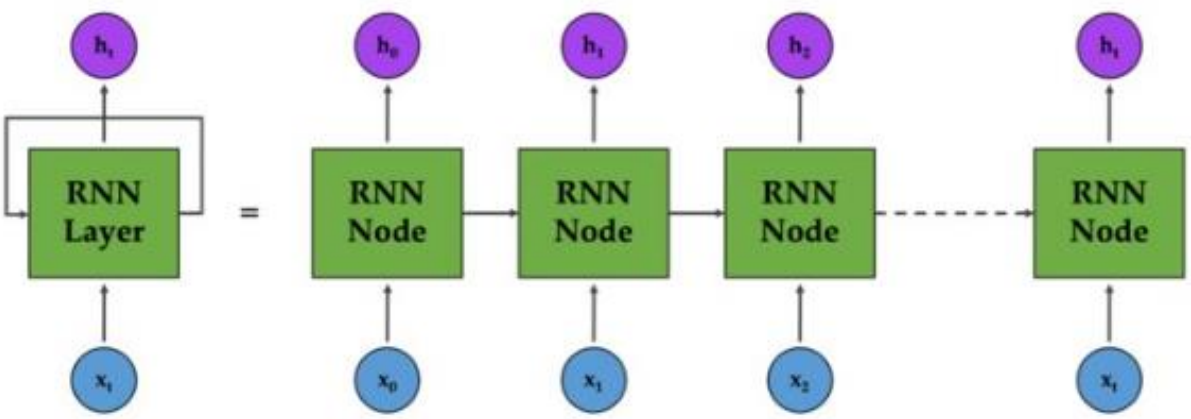


Audio

Overview of Transformer Neural Networks

➢ From RNNs to Transformers

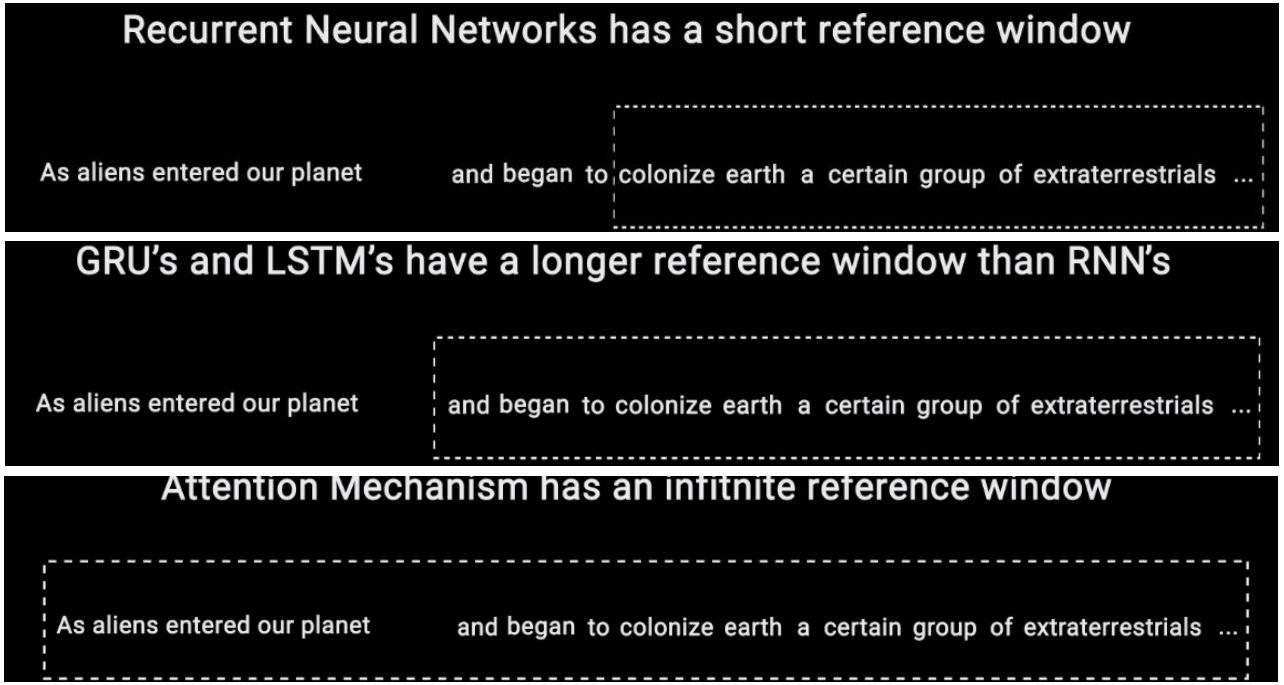
- RNN = Recurrent Neural Network
- Widely used in Natural Language Processing (until 2017)
- Processing text sequentially, token by token.

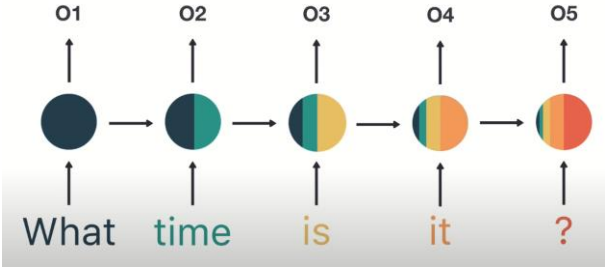


Overview of Transformer Neural Networks

➢ From RNNs to Transformers

- Why transformers?
  - RNN sequential processing – time consuming → Transformer – parallelization
  - RNN fail to obtain global information/ context → Transformers attend to every tokens (Because x0 is very far from xn)
  - RNN – single direction from left to right → Transformer both direction
  - Reference window is larger in Transformers.





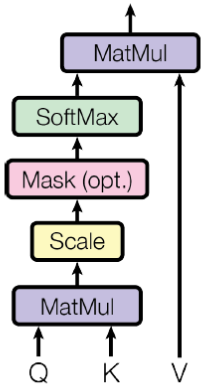
UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

A Transformer is Born

- Architecture of Transformer Neural Network

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

Scaled Dot-Product Attention



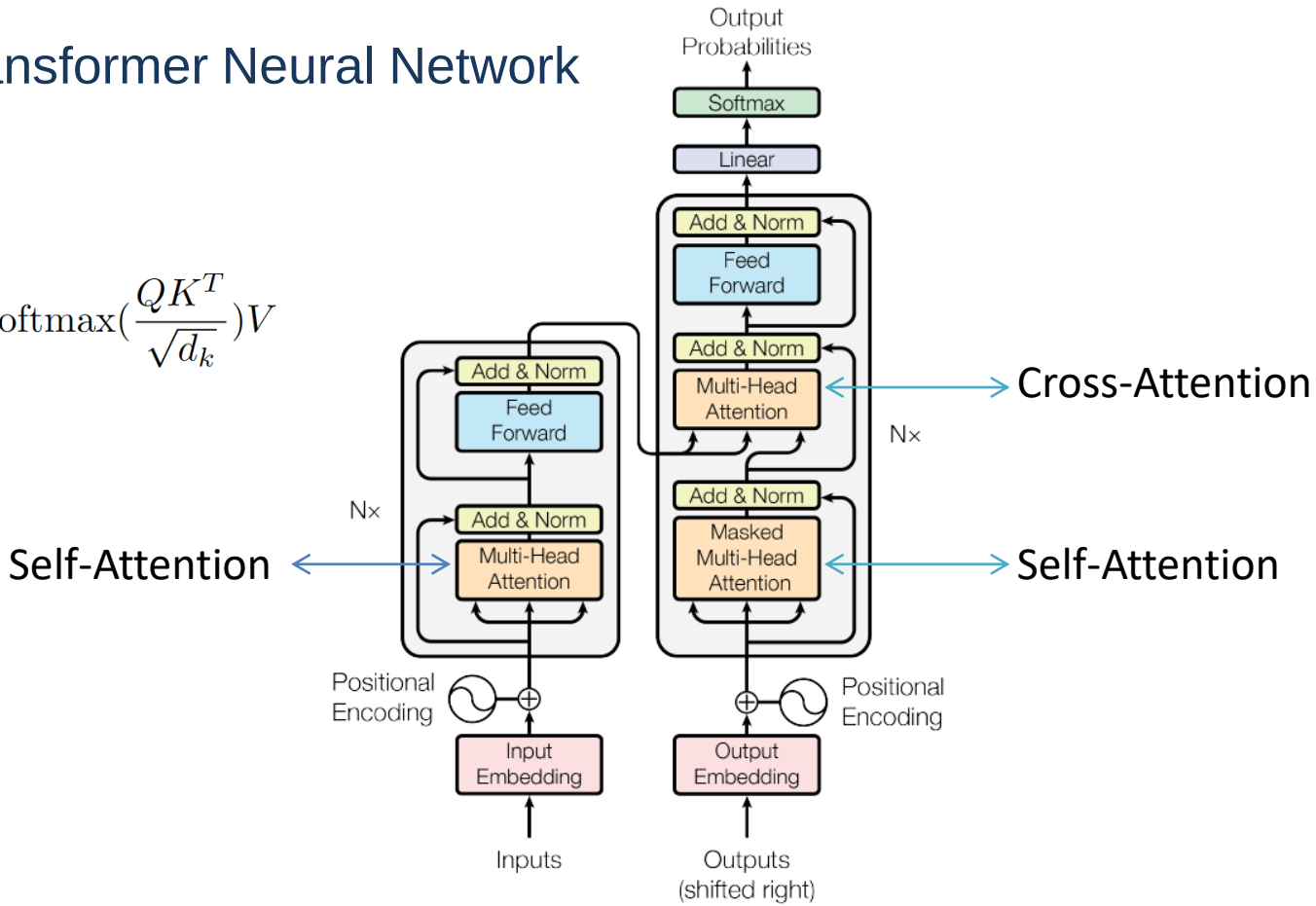


Figure 1: The Transformer - model architecture.

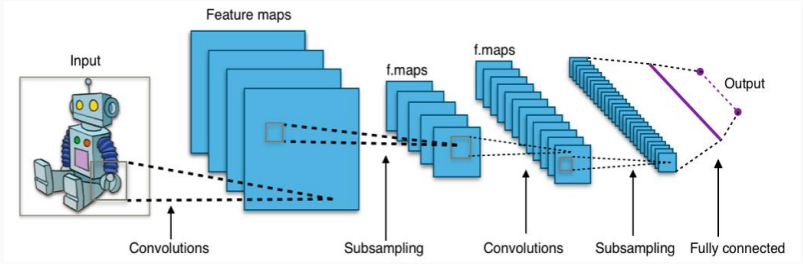
Attention is all you Need, Vaswani et al., 2017

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

A Transformer is Born

Computer Vision

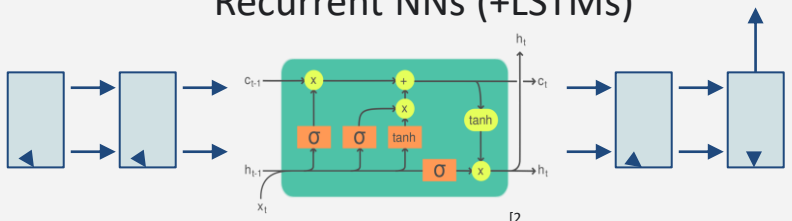
Convolutional NNs (+ResNets)



[1]

Natural Lang. Proc.

Recurrent NNs (+LSTMs)

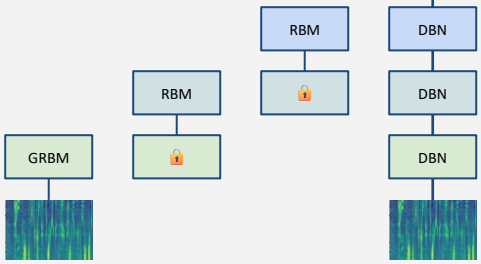


[2]

h e l l o

Speech

Deep Belief Nets (+non-DL)



Translation

Seq2Seq

s a l u t

h e l l o </s> s a l u t

RL

BC/GAIL

Algorithm 1 Generative adversarial imitation learning

1: **Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0
2: **for** $i = 0, 1, 2, \ldots$ **do**
3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$
4: Update the discriminator parameters from w_i to w_{i+1} with the gradient

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))] \quad (17)$$

5: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$. Specifically, take a KL-constrained natural gradient step with

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)] - \lambda \nabla_\theta H(\pi_\theta), \quad (18)$$
$$\text{where } Q(\bar{s},\bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \mid s_0 = \bar{s}, a_0 = \bar{a}]$$

6: **end for**

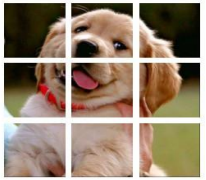
[1] CNN image CC-BY-SA by Aphex34 for Wikipedia https://commons.wikimedia.org/wiki/File:Typical_cnn.png
[2] RNN image CC-BY-SA by GChe for Wikipedia https://commons.wikimedia.org/wiki/File:The_LSTM_Cell.svg

- They all are single modality

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

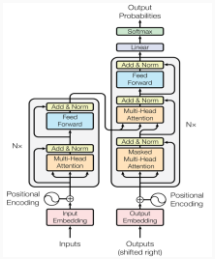
A Transformer is Born

- The Transformer's Takeover

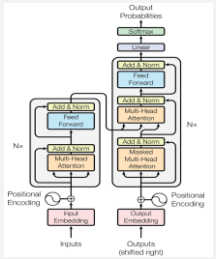


Input Tokens

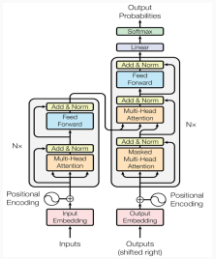
Computer Vision



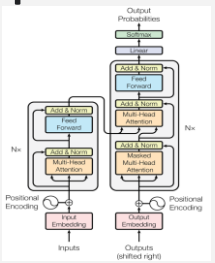
Natural Lang.



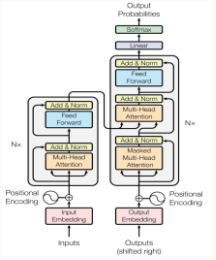
Reinf. Learning



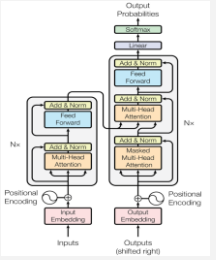
Speech



Biomedical



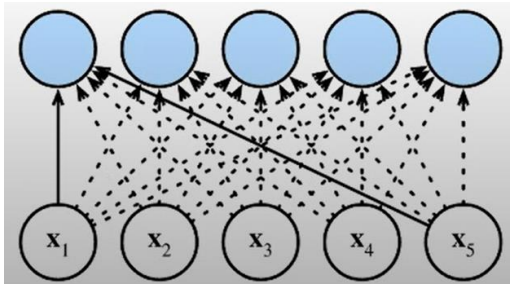
Graphs/Science



UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Attention is all you Need, Vaswani et al., 2017

Attention



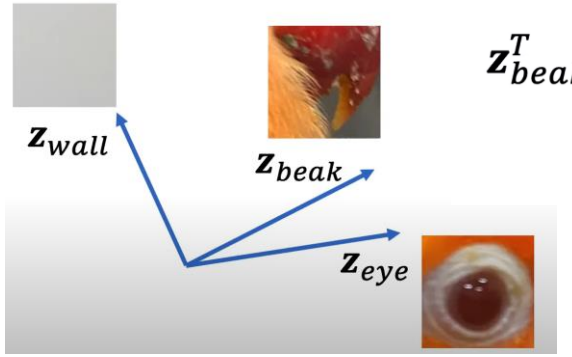
- At the heart of the Transformer lies the simple attention mechanism.
  - Goal of attention mechanism is to take each words (tokens) as shown in figure and create a representation that that contains information from the other relevant parts (other words/ tokens).
    (Creates a contextualized representation for each input token)

In natural language processing:

The quick brown fox jumps over the lazy dog

Think of Dot product!!

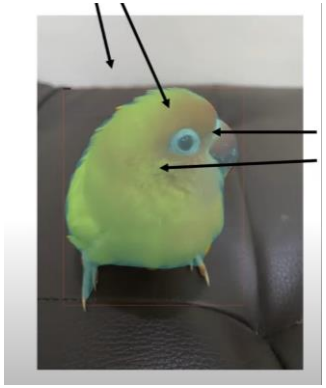
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



$$z_{beak}^T z_{eye} \gg z_{wall}^T z_{eye}$$

Attention

Low Attention

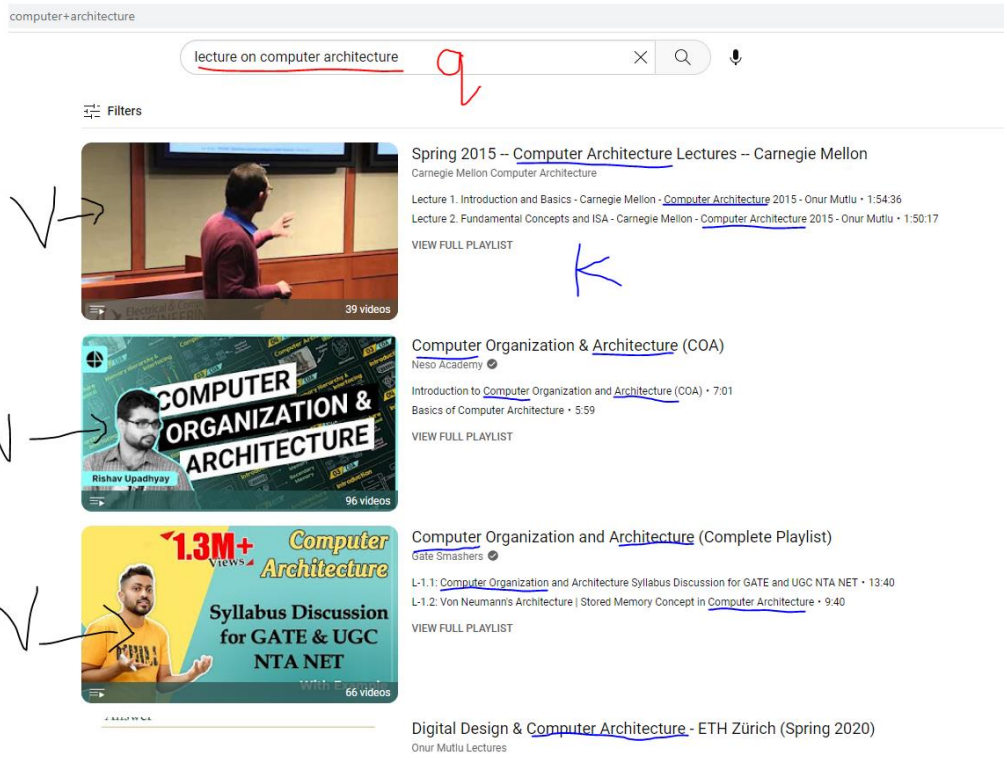


High Attention

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

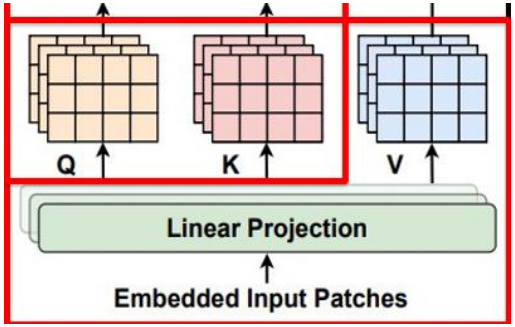
Attention - Intuition

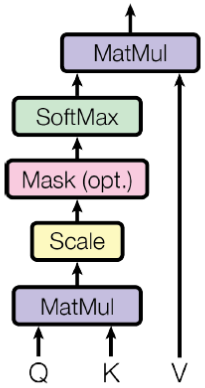
- Similar to retrieval from databases:
  - Query = a query we wish to run on a database
  - Key = the keys to search on in the database
  - Value = values corresponding to each key in the database



Scaled Dot-Product Attention

- Intuition – each token "searches the database" for tokens related to it.





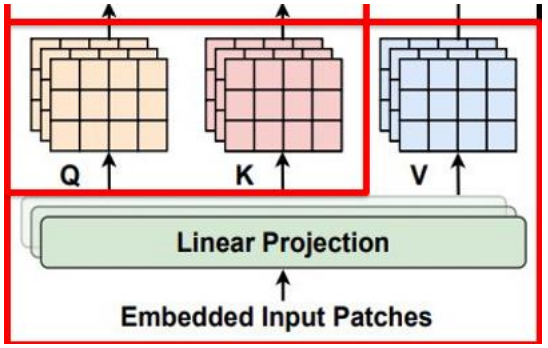
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

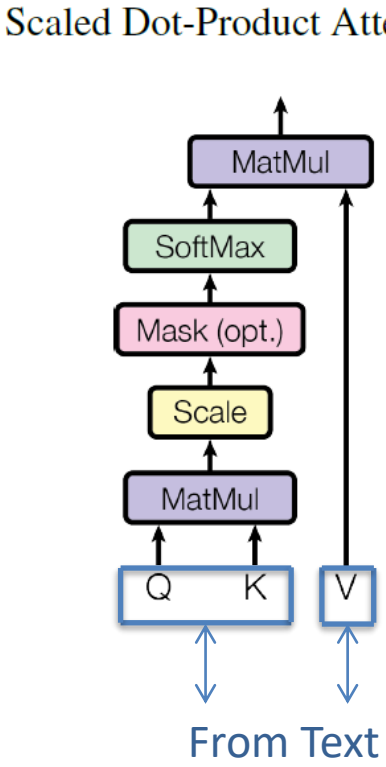
QUERY → Q
VALUE → V
KEY → K

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

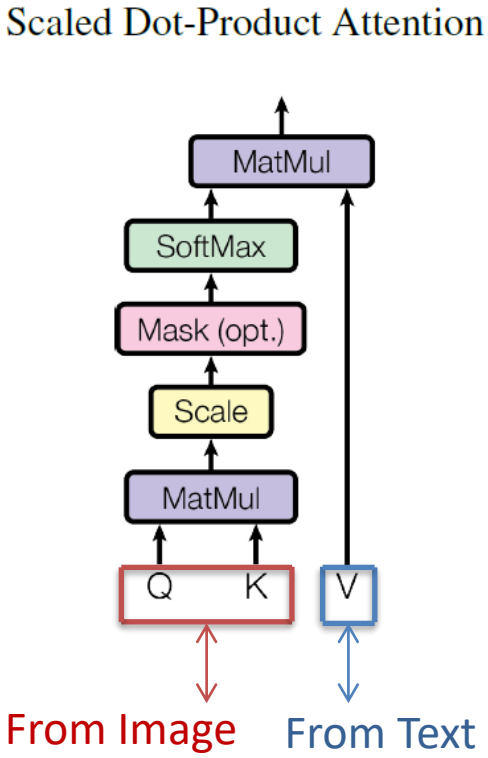
From Self-Attention to Cross-Attention

- Cross-attention is used to gain context from **another modality/ input type**
  - For example – gain context from text for image processing
  - Simply extract the queries, keys matrix from the other modality





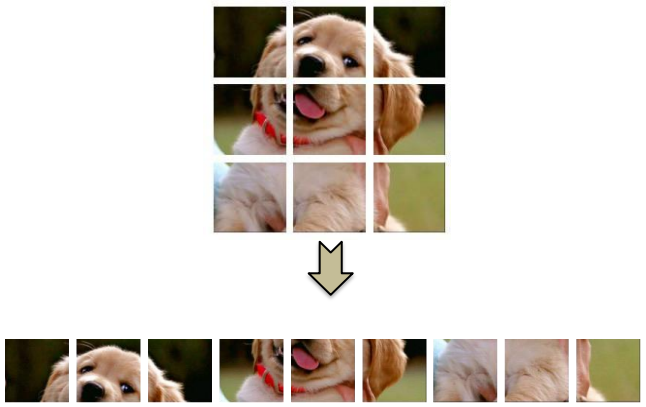
Self-Attention



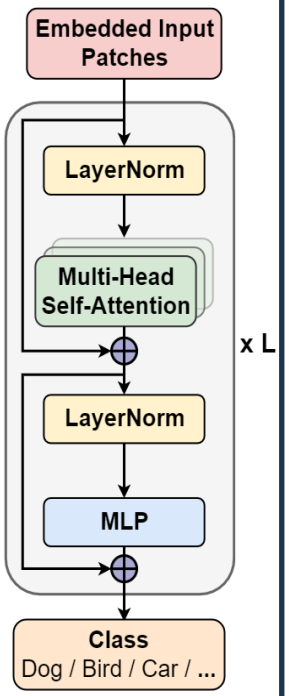
Cross-Attention

Background of Vision Transformer (ViTs)

■ **Input**: 2D image ➔ input tokens/patches



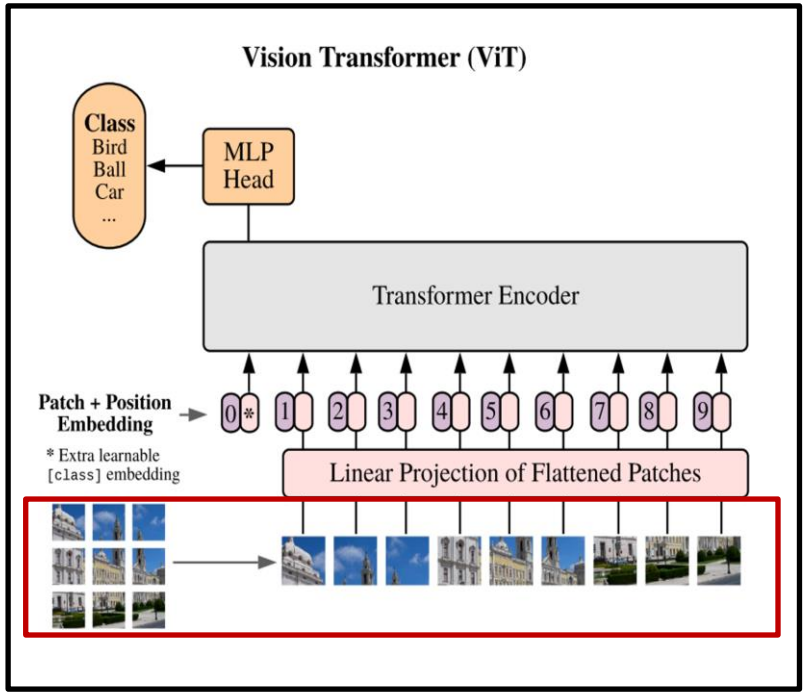
Input Tokens



ViT Models

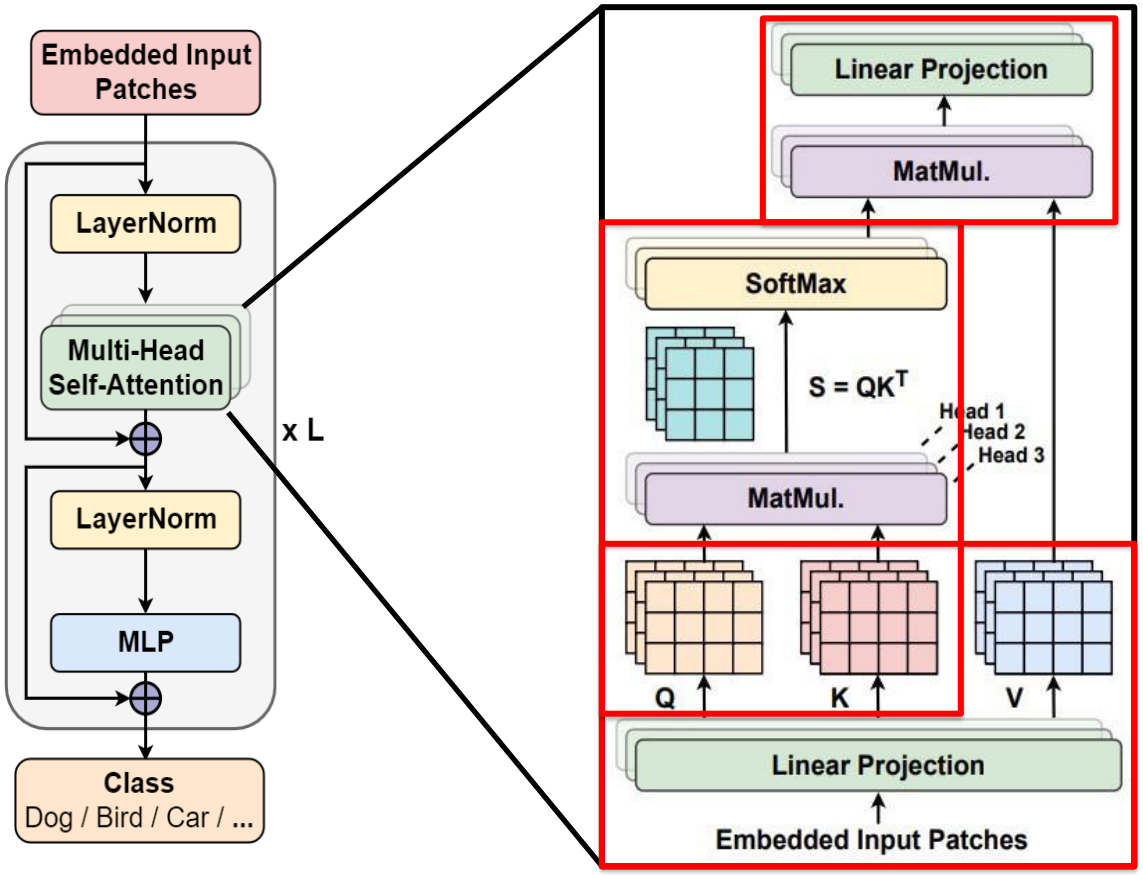
■ **Core Model**: Self-Attention and MLP

≈



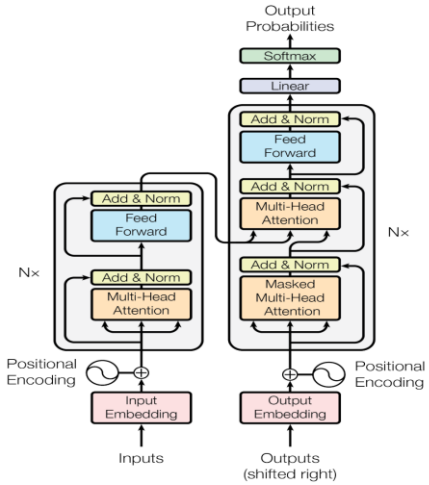
UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Background of Vision Transformer (ViTs)



QUERY → Q
VALUE → V
KEY → K

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

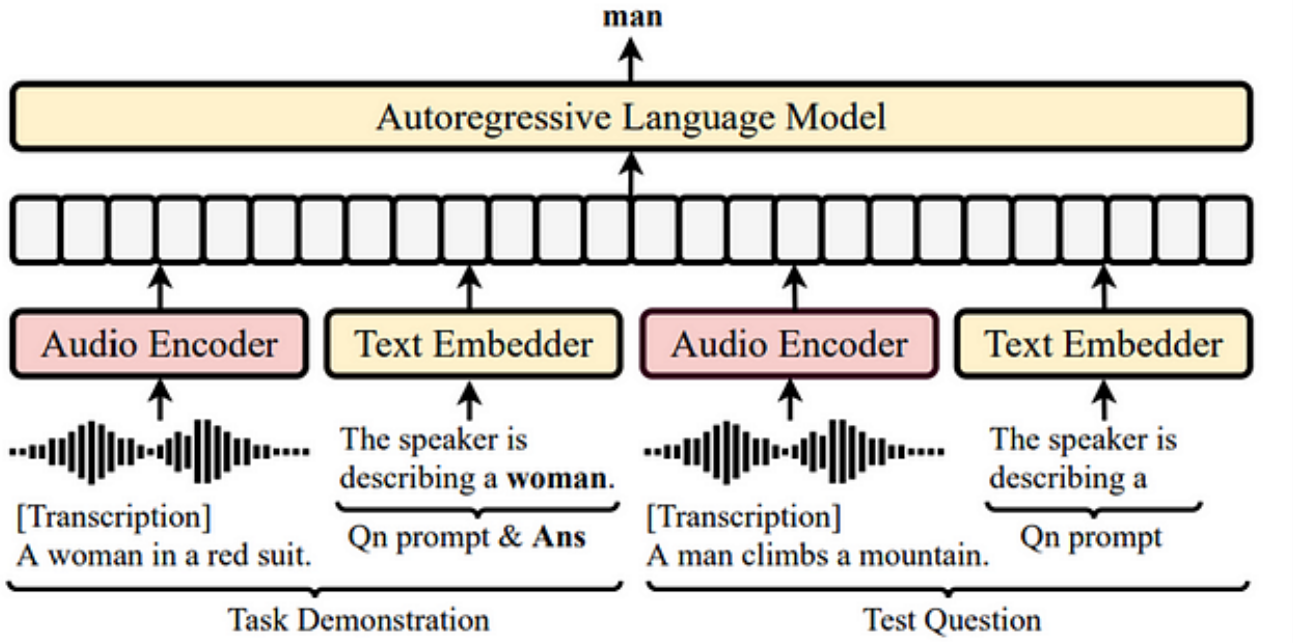


Examples of Cross-modal interactions

- Audio-Text
- Audio-Visual
- Vision-Language
- Video-Audio-Text

Audio-Text

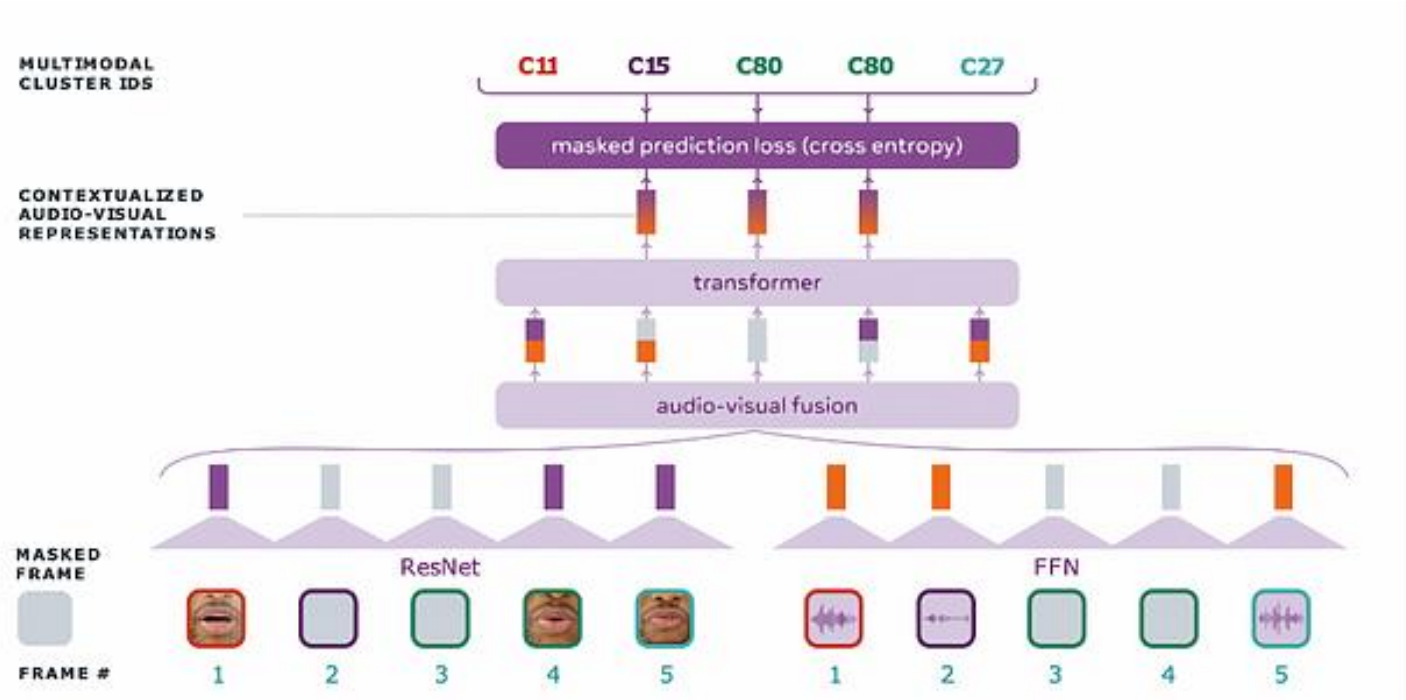
- Using the audio-text pair to learn spoken language understanding.
- In this work, the audio embeddings and text embeddings are concatenated and fed into an autoregressive language model.
- The audio encoder used is wav2vec 2.0 and the language model is GPT-2.



WAVPROMPT model architecture. [2]

Audio-Visual

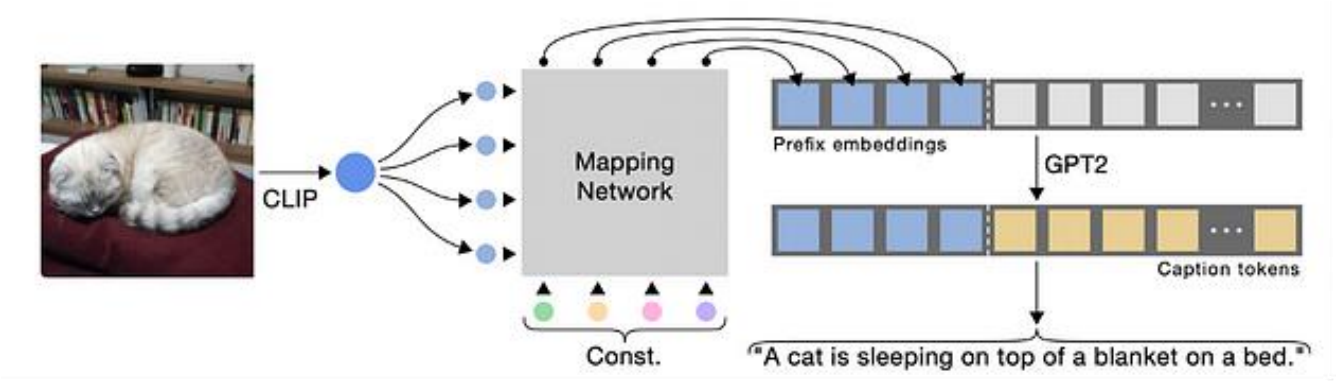
- Automatic speech recognition combined with lip reading.
- Lip reading is utilized to aid in automatic speech recognition tasks.



Audio-Visual Hidden Unit BERT (AV-HuBERT). [3]

Vision-Language

- Generate corresponding captions conditioned on the given image.
- Features of images are extracted by CLIP encoder since CLIP is designed to impose a shared representation for both images and text prompts.
- Then, a mapping network is applied to produce prefix embeddings which are then combined with the caption embeddings.
- These are then fed to the language model.



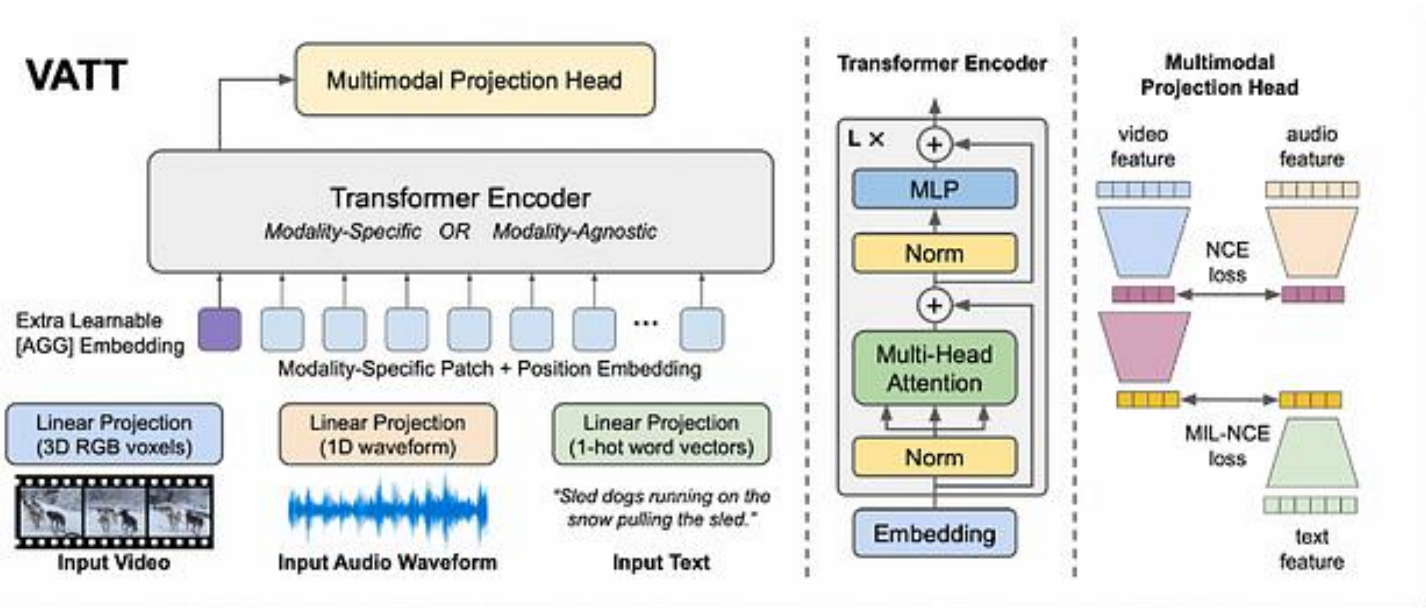
ClipCap model architecture. [4]

Video-Audio-Text

1. Zero-Shot Text-to-Video Retrieval: Given a particular text query and a pool of candidate videos, the task is to select the video that best corresponds to the text query.

2. Video Action Recognition: The task is to recognize common human actions in a video.

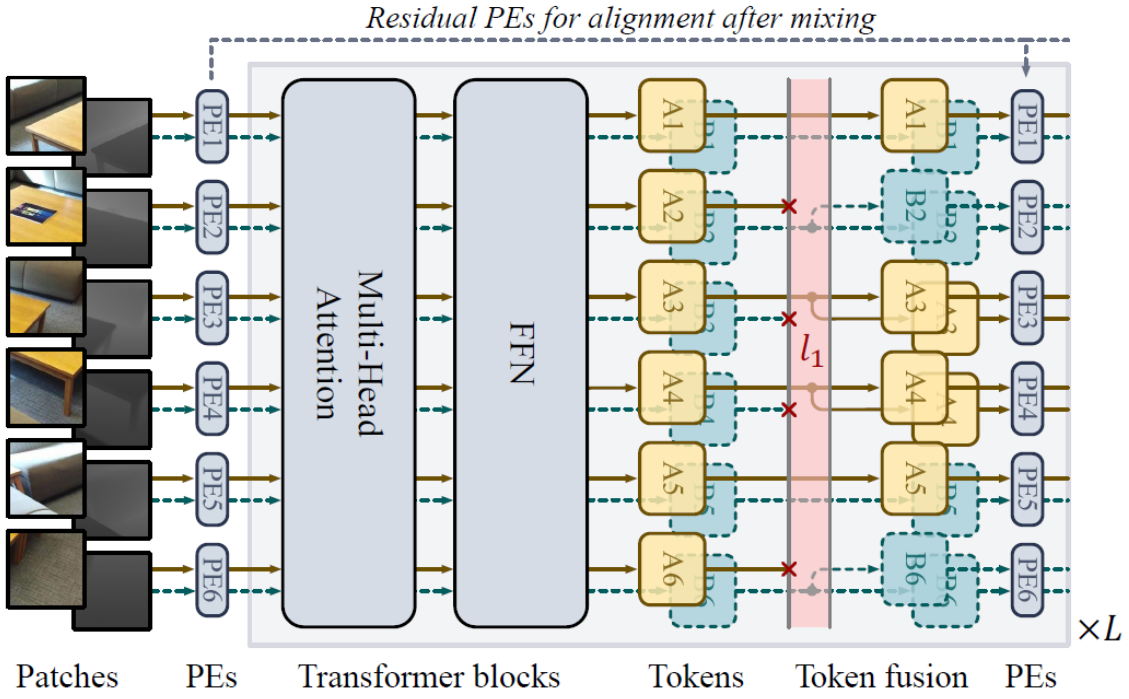
3. Audio Event Classification: The task is to recognize the respective audio events and their temporal start and end times in a recording.



VATT model architecture. [5]

Higher Modalities: Multimodal Token Fusion for Vision Transformers

- Proposes TokenFusion for fusing multiple vision transformer models handling different modalities (e.g., images, point clouds).
- Dynamically detects and prunes uninformative tokens from each transformer. The pruned tokens are substituted with projected features from other modalities.
- This allows combining off-the-shelf transformers without much modification to their original designs.



Model architecture. [10]

Homogeneous Multimodal Fusion

- Homogeneous modalities refer to multiple sources of data that are of the same type, such as multiple images or multiple audio files.
- Both RGB and depth data are sent to a shared transformer. Overall, this framework allows the model to learn correlations between the RGB and depth data.
- A pre-allocation is carried out so the fusion process can be of following form.

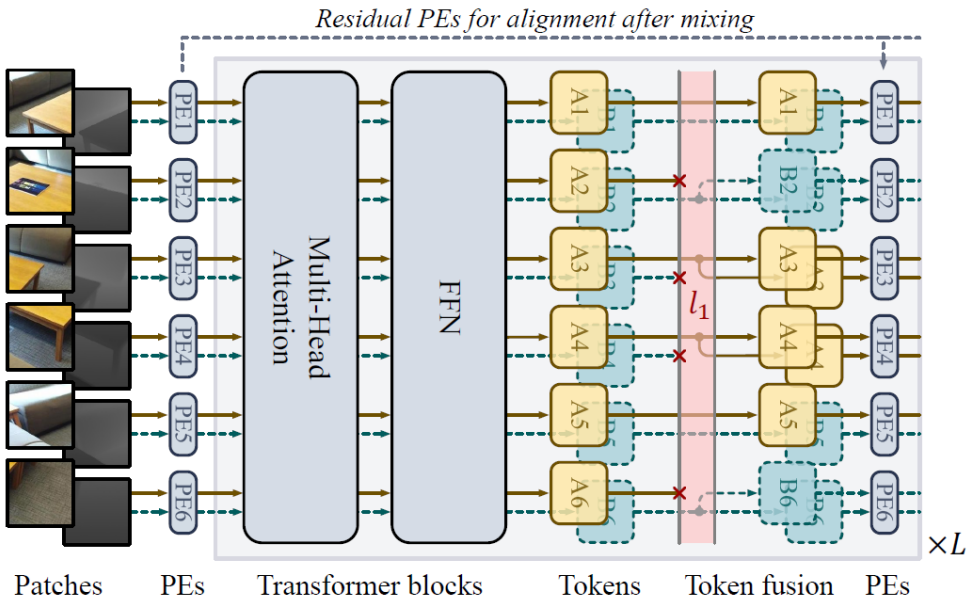


Figure 1. Framework of TokenFusion for homogeneous modalities with RGB and depth as an example. Both modalities are sent to a shared transformer with also shared positional embeddings.

$$e_m^l = e_m^l \odot \mathbb{I}_{s^l(e_m^l)\geq\theta} + \sum_{m'=1,m'\neq m}^{M} a_{m'}(m) \odot \mathrm{Proj}_{m'}^{M}(e_m^l) \odot \mathbb{I}_{s^l(e_m^l)<\theta}$$

Pre-allocation

$$a_{m'}(m) \in \{0,1\}^N$$

Heterogeneous Fusion for Vision Transformers (e.g., 3D object detection based on point cloud and 2D image

- Additional inter-modal projections (Proj) are needed.
- Image and 3d point data are mapped to the n-th image patch. $N_{\text{point}} - N_{\text{img}}$

$$[u, v, z]^{\mathrm{T}} = \mathbf{K} \cdot \mathbf{R}_t \cdot \left[x_{n_{\text{point}}}, y_{n_{\text{point}}}, z_{n_{\text{point}}}, 1\right]^{\mathrm{T}}, \quad n_{\text{img}} = \left\lfloor \frac{\lfloor v/z \rfloor}{P} \right\rfloor \times \left\lfloor \frac{W}{P} \right\rfloor + \left\lfloor \frac{\lfloor u/z \rfloor}{P} \right\rfloor$$

- The TokenFusion method dynamically detects uninformative tokens and substitutes them with projected and aggregated inter-modal features.

$\left[x_{n_{\text{point}}}, y_{n_{\text{point}}}, z_{n_{\text{point}}}\right]$

3-d point data

$[\lfloor u/z \rfloor, \lfloor v/z \rfloor]$

2D pixel data

$\mathbf{K}$ and $\mathbf{R}_t$

Camera parameters

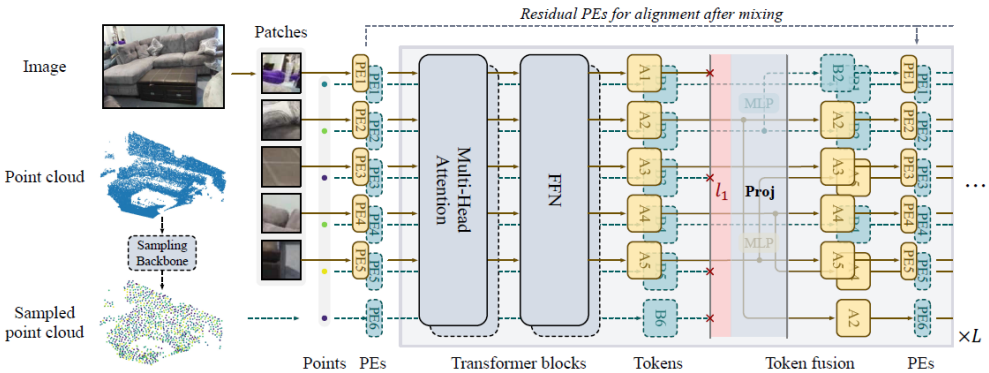


Figure 2. Framework of TokenFusion for heterogeneous modalities with point clouds and images. Both modalities are sent to individual transformer modules with also individual positional embeddings. Additional inter-modal projections (Proj) are needed which is different from the fusion for homogeneous modalities.

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Some other Applications

- Multi-Modal Retrieval
- Image Captioning
- Image Question Answering

Multi-Modal Retrieval

- This model has two capabilities: image-to-text retrieval and text-to-image retrieval.
- The former provides text from an image query, while the latter provides an image from a text query.

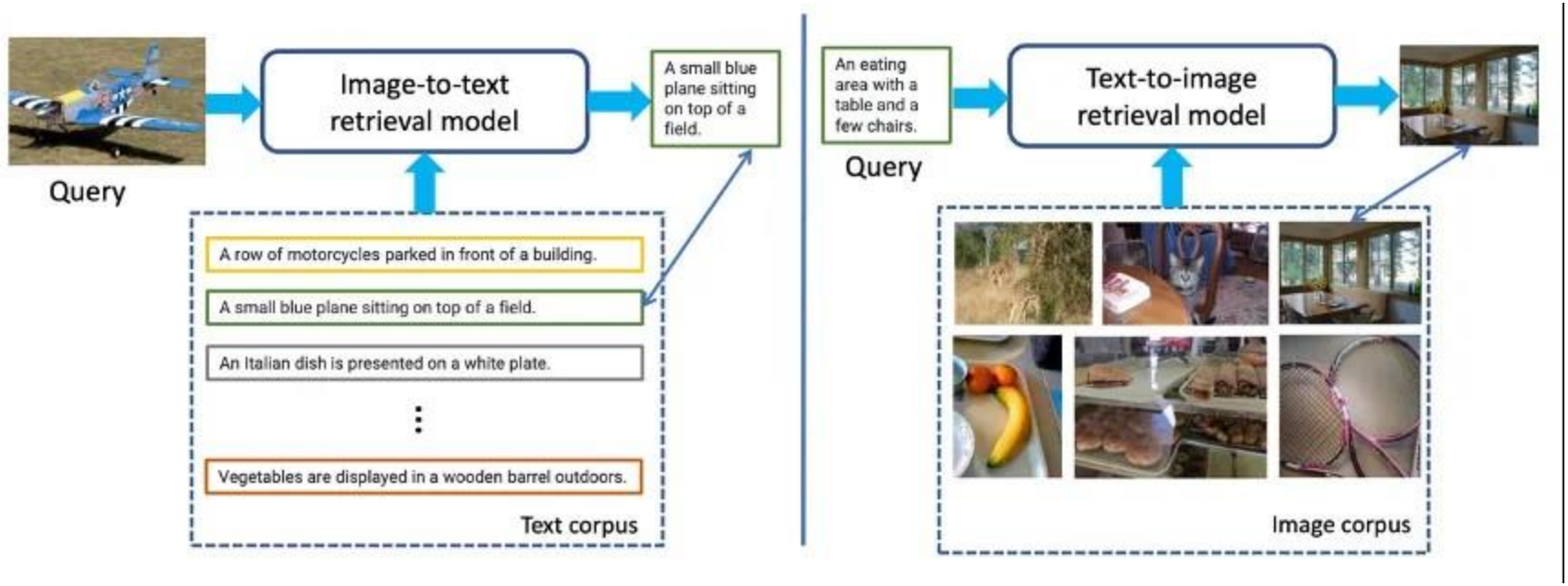
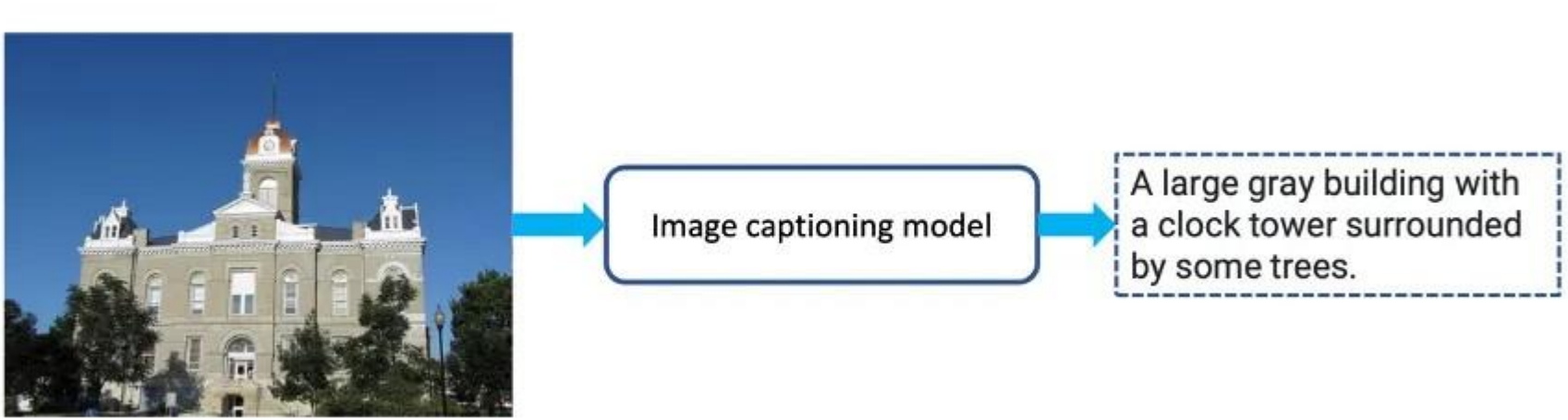


Image-to-text and text-to-image retrieval. [7]

Image Captioning

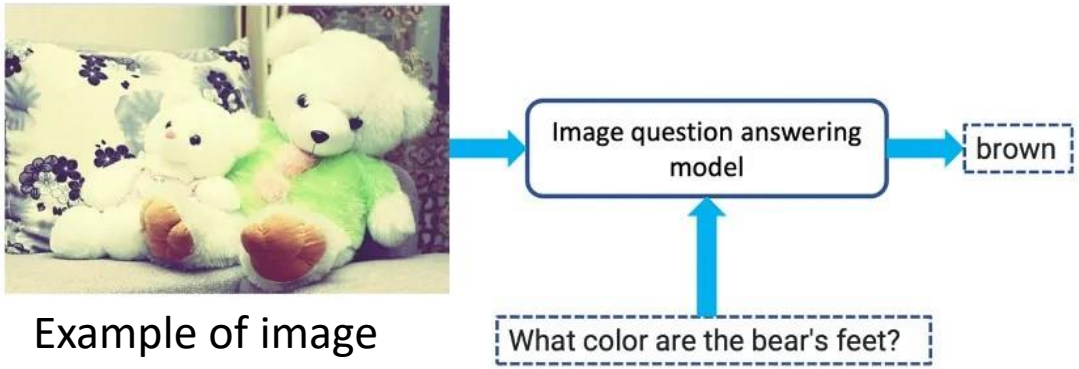
- The image captioning model generates a corresponding description based on a given image.



Example of image captioning. [7]

Image Question Answering

- Image question answering is like image captioning. The model answers a question based on the image.
- The challenge is to help the model understand the relationship between image and text. This can be done with large datasets of image-text pairs.



Example of image question answering. [8]

Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

Conceptual Captions: a worker helps to clear the debris.

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Example of image-caption(text) pairs. [9]


a very typical bus station

functions of government : 1 . form a more perfect union


safe deposit with money around it on a white background photo


emergency services were called after a car smashed through a set of traffic lights

Example of image-text pairs. [7]

Conclusion

- **Enhanced Information:** Multimodal data provides richer information than unimodal data, offering AI models a broader context for decision-making.
- **Challenges:** Training multimodal networks is complex due to issues like data alignment and modality fusion, presenting significant research challenges.
- **Performance Improvement:** The integration of multiple modalities can potentially lead to improved predictive performance in AI models.
- **New Possibilities:** Multimodal AI unlocks capabilities that were previously challenging for unimodal models, enabling more human-like understanding and problem-solving.
- **Active Research:** Multimodal deep learning is a dynamic and evolving research area with applications spanning various domains.

References

1. Morency, L. P., Liang, P., & Zadeh, A. (2022). Multimodal Machine Learning | CVPR 2022 Tutorial.
2. Chen, Y., Chen, Y., & Cheng, K. (2022). WAVPROMPT: Towards Few-Shot Spoken Language Understanding with Frozen Language Models.
3. Gao, L., Wang, J., & Zhang, H. (2022). Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction.
4. Zhou, L., Liu, X., & Zhang, Z. (2021). ClipCap: CLIP Prefix for Image Captioning.
5. Sun, L., Jia, K., & Qi, J. (2021). VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text.
6. CVPR 2022 Tutorial on “Recent Advances in Vision-and-Language Pre-training”.
7. CVPR 2022 Tutorial on “Recent Advances in Vision-and-Language Pre-training” | Overview of Image-Text Pre-training Slides.
8. Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering.
9. Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning.
10. Wang, Yikai, et al. "Multimodal token fusion for vision transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Thank you for your attention!

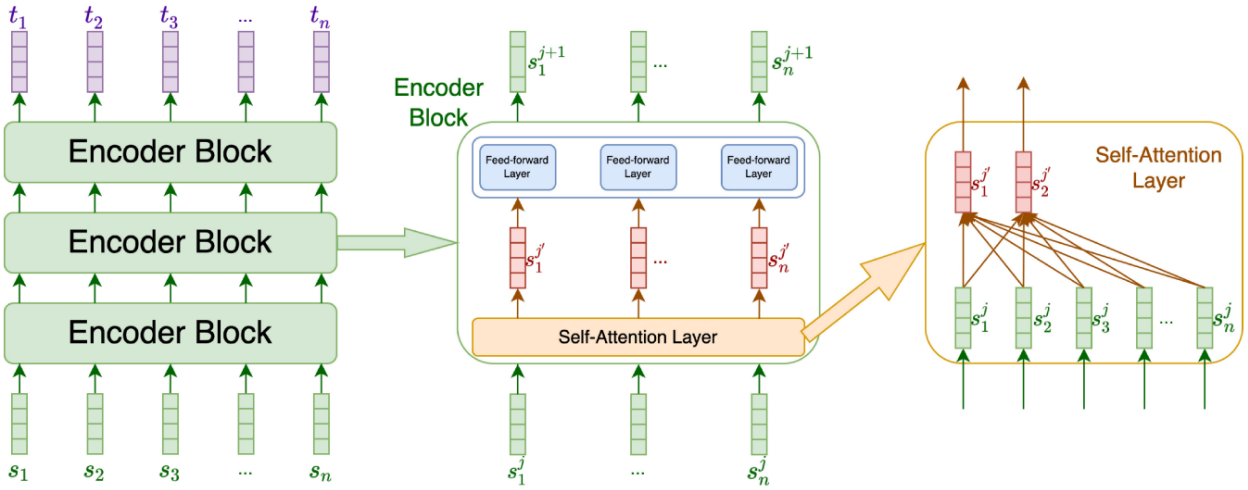
Questions ??

adevkota2@uh.edu

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Extras

Some EM papers on Transformers



[1]

FIGURE 2. The neural network illustration of the encoder-only transformer. The output of the transformer encoder is normally referred to as the hidden state H, but in the case of the encoder-only transformer, H is the final output target sequence T. This Figure is based on [33].

Theta	-90	-89	...	-60	...	-53	-52	-51	-50	-49	-48	-47	...	90
Beam	0	0	...	0	...	0	0	0	1	0	0	0	...	0
HPBW	0	0	...	0	...	1	1	1	1	1	1	1	...	0
SLL	-12	-12	...	-12	...	-12	-12	-12	-12	-12	-12	-12	...	-12
MNL	-21	-21	...	-21	...	-21	-21	-21	-21	-21	-21	-21	...	-21
Nulls	0	0	...	1	...	0	0	0	0	0	0	0	...	0

(a) An example of modified far-field pattern specifications as input sequence.

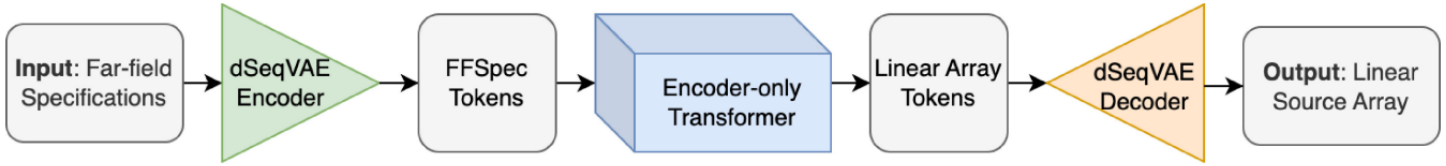
Theta	-90	-89	...	-60	...	-53	-52	-51	-50	-49	-48	-47	...	90
Power	-50	-36	...	-50	...	-3	-2	-1	0	-1	-2	-3	...	-50

(b) An example of far-field power pattern as output sequence.

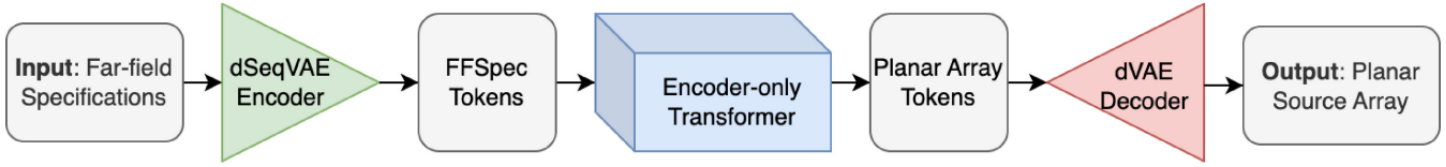
Fig. 2. An example of the modified far-field pattern specifications as the input sequence and far-field power pattern as the output sequence. In order for the input sequence to have the same length as the output sequence, we created the table shown in (a) where the first row is the theta (θ) angles of the far-field cut. The second, third and last row are binary arrays that represent the main beam directions, HPBWs, and null positions respectively where 1 represents the presence of the quantity and 0 represents the absence of the same quantity. The fourth and fifth row are arrays of constant values of side lobe level (SLL) and maximum null levels (MNL).

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Some EM papers on Transformers



(a) The proposed neural network architecture of far-field specifications to linear source array synthesis.



(b) The proposed neural network architecture of far-field specifications to planar source array synthesis.

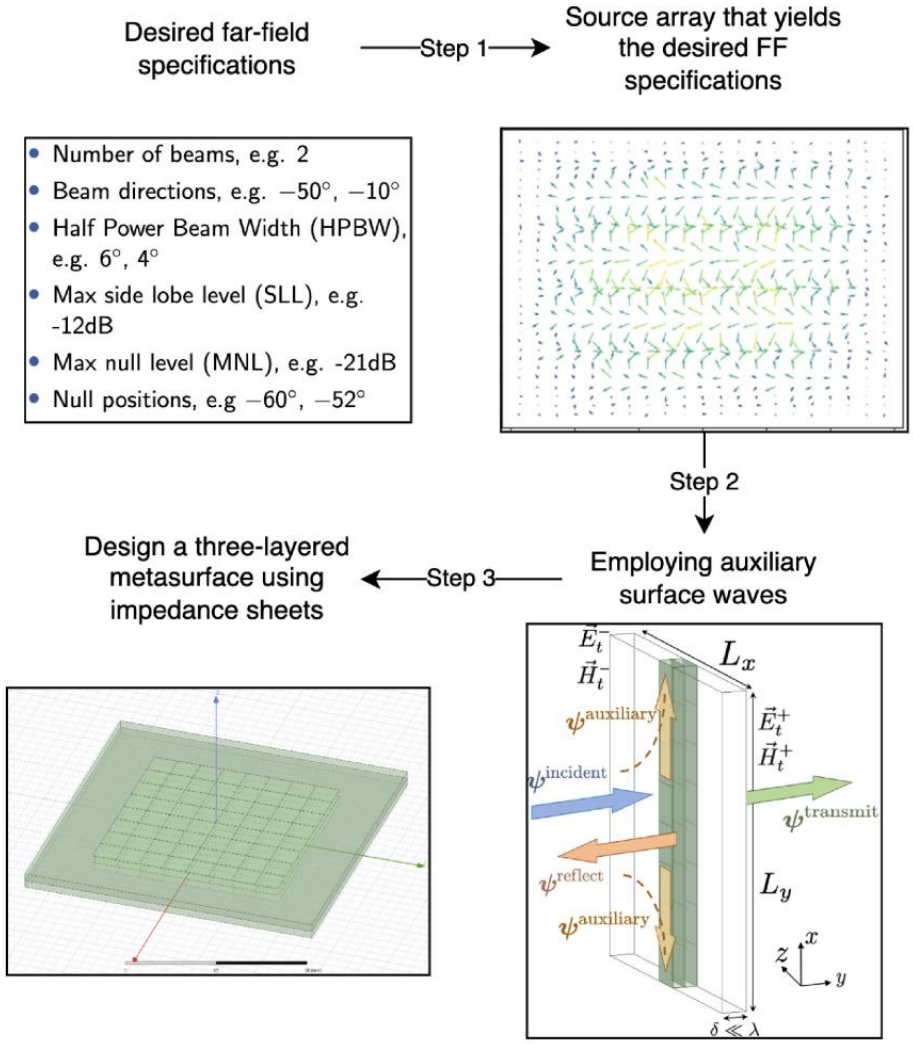
[2]

FIGURE 6. The proposed neural network architecture of far-field specifications to source array synthesis for both (a) 2D and (b) 3D scenarios. It should be noted that in (a) the dSeqVAE encoder and the dSeqVAE decoder on the two sides of the transformer are of two separate models as indicated by the different colors. On the other hand, the dSeqVAE encoders in (a) and (b) are of the same model, and are therefore shown with the same color.

Theta (deg)	0.0	1.0	...	59.0	60.0	61.0	...	179.0	180	90.0	...	90.0	90.0	90.0	90.0	90.0	90.0	90.0	...	90.0
Phi (deg)	90.0	90.0	...	90.0	90.0	90.0	...	90.0	90.0	0.0	...	67.0	68.0	69.0	70.0	71.0	72.0	73.0	...	180.0
Beams	0	0	...	0	1	0	...	0	0	0	...	0	0	0	1	0	0	0	...	0
HPBW	0	0	...	1	1	1	...	0	0	0	...	0	1	1	1	1	1	0	...	0
Nulls	0	1	...	0	0	0	...	1	0	0	...	1	0	0	0	0	0	1	...	0
SLL (dB)	-12.3	-12.3	...	-12.3	-12.3	-12.3	...	-12.3	-12.3	-13.8	...	-13.8	-13.8	-13.8	-13.8	-13.8	-13.8	-13.8	...	-13.8
MNL (dB)	-21.5	-21.5	...	-21.5	-21.5	-21.5	...	-21.5	-21.5	-37.6	...	-37.6	-37.6	-37.6	-37.6	-37.6	-37.6	-37.6	...	-37.6

FIGURE 5. The vectorized far-field performance criteria where the first row lists the theta (θ) angles of the far-field cut. The main beam directions, HPBWs of each beam, and null locations are represented in binary format, where 1 represents the presence and 0 represents the absence of the quantity, and are located at the second, third, and last rows respectively. The side lobe level (SLL) and maximum null levels (MNL) are arrays of constant value and are located at the fourth and fifth rows respectively.

Some EM papers on Transformers



[2]

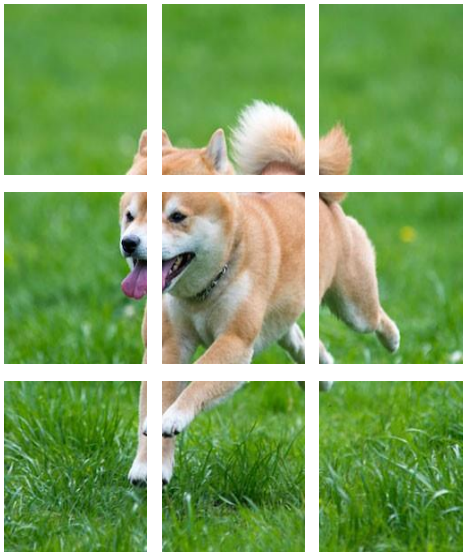
FIGURE 1. The end-to-end metasurface macroscopic design pipeline utilized in this work. The lengths and directions of the arrows in the source array image depict the amplitude and phase of different infinitesimal dipole antennas respectively.

Ref: Some EM papers on Transformers

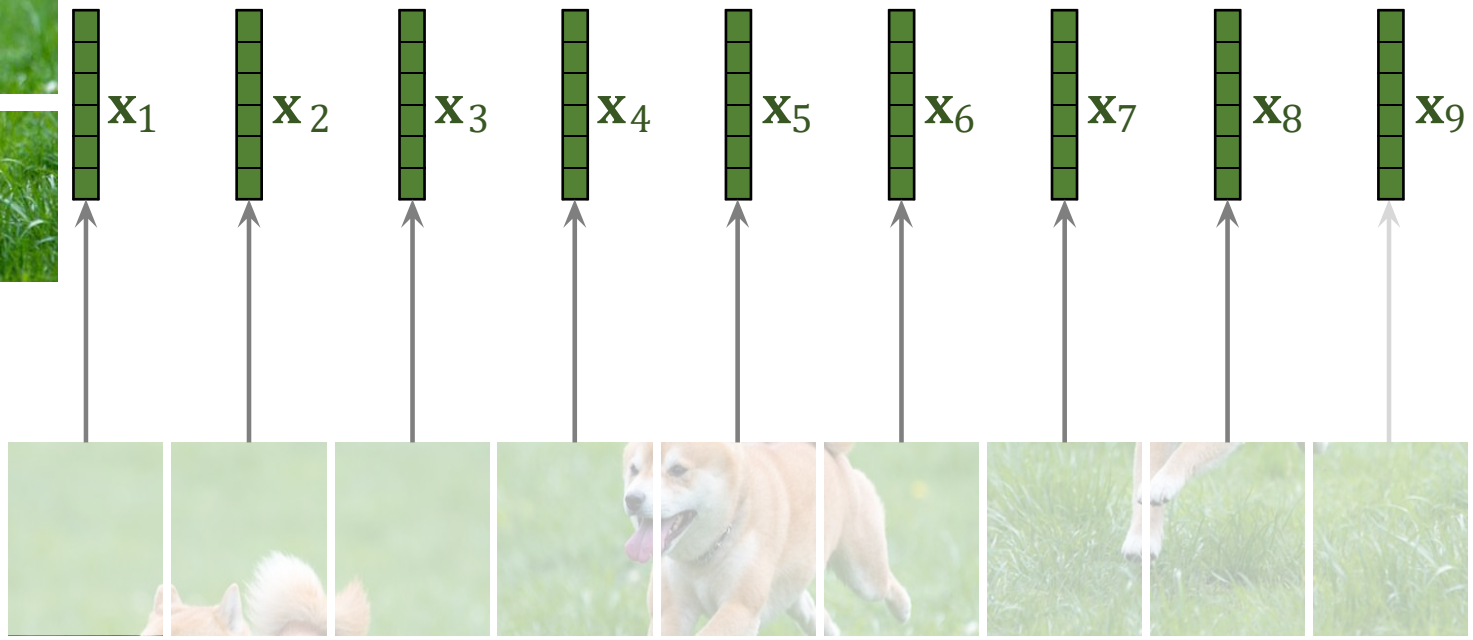
1. Niu, Chen, et al. "A deep learning based approach to design metasurfaces from desired far-field specifications." IEEE Open Journal of Antennas and Propagation (2023).
2. Niu, Chen, Max Kelly, and Puyan Mojabi. "An encoder-only transformer to generate power patterns from far-field performance criteria." 2022 16th European Conference on Antennas and Propagation (EuCAP). IEEE, 2022.

How attention is calculated

Vectorization

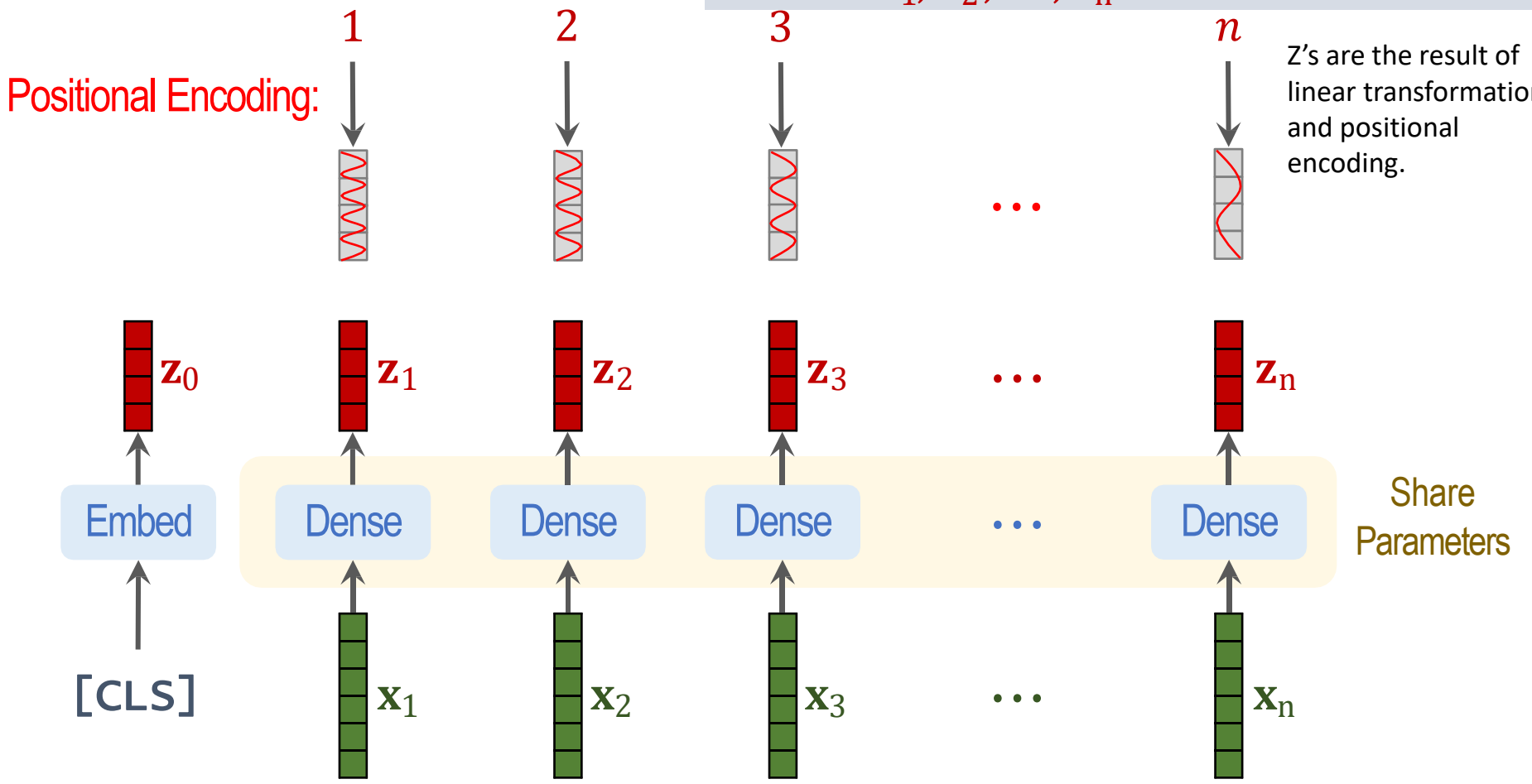


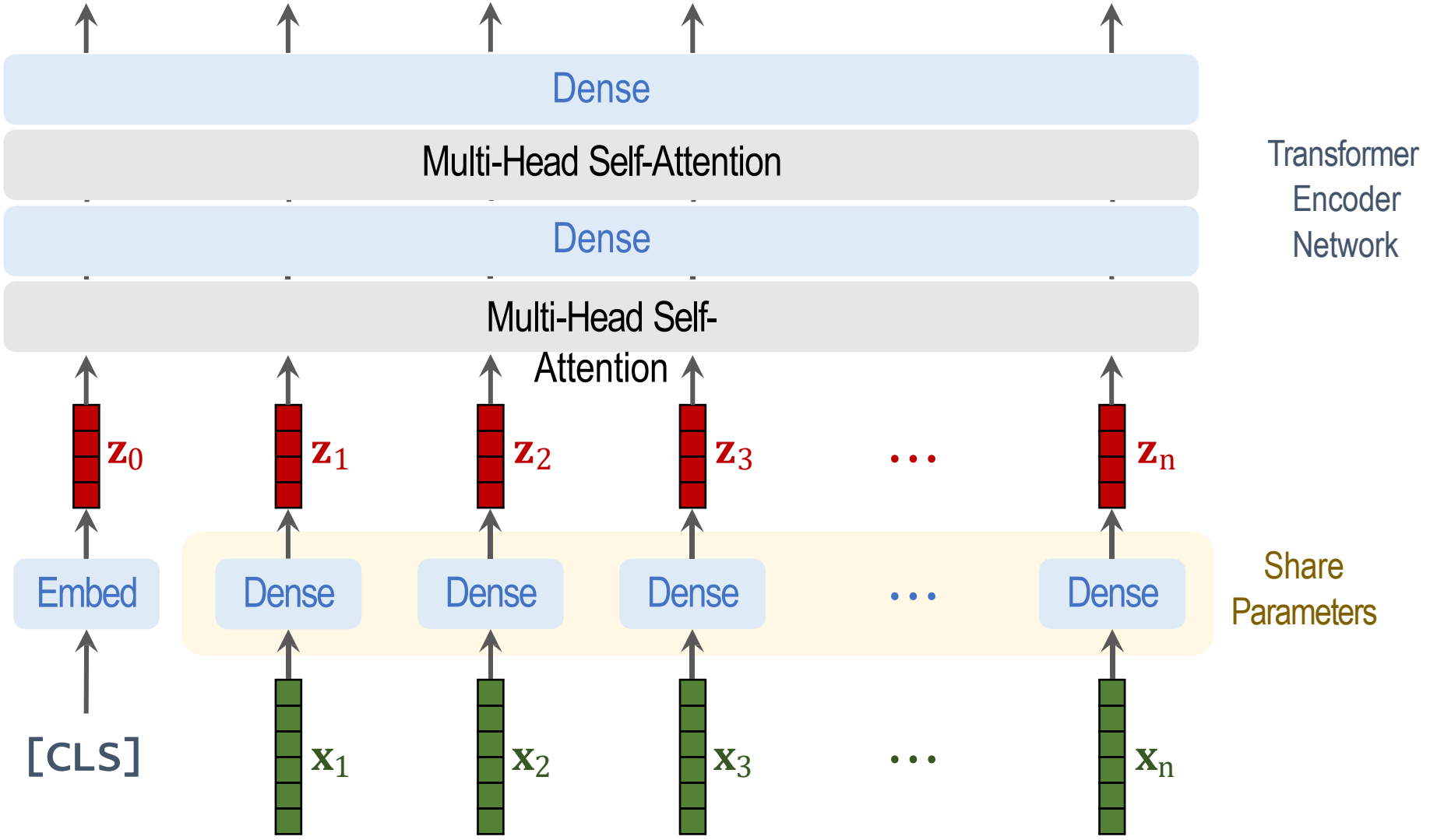
If the patches are $d_1 \times d_2 \times d_3$ tensors, then the vectors are $d_1 d_2 d_3 \times 1$.

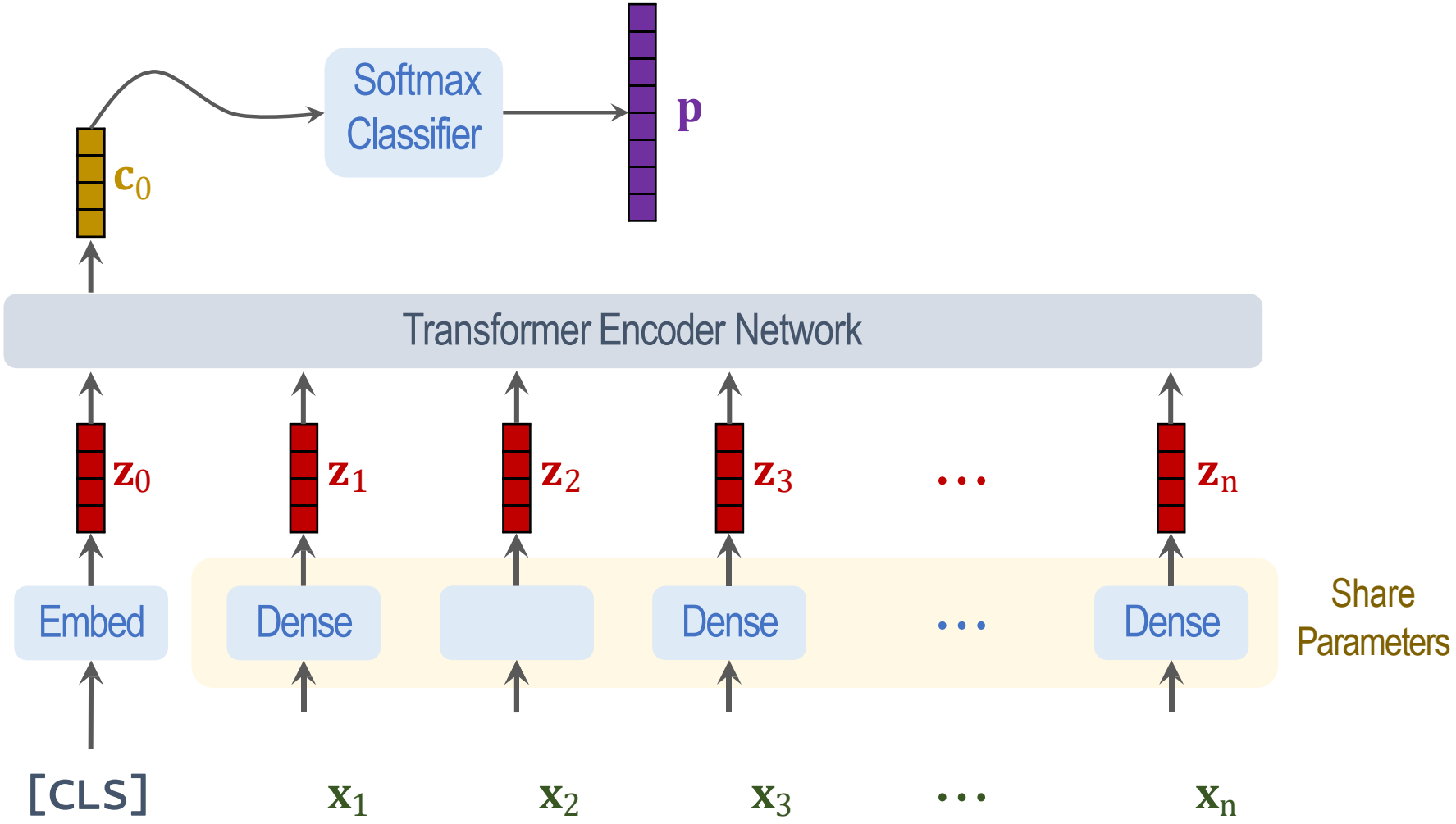


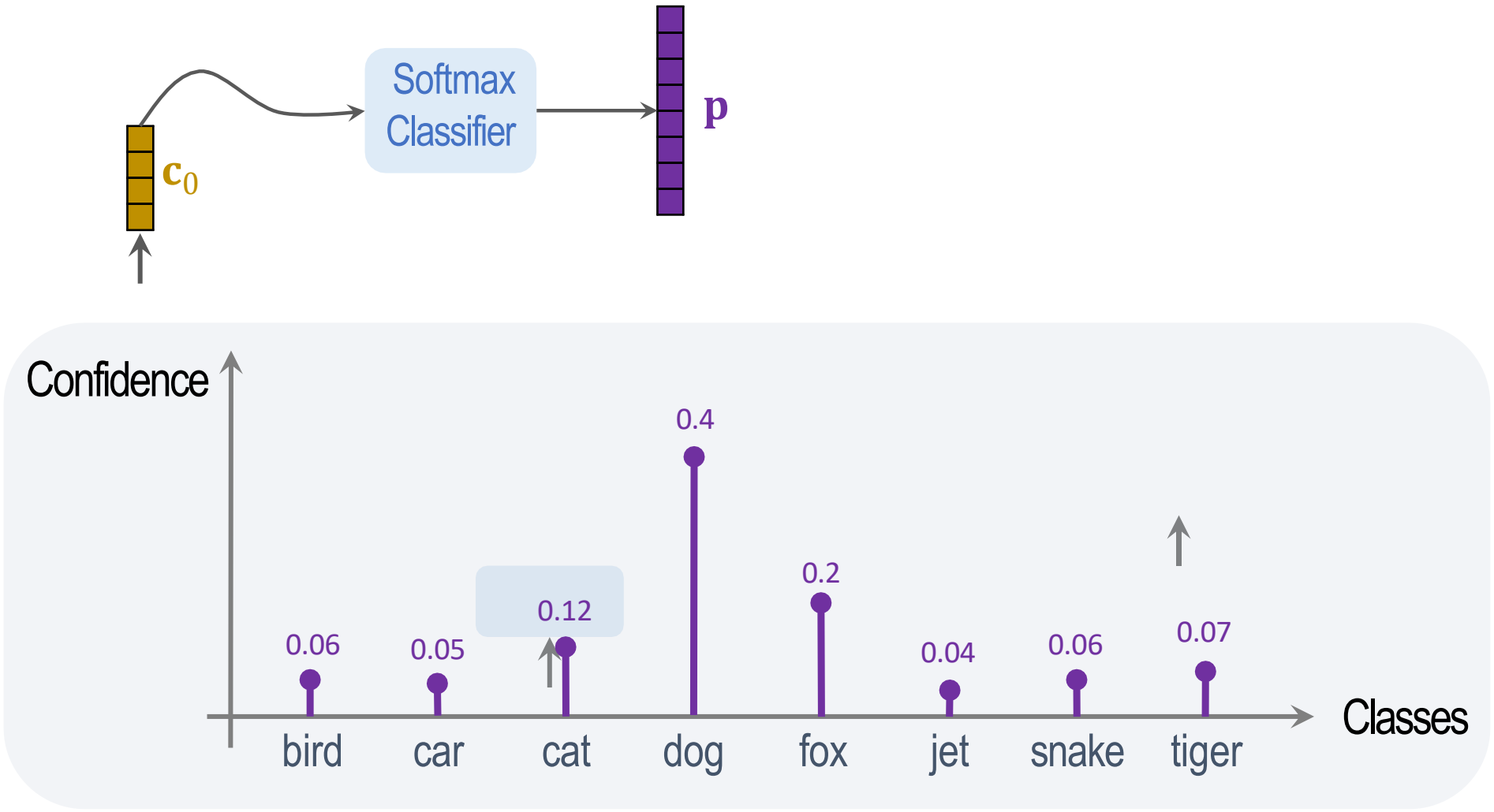
$\mathbf{z}_1 = \mathbf{W}\,\mathbf{x}_1 + \mathbf{b}$ $\mathbf{z}_2 = \mathbf{W}\,\mathbf{x}_2 + \mathbf{b}$

Also, add positional encoding vectors to $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n$.




Dense
Multi-Head Self-Attention
Dense
Multi-Head Self-Attention
Transformer Encoder Network
z_0
z_1
z_2
z_3
z_n
Embed
Dense
Dense
Dense
Dense
Share Parameters
[CLS]
x_1
x_2
x_3
x_n


Softmax Classifier
p
c_0
Transformer Encoder Network
z_0
z_1
z_2
z_3
z_n
Embed
Dense
Dense
Dense
Share Parameters
[CLS]
x_1
x_2
x_3
x_n


Softmax
Classifier
$\mathbf{c}_0$
$\mathbf{p}$
Confidence
0.06
0.05
0.12
0.4
0.2
0.04
0.06
0.07
bird
car
cat
dog
fox
jet
snake
tiger
Classes

CLIP encoder