STARS RATED FOR EXCELLENCE 2012

Fpt University

TRƯỜNG ĐẠI HỌC FPT

Visual Question Answering for Medical Data Using a Visio-Linguistic Model

Students

Tran Quang Duc

Le Viet Tien

Tran Thi Kim Thanh

Advisor

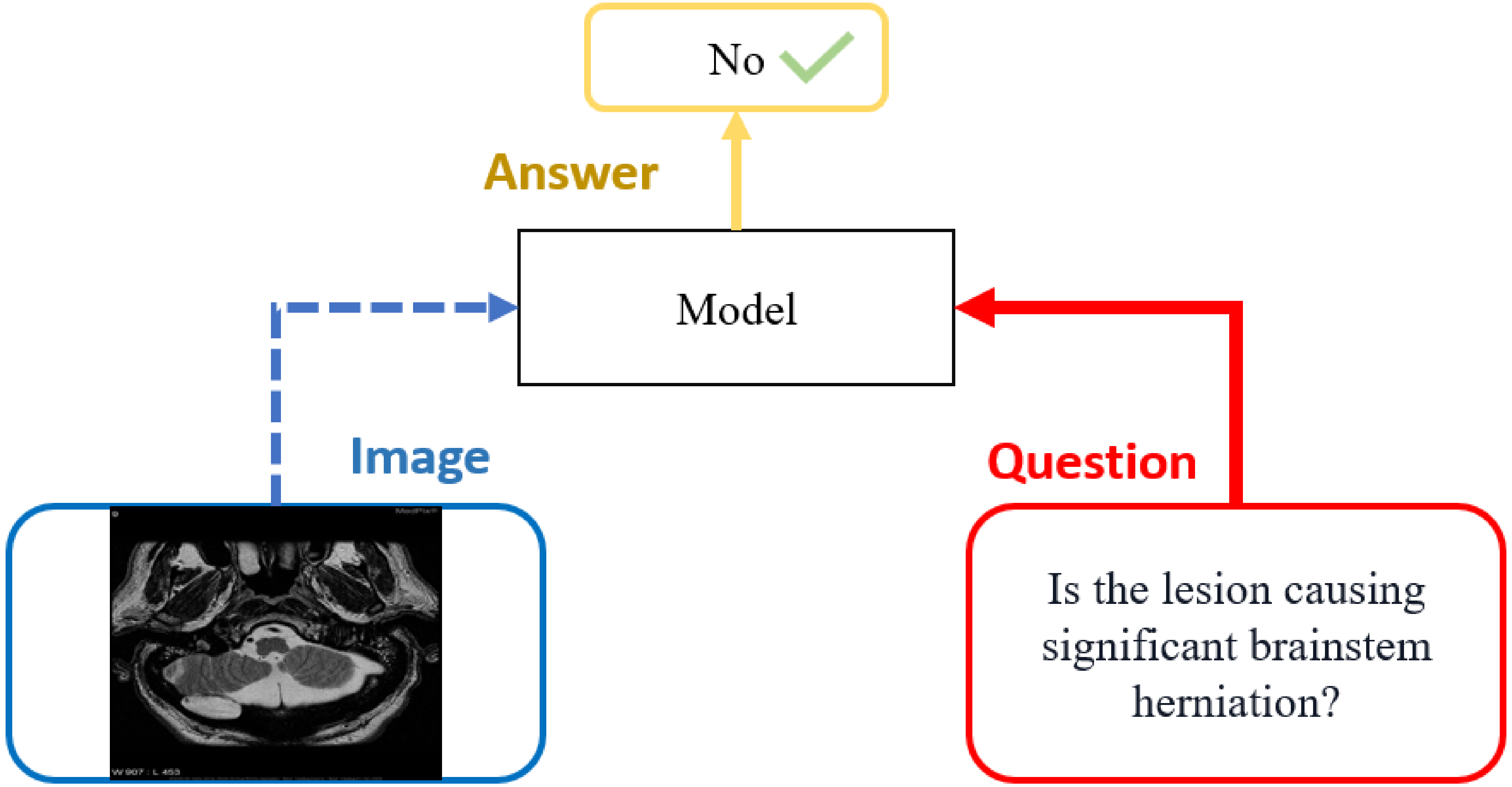
Bui Van Hieu

Table of content

INTRODUCTION

Problem & Motivation

- Medical Visual Question Answering (Med-VQA) is a challenging task that combines the fields of CV and NLP.



Problem & Motivation

QS STARS RATED FOR EXCELLENCE 2012

FPT Fpt University

TRƯỜNG ĐẠI HỌC FPT

- Med-VQA is still in its infancy and is far from practical use[1].

[1] Bazi, Y., Rahhal, M. M., Bashmal, L., & Zuair, M. (2023). Vision–Language Model for Visual Question Answering in Medical Imagery. Bioengineering, 10(3), 380. https://doi.org/10.3390/bioengineering10030380

QS STARS RATED FOR EXCELLENCE 2012

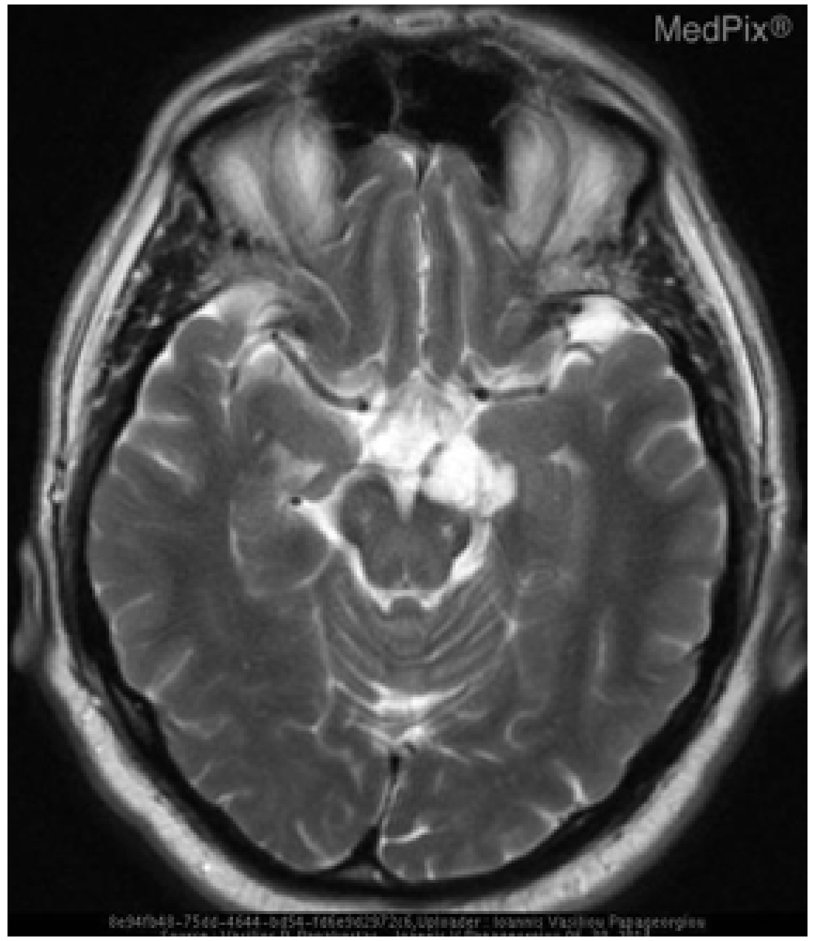
FPT Fpt University

TRƯỜNG ĐẠI HỌC FPT

Problem & Motivation

- The current medical data is limited. [2]

--->The efficacy of medical models is suboptimal.



Question: Is this a singular or multilobulated lesion?

Answer: Multilobulated

[2] Nguyen, B. D., Do, T., Nguyen, B. X., Do, T., Tjiputra, E., & Tran, Q. D. (2019). Overcoming Data Limitation in Medical Visual Question Answering. ArXiv. /abs/1909.11867

Related work

TRƯỜNG ĐẠI HỌC FPT

VQA-RAD

Team/Method	Image Encoder	Language Encoder	Fusion	Output Mode	Other Technique(s)
BAN-VQAMix	CNN	LSTM	BAN	Classification	Triplet Mixup Scheme
MTPT-CMSA	Multi- ResNet-34	LSTM	CSMA	Classification	Cross-modal self-attention, Multi-task pre-training with extra data
hi-VQA	EfficientNet-b5	RadBERT	Multi-head attention(Transformer)	Classification	
MMQ-BAN	MMQ	LSTM	BAN/SAN	Classification	Multiple Meta-model Quantifying
Q2ATransformer	Swin Transformer	BERT	Multi-head attention(Transformer)	Classification	

Objective

1. Introduce an architecture Med-VQA with Associative Memory Module (AMM)
2. Practical Prototype Learning in features fusion.
3. We achieved an improved result on VQA-RAD.

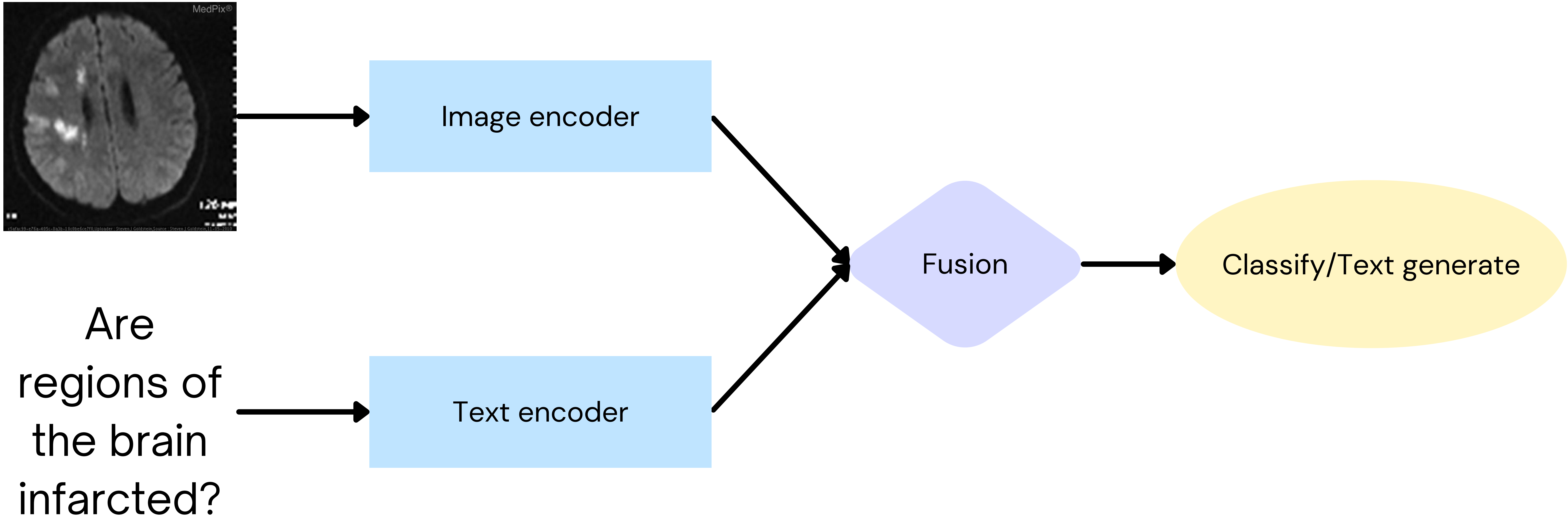
METHODOLOGY

QS STARS RATED FOR EXCELLENCE 2012

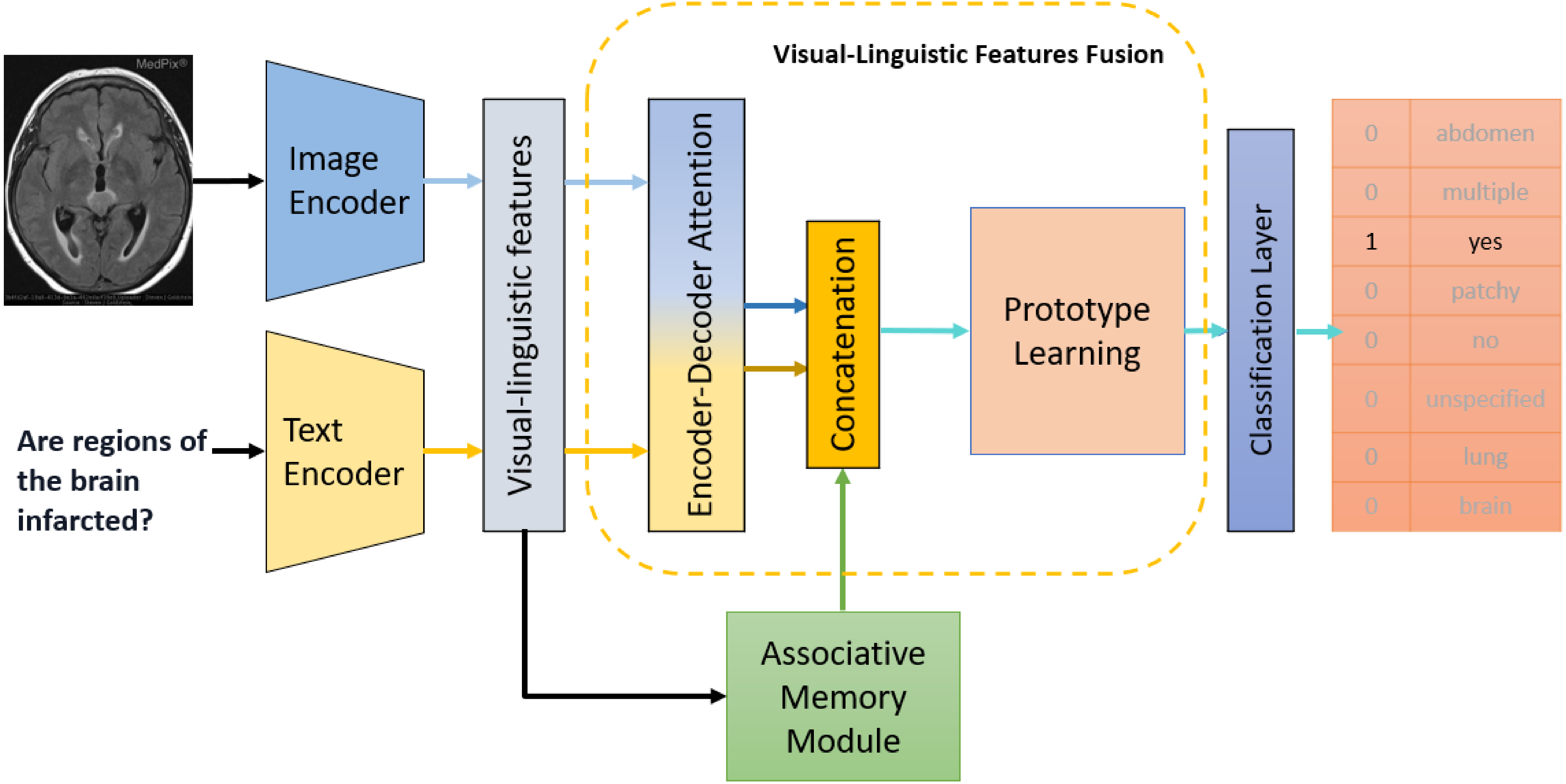
Fpt University

TRƯỜNG ĐẠI HỌC FPT

Methodology



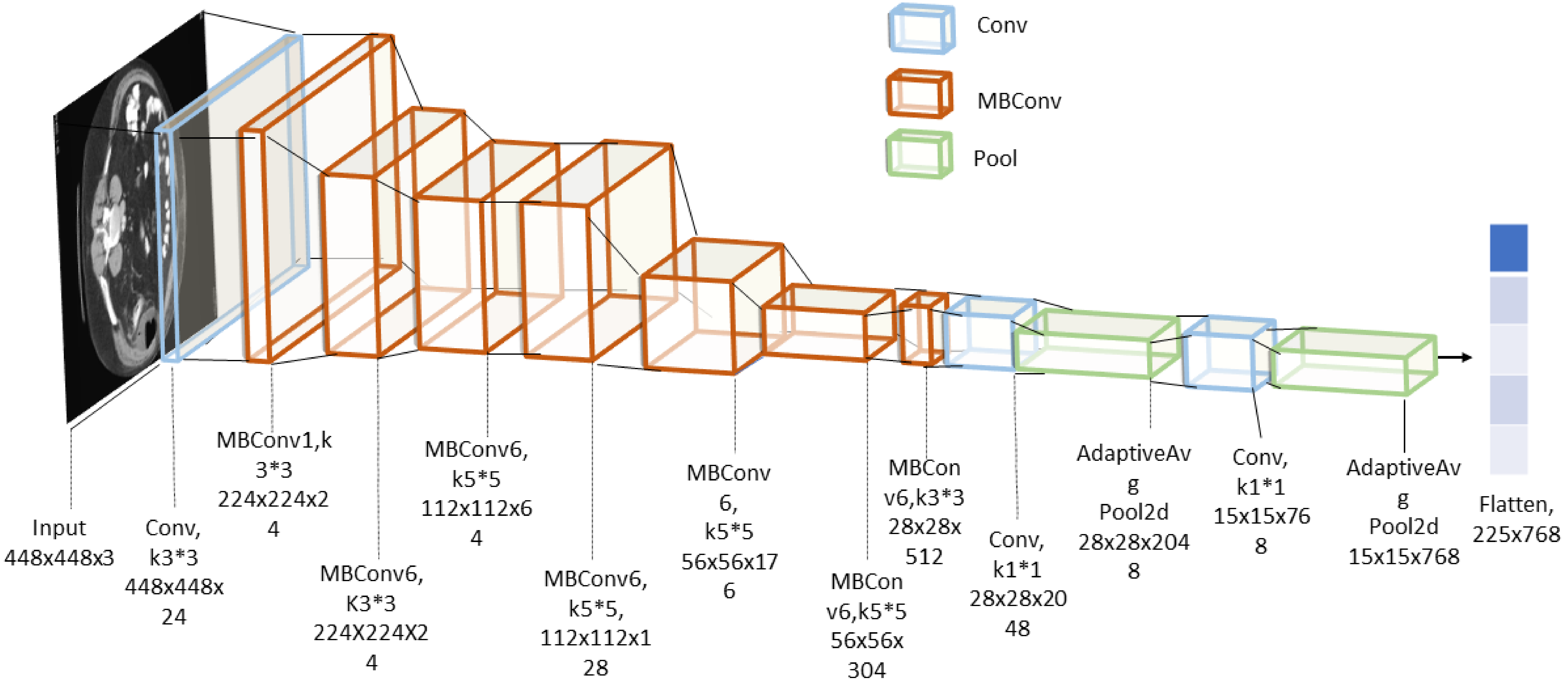
Methodology



Overview of model architecture

Methodology

Image Encoder



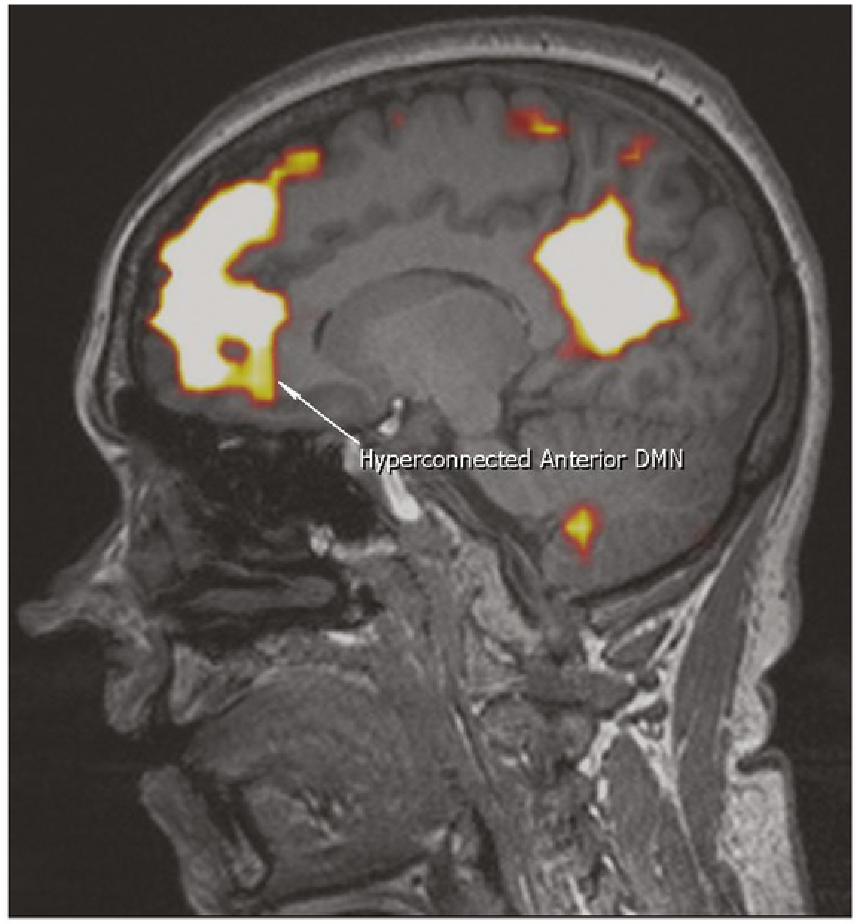
The architecture of EfficientNet-b5 model

Methodology

Text Encoder

Pre-trained: RadBERT-RoBERTa-4m
By: UCSD-VA-health

- Trained with 4 million radiology reports deidentified from US VA hospital



A

MRI OF THE BRAIN

HISTORY

Name: DOB: 12/28/1954 Female
Exam Date: 9/8/14
Referring Phys.: GunnarHeuser,M.D.

This is a 59-year-old female with exposure to mold and mercury. The patient has symptoms of seizures, memory loss, and numbness in hands and left arm.

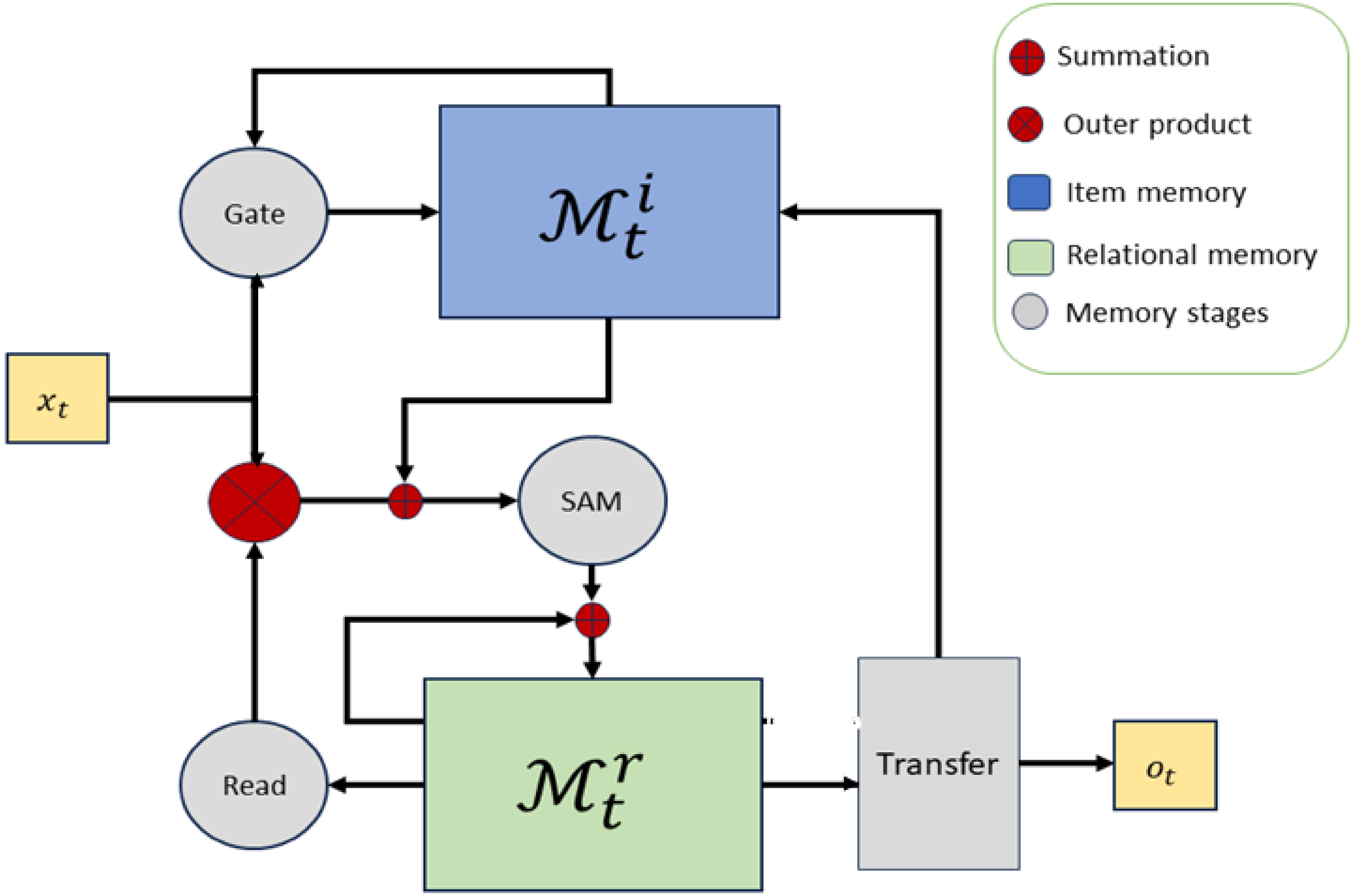
PROCEDURE

Using a 3 Tesla Siemens Verio MRI Open system, the following sequences were obtained:

1) Localizer.
2) T1 3D sagittal MPRAGE.
3) T2 FLAIR sagittal.
4) DWI axial.
5) SWI axial.
6) T2 FLAIR axial. 7) T2 TSE axial.

Source

Methodology



Overview of Attentive Memory Module

Methodology

Self-Attentive Memory

Outer Product Attention

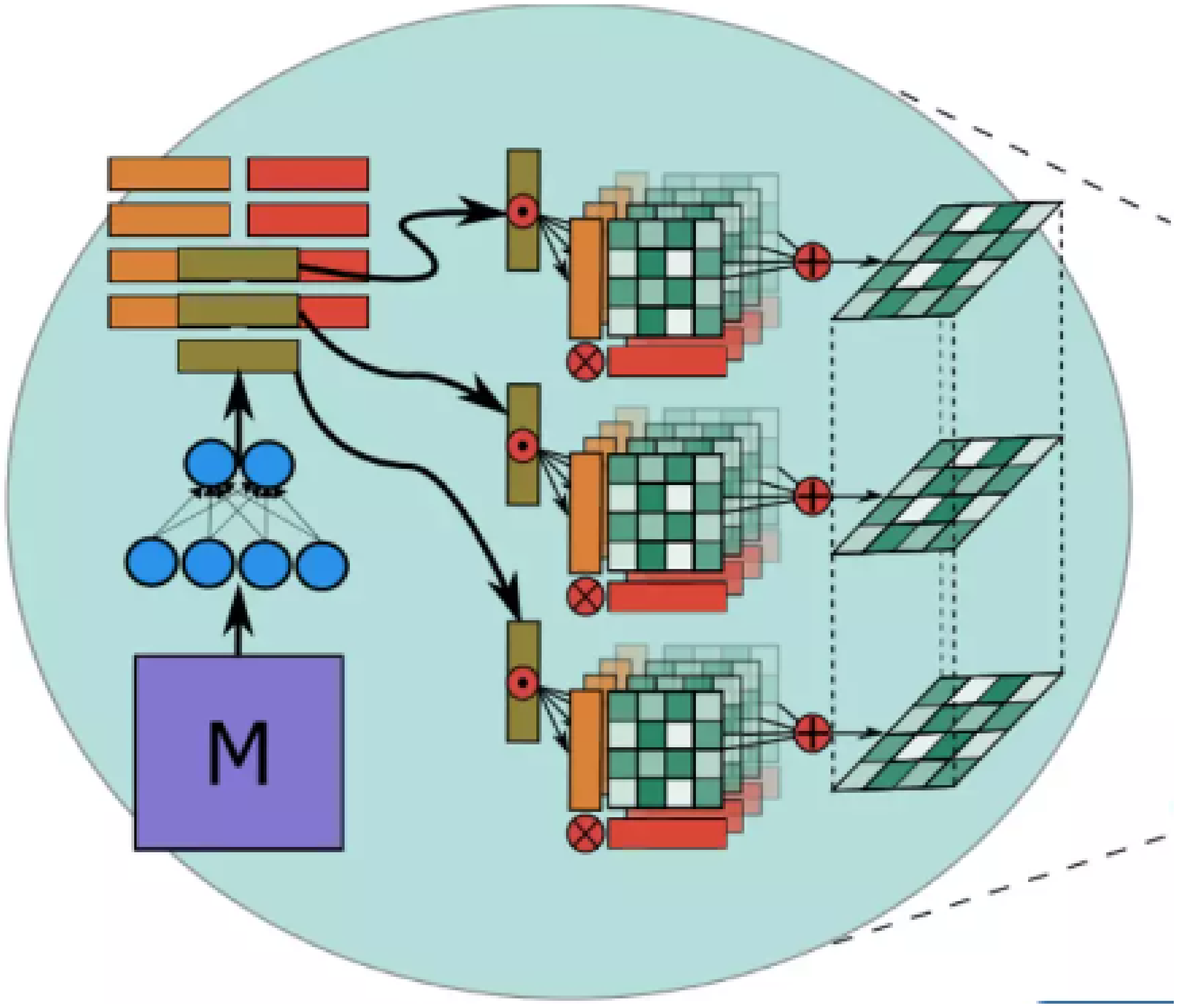
$$A^{\otimes}(q, K, V) = \sum_{i=1}^{n_{kv}} F(q \odot k_i) \otimes v_i$$

Where $A^{\otimes} \in \mathbb{R}^{d_{qk} \times d_v}; q, k_i \in \mathbb{R}^{d_{qk}}, v \in \mathbb{R}^{d_v}$, $\otimes$ is outer product, $\odot$ is element-wise multiplication and F is chosen as the tanh function.

Methodology

Self-Attentive Memory

Given memory input M:



$$M_q = LN(W_q M)$$

$$M_k = LN(W_k M)$$

$$M_v = LN(W_v M)$$

Extract items

$$SAM_\theta(M)[l] = A^{\otimes}(M_q[l], M_k, M_v)$$

Associate items

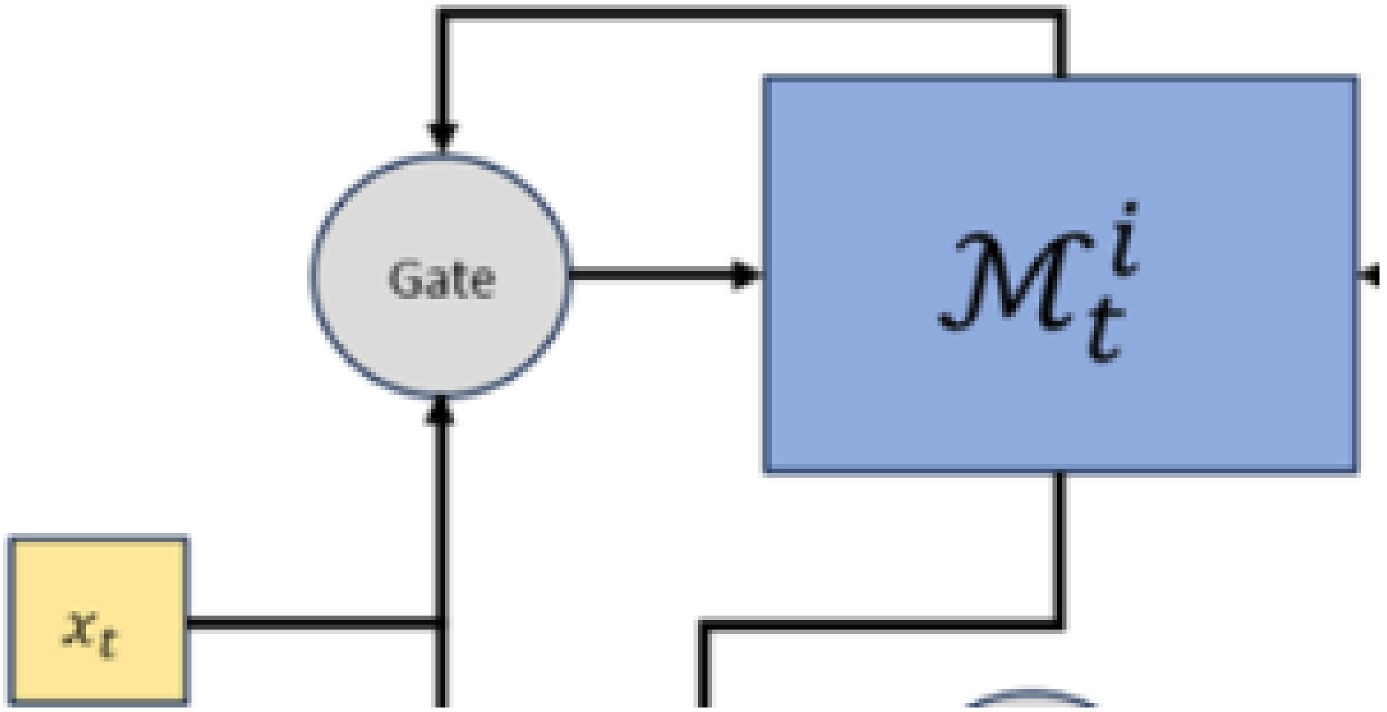
Where W_q, W_k, W_v is weight parameter, LN is Layer Normlization

Source: Hung el al (2020)

Methodology

Associative Memory Module

Construct item memory



$$X_t = f_1(x_t) \otimes f_2(x_t)$$

$$\mathcal{M}_t^i = \mathcal{M}_{t-1}^i + X_t$$

$$\mathcal{M}_t^i = F_t(\mathcal{M}_{t-1}^i, x_t) \odot \mathcal{M}_{t-1}^i + I_t(\mathcal{M}_{t-1}^i, x_t) \odot X_t$$

where $f1$ and $f2$ are fully connected neural networks

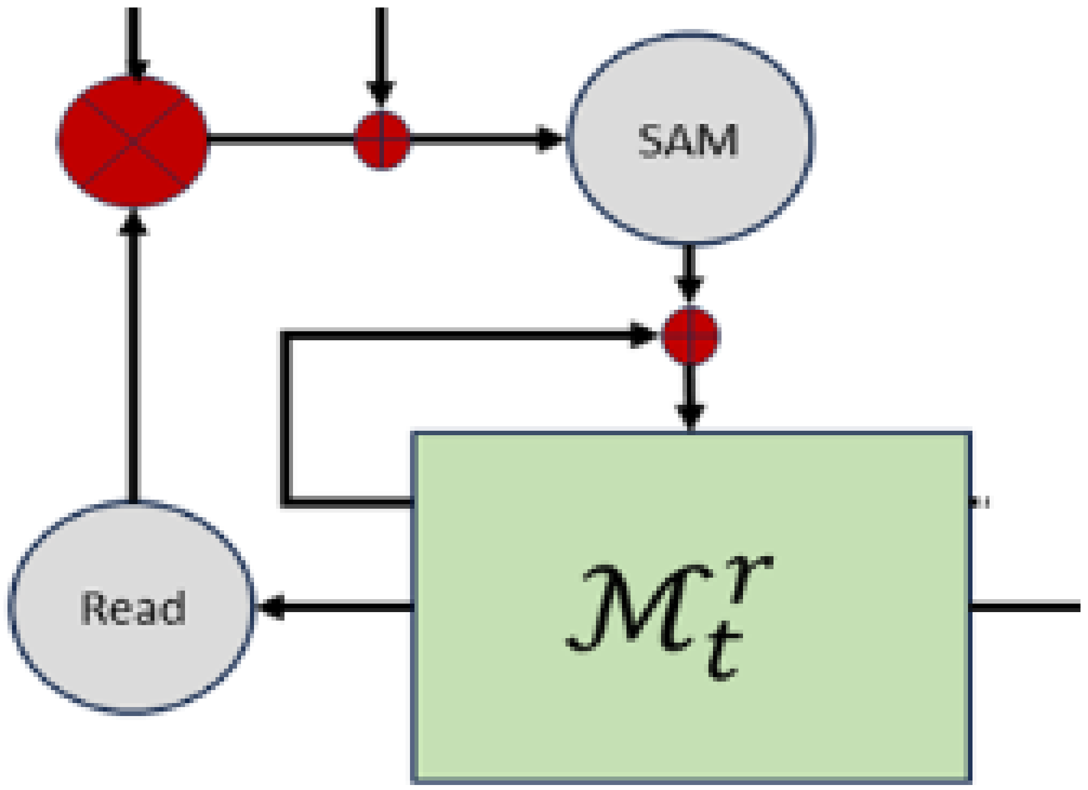
I_t and F_t are input and forget gate

and current input data x_t.

Methodology

Associative Memory Module

Construct relation memory



$$v_t^r = softmax(f_3(x_t)^\top)\mathcal{M}_{t-1}^r f_2(x_t)$$

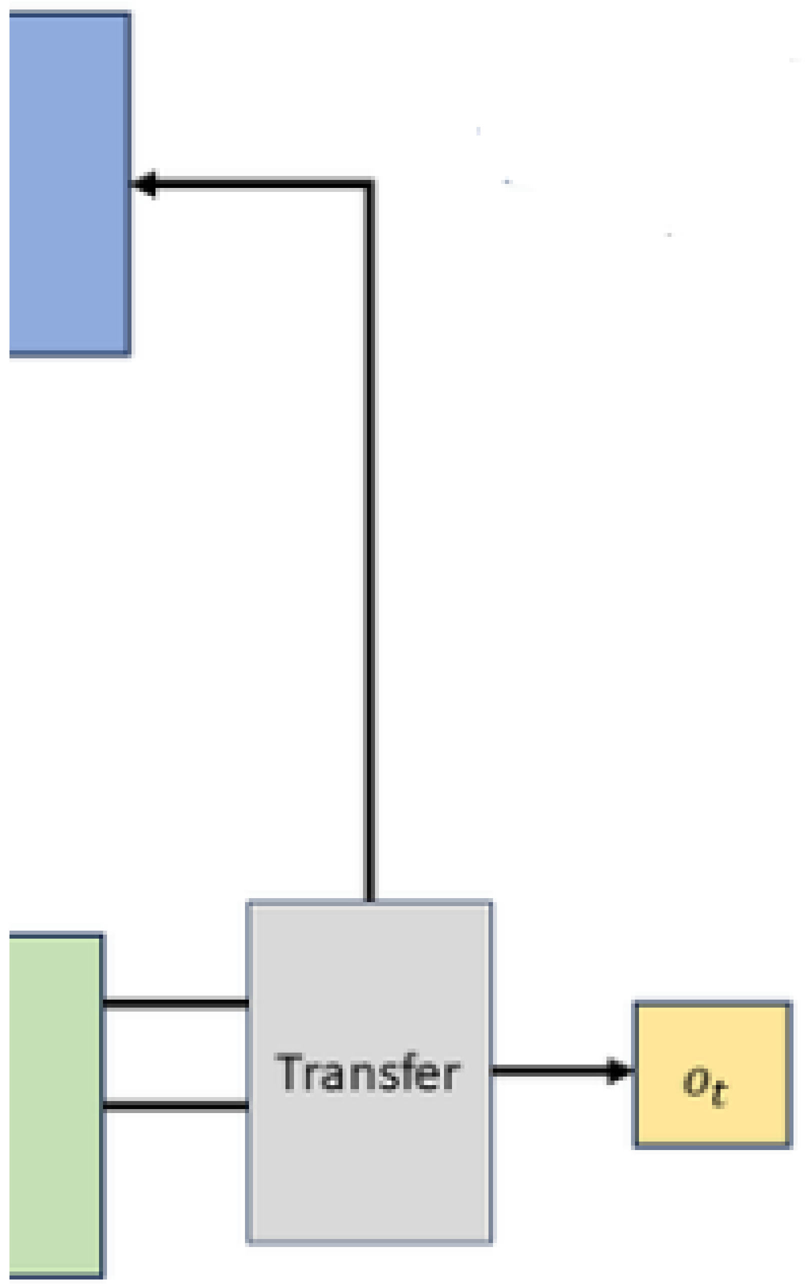
where f_3 is a fully connected neural network

$$\mathcal{M}_t^r = \mathcal{M}_{t-1}^r + \alpha_1 SAM_\theta(\mathcal{M}_t^i + \alpha_2\, v_t^r \otimes f_2(x_t))$$

where α_1 and α_2 are scaling hyper-parameters

Methodology

Associative Memory Module



$$\mathcal{M}_t^i = \mathcal{M}_t^i + \alpha_3 G_1 \circ V_f \circ \mathcal{M}_t^r$$

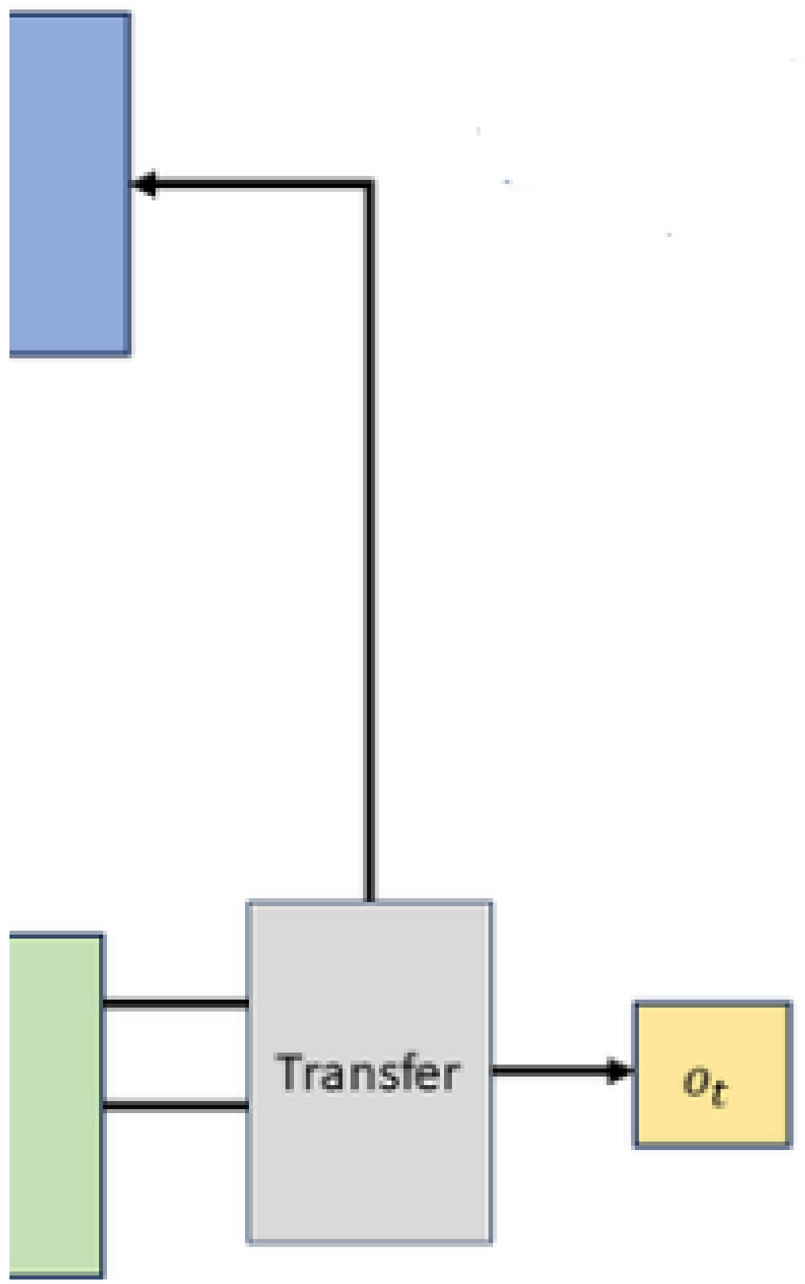
where V_f is a function use to the input tensor be flattens the first two dimensions.

G_1 is a Multilayer perceptron neural network that maps $\mathbb{R}^{(n_{kv} \times d) \times d} \rightarrow \mathbb{R}^{d \times d}$

α_3 is a combining hyper-parameter

Methodology

Associative Memory Module



$$o_t = G_2 \circ V_l \circ G_3 \circ V_l \circ \mathcal{M}_t^r$$

where V_l is a function that the input tensor flattens the last two dimensions

G_2 and G_3 are Fully Connected neural networks

Methodology

Fusion Module

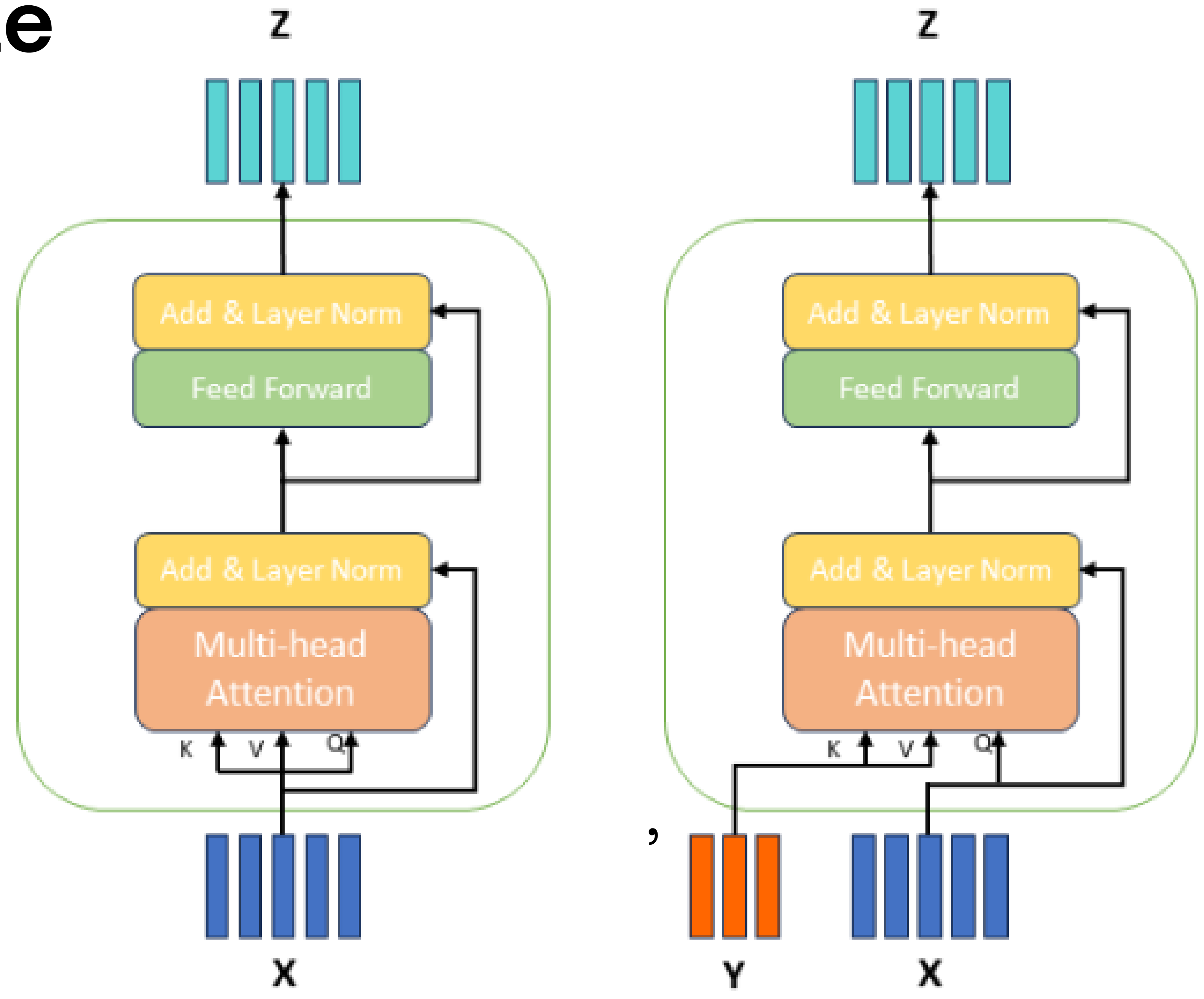


Figure: Self-Attention (left) and Cross-Attention (Right).

Methodology

Encoder-Decoder attention

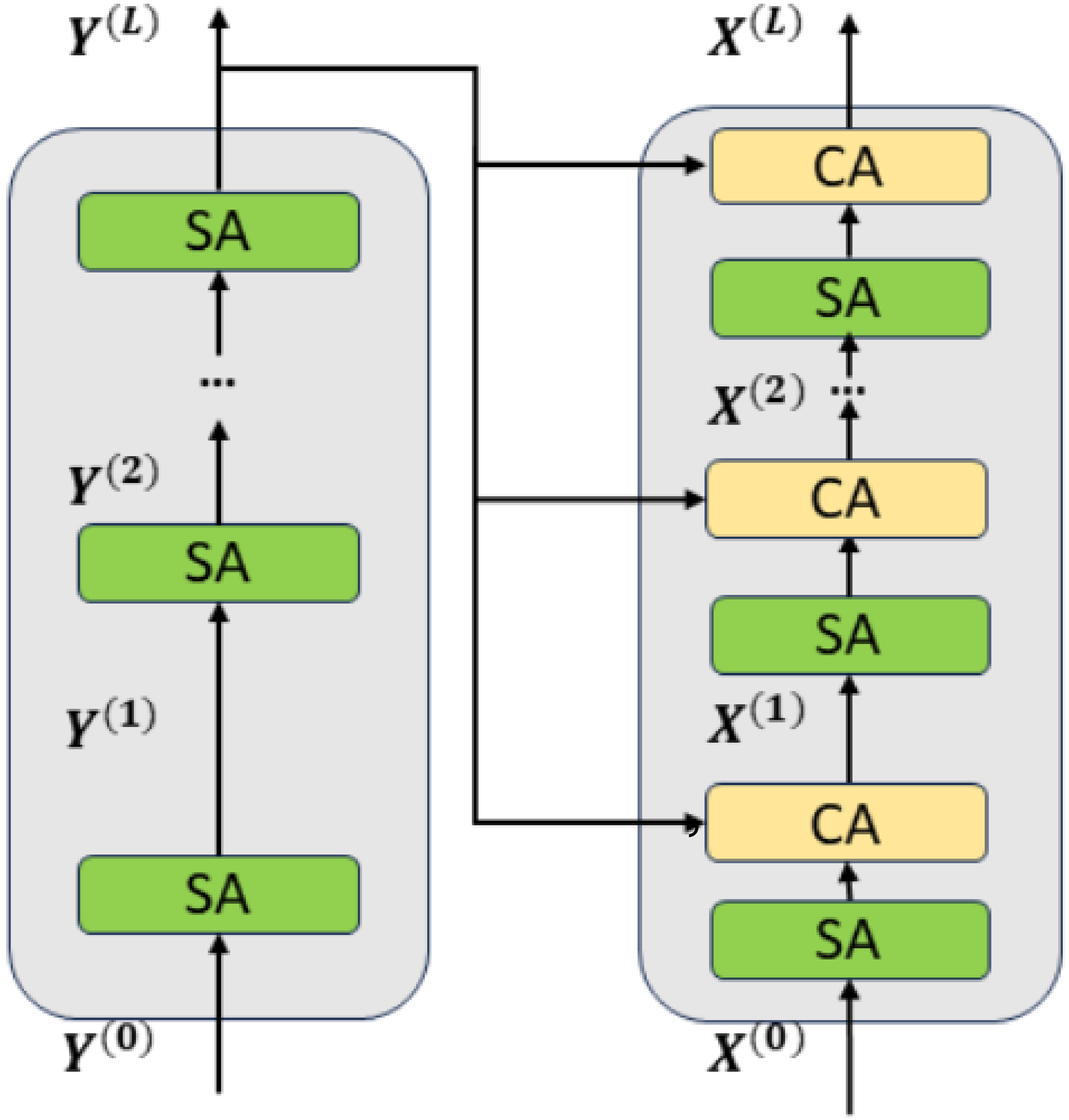


Figure: Encoder-Decoder attention.

Methodology

Prototype Learning Block

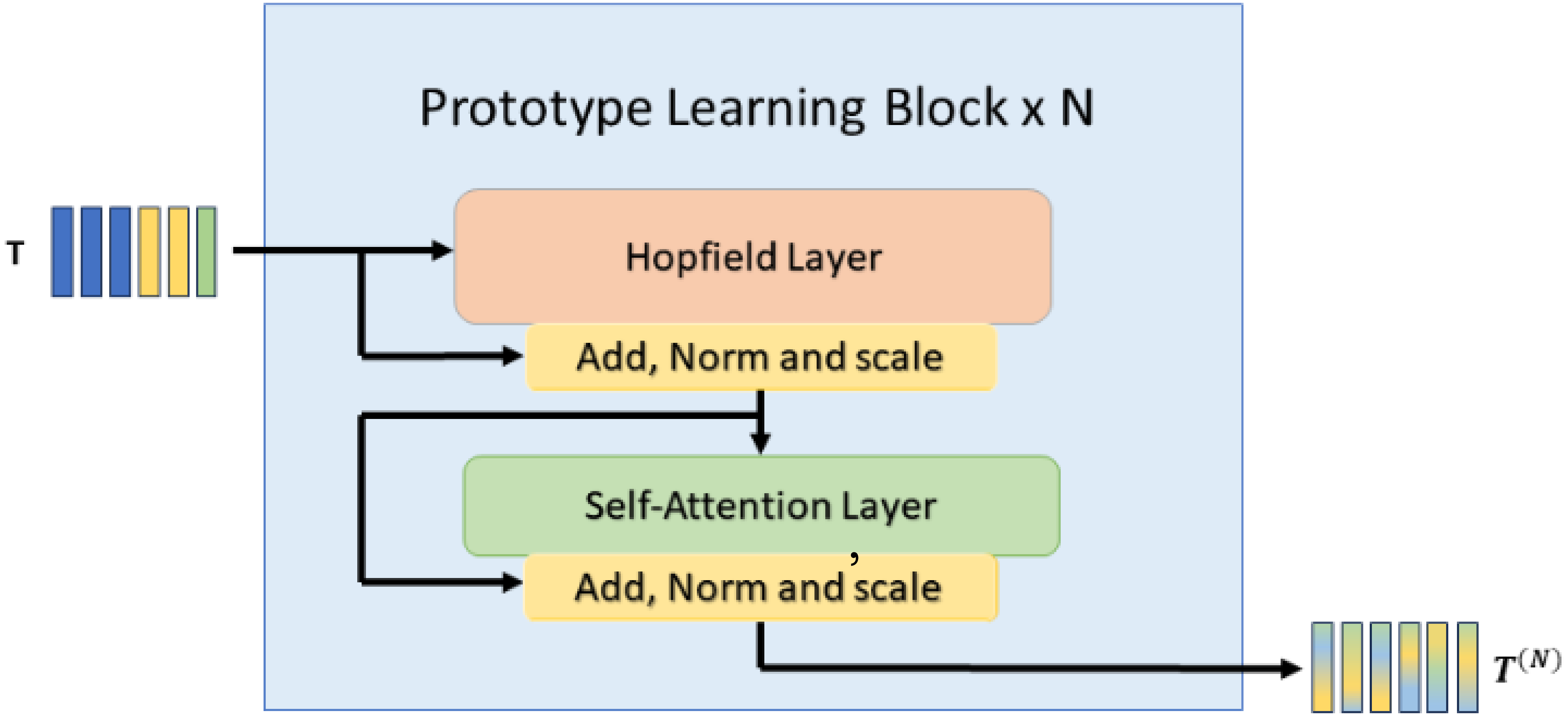


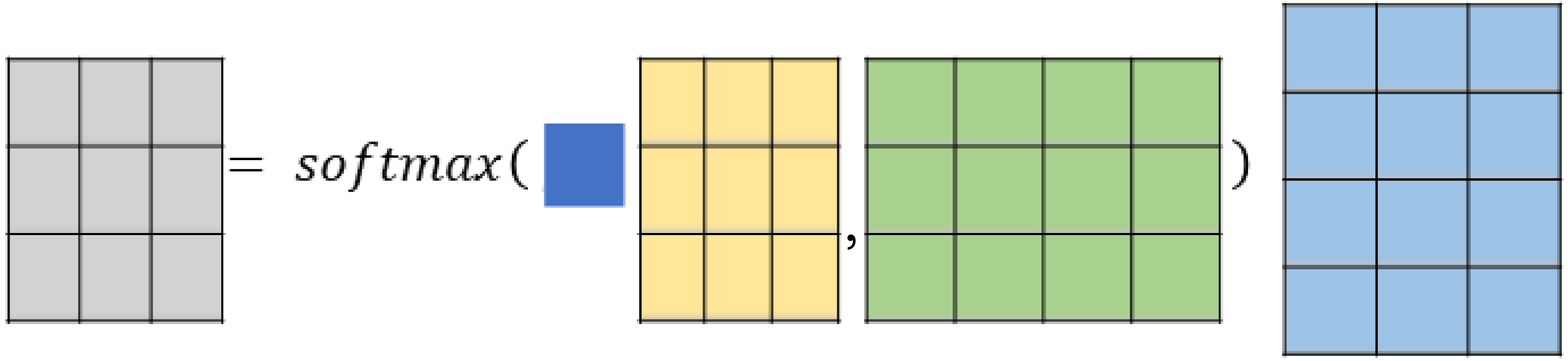
Figure: Detail of Prototype Leaning Block.

Methodology

Prototype Learning Block

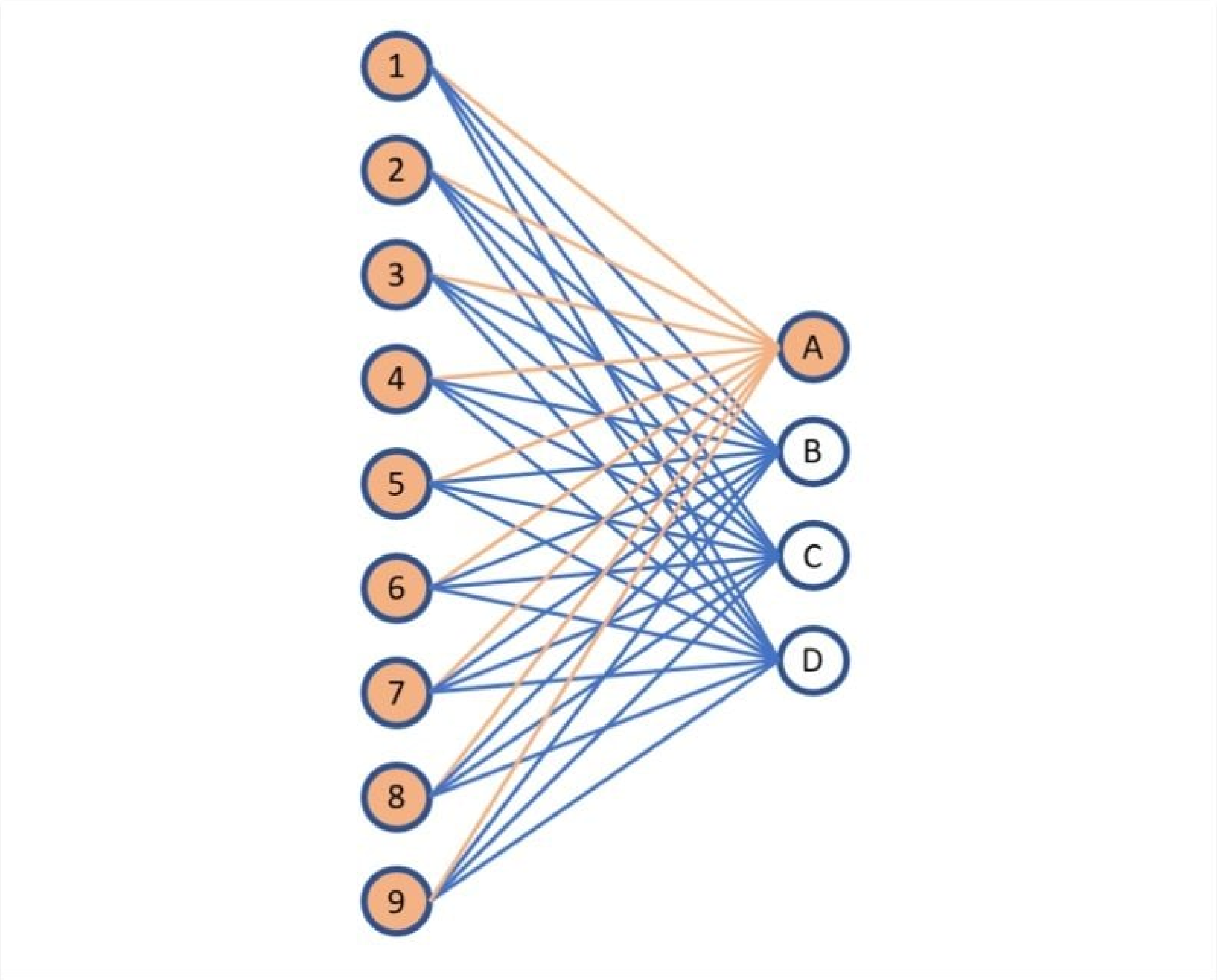
Formula of Hopfield layer with R is input

$$Z = softmax\left(\beta \; R \; W_{lookup}^{T}\right) W_{Store}$$



Methodology

Answer components



Fully Connected layer for classification

Image source: https://builtin.com/machine-learning/

Methodology

Loss function

Focal Loss:

$$L_{Focal}(p_t) = -(1-p_t)^{\gamma}log(p_t)$$

Image source: Lin el al (2017)

EXPERIMENTAL RESULT

EXPERIMENTAL RESULT

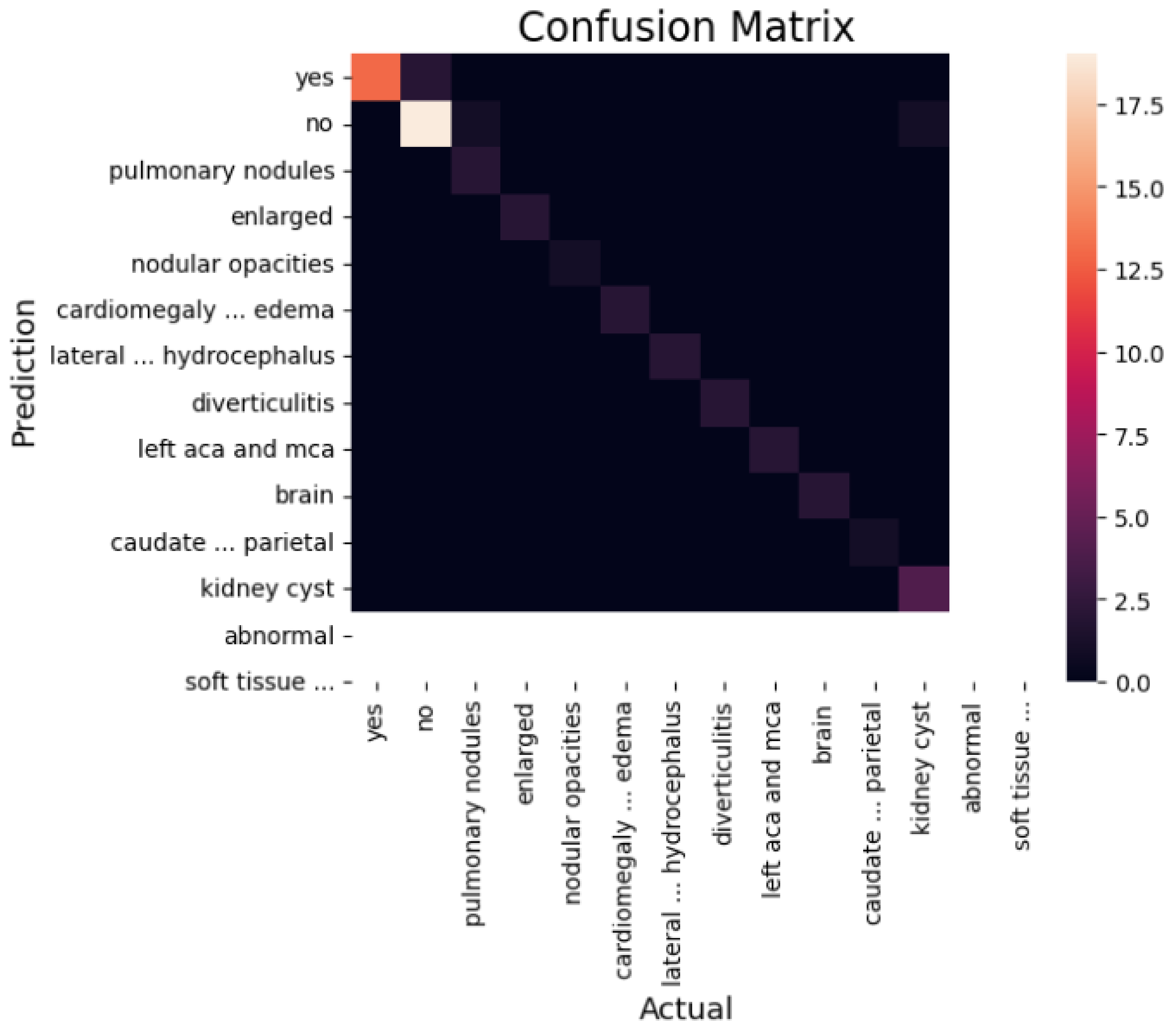
QS STARS RATED FOR EXCELLENCE 2012

Fpt University
TRƯỜNG ĐẠI HỌC FPT

Table 3. Comparisons our method with the state-of-the-art methods on the VQA-RAD test set

Methods	Closed	Open	Overall
BAN-VQAMix [*]	74.0	53.8	65.9
CMSA-MTPT [*]	77.3	56.1	68.8
MMQ-BAN [*]	75.8	53.7	67.0
FITS [*]	**82.0**	68.2	76.5
hi-VQA	-	-	76.3
Q2ATransformer	81.2	<u>79.19</u>	<u>80.48</u>
Ours	<u>81.98</u>	**79.39**	**80.93**

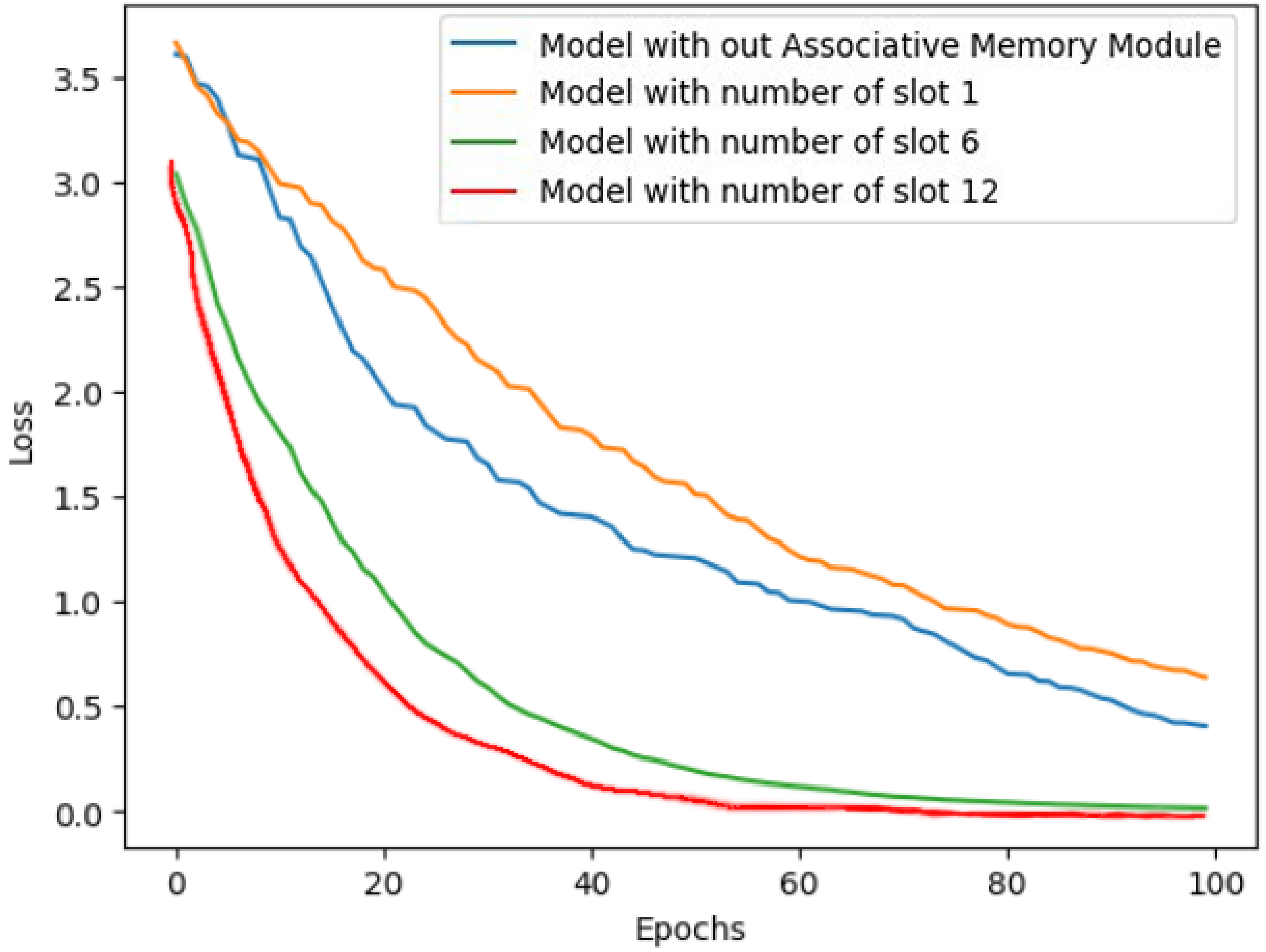
EXPERIMENTAL RESULT



EXPERIMENTAL RESULT

Model	Accuracy (%)	Average training time (s/epoch)
w/o AMM	62.4	61
$n_q = 1$	68.8	65
$n_q = 6$	75.2	96
$n_q = 12$	**79.7**	**119**

Comparison of models with different hyper parameters of AMM



Training process of model with/without AMM hyper-parameter modidication

EXPERIMENTAL RESULT

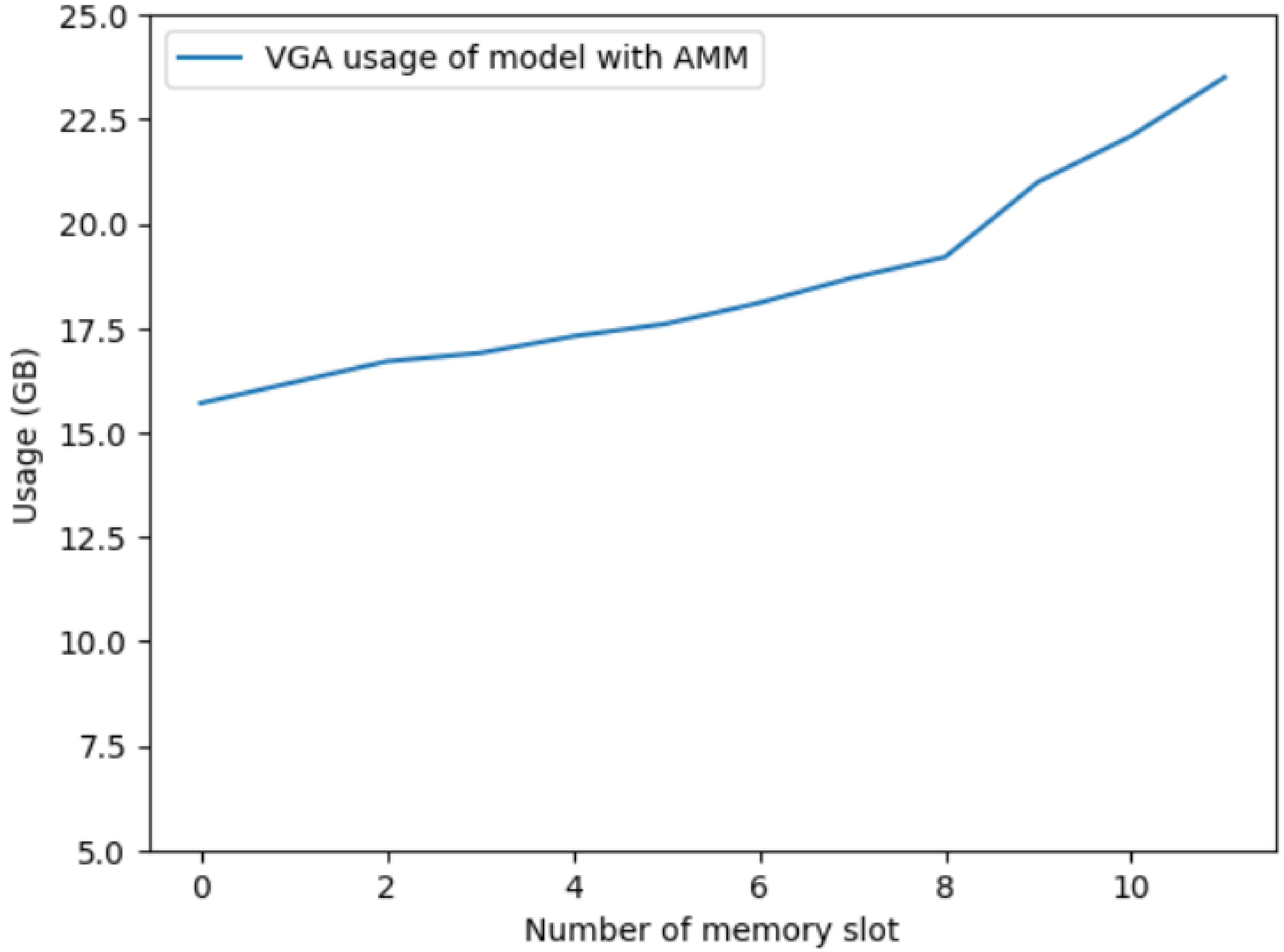


Figure 4.4: GPU consumption of model on VQA-RAD. The usage is calculated on the entire model process with batch size 16 and similar to the above hyper-parameter.

EXPERIMENTAL RESULT

QS STARS RATED FOR EXCELLENCE 2012

Fpt University

TRƯỜNG ĐẠI HỌC FPT

No of prototype/block	5	10	15
500	80.1	80.47	79.96
1000	80.24	**80.93**	80.24
1500	80.18	80.51	80.04

The model accuracy (%) of each set number prototype and number of block prototype learning.

CONCLUSION

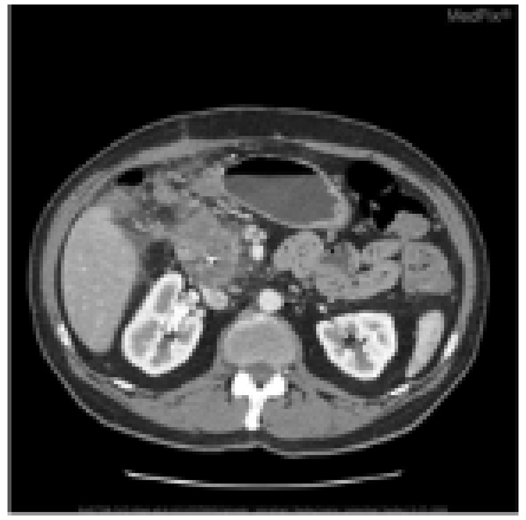
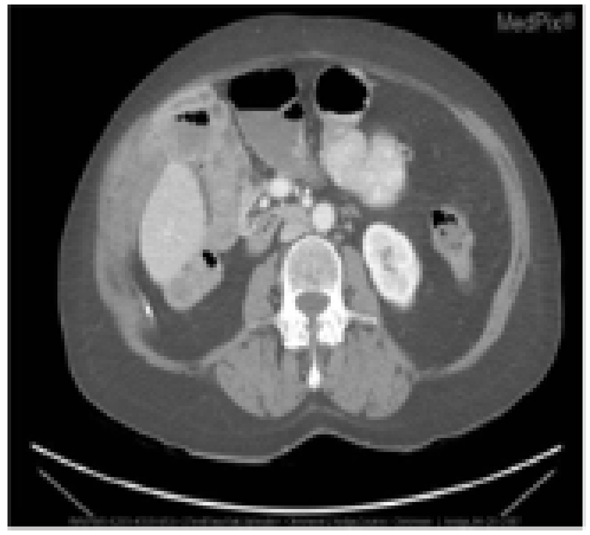
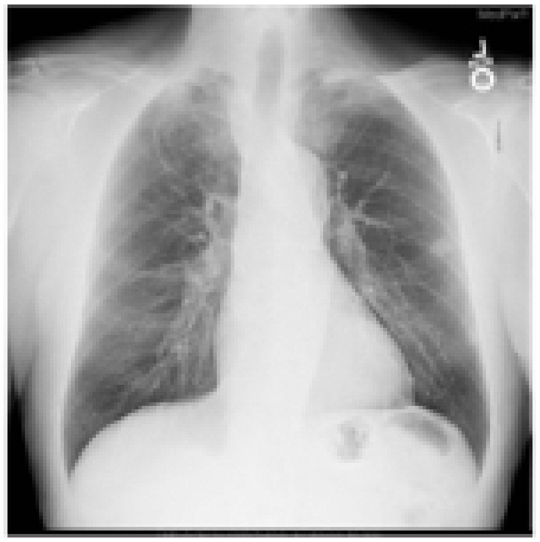
CONCLUSION

- An architecture in medical VQA based on Associative Memory and Prototype Learning.
- The result is not significantly improved.

FUTURE WORK

- Experiment on other datasets with similar limitations and improve the model.
- Experiment on some data augmentation techniques to enrich the datasets.

Visualization

				
Question:	What is the location of the mass?	Where is the colon most prominent from this view?	which organ system is abnormal in this image?	Is the diaphragm flat on either side?
Answer:	Head of the pancreas	Left	cardiovascular	No
Question Category:	Positional	Location	Modality	Yes/No
Q2A-Tranformer	Head of the pancreas	Right	Lung	Yes
Our Model:	Head of the pancreas	Left	Right lung	Yes