TYPE-AWARE MEDICAL VISUAL QUESTION ANSWERING

Anda Zhang[1] *Wei Tao*[1] *Ziyan Li*[1] *Haofen Wang*[2⋆] *Wenqiang Zhang*[1⋆]

[1]Academy for Engineering and Technology, Fudan University, Shanghai, China
[2]College of Design and Innovation, Tongji University, Shanghai, China

ABSTRACT

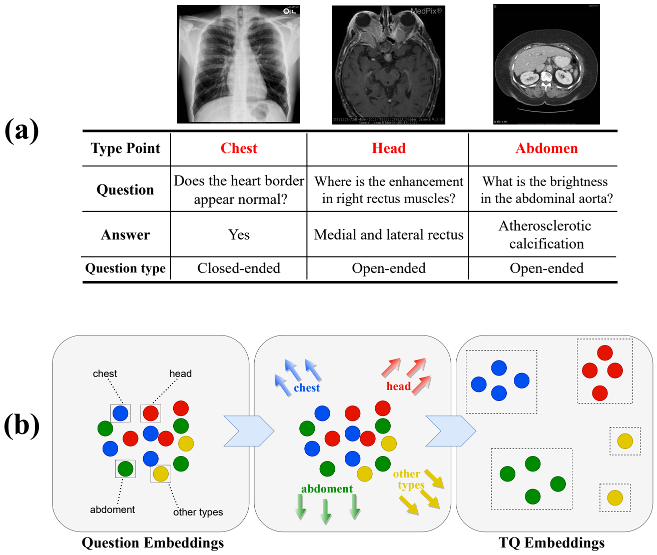
Medical Visual Question Answering (Med-VQA) helps answer medical questions raised by patients automatically so as to relieve the shortage of experienced doctors. Cross-modal feature alignment is a major challenge of Med-VQA. Moreover, it is critical to exploit sufficient semantic features with the consideration of characteristic of medical images and language. In this paper, we propose a novel From Image type point To Sentence (FITS) method to tackle the above challenge. In particular, the type of the medical images is represented as a type point which is further considered in the question sentence representation. The combined representation aims to optimize the feature distribution in an embedding space and thus enhances the ability of semantic alignment. Type point is also used in two feature extraction modules for medical questions and images respectively, which can efficiently improve the reasoning ability of different modalities, and further enhance the applicability of the fusion method for Med-VQA. The experimental results show that FITS outperforms all the previous approaches in terms of accuracy especially in open-ended questions significantly.

Index Terms— Medical visual question answering, multi-modal transformer, feature alignment, representation learning

1. INTRODUCTION

Medical **V**isual **Q**uestion **A**nswering (Med-VQA) aims to predict the answers to medical questions through the understanding and reasoning of vision (image) and natural language (question). It can provide convenience for patients and ease the shortage of medical resources.

The current Med-VQA framework is mainly composed of feature extraction and cross-modal fusion. Most Med-VQA methods [1, 2, 3, 4, 5] used LSTM [6] or GRU [7] to learn the representation of questions and ResNet [8] to learn the the representation of images. The methods based on the attention mechanism were applied as cross-modal fusion strategy, such as SAN [9] or BAN [10]. Particularly, Li et al. [3] proposed the conditioned reasoning method and classified medical questions into two types to train separate models. Gong et al. [4] introduced a cross-modal self-attention (CMSA) module and formulated multi-task pre-training for Med-VQA. Inspired by the widespread application of transformer [11] in NLP tasks, Khare et al. proposed MMBERT [12] which used a small-scale vision and language pre-trained method. MMBERT concatenates image features and text features directly, and this simple concatenation is served as an input for semantic alignments.



Type Point	Chest	Head	Abdomen
Question	Does the heart border appear normal?	Where is the enhancement in right rectus muscles?	What is the brightness in the abdominal aorta?
Answer	Yes	Medial and lateral rectus	Atherosclerotic calcification
Question type	Closed-ended	Open-ended	Open-ended

Fig. 1. (a) Examples of Med-VQA. Type Point (red font) means the type of medical images. (b) The distribution of different type questions in the embedding space.

Although recent works improved performance on Med-VQA, the researchers paid little attention to semantic alignment and overlooked some characteristics of Med-VQA. In Med-VQA, images are taken with medical equipment for a specified part of the human body, which is different from images in the general VQA task. As shown in Fig 1 (a), medical images can be classified by parts of the human body. Just as many medical terms are only used for a specific part, images of different types correspond to different questions. From that observation, we encode these questions into the low-dimensional embedding space [13] and find the distribution of questions is chaotic and irregular as shown in the left part of Fig 1 (b). We find that the embeddings achieve

⋆ Corresponding author. This work was supported by National Natural Science Foundation of China (No.62176185)

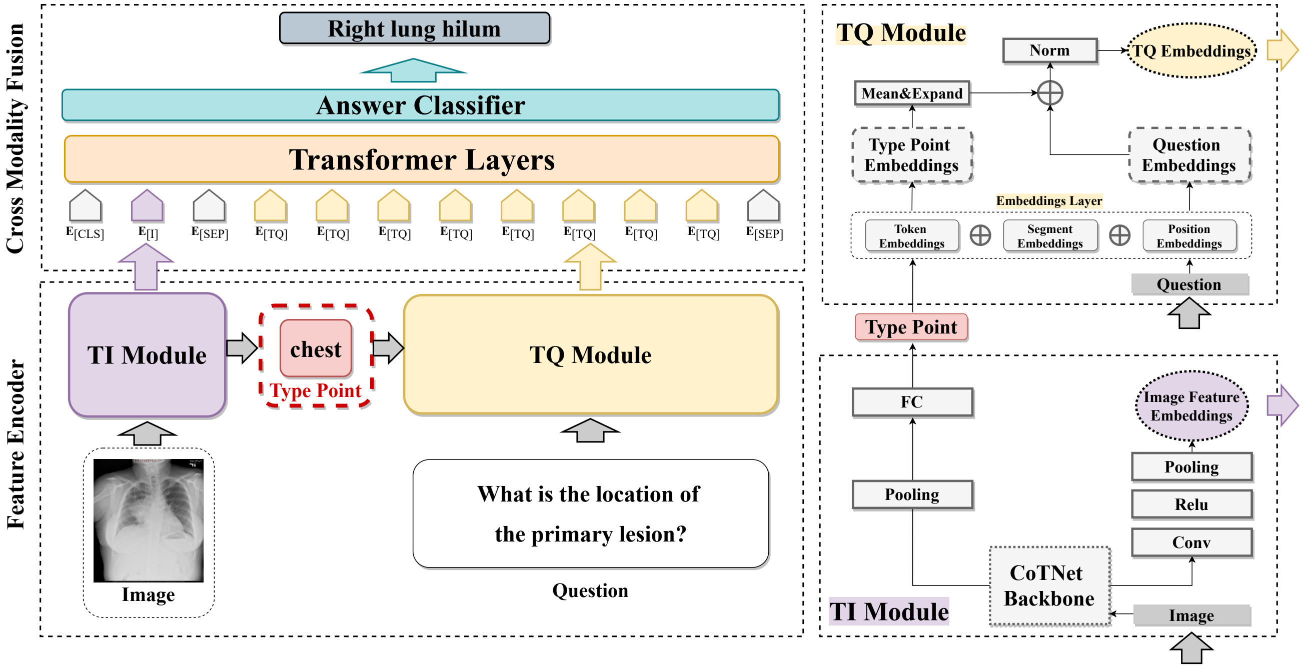


Fig. 2. Model architecture. The two dashed boxes on the left represent Feature Encoder and Cross Modality Fusion. The two modules on the right represent TQ and TI modules. In TI module, the left branch predicts type point, and the right generates image representation. TQ module generates TQ Embeddings from question and type point.

spatial relativity and a more regular distribution as shown in the right part of Fig 1 (b) after adding an offset to the question embeddings of each type. This distribution can help the model learn implicit knowledge of the image type in advance and further enhance the ability of semantic alignment. This finding enlightens us to consider the importance of "type" information in the representation and fusion module. The type of images is employed as a bridge between text features and image features and we call it as **type point**. The new embedding includes both **T**ype point and **Q**uestion, so we refer it as TQ Embedding. Based on the type point, we design TQ module as text feature encoders and TI (**T**ype point and **I**mage) module as image feature encoders for Med-VQA.

To summarize, our contributions in this paper include:

- We design an image feature extraction module with double-branch for medical images. In this module, we fully capitalize the image contextual features by the first branch, and predict the type point of the image as the input for the text representation module by another branch. Thus, the characteristics of medical images are exploited efficiently.

- We design a text representation module for Med-VQA. In this module, we jointly learn the type point embeddings and question embeddings to generate a novel representation, which improves the ability of semantic alignment between different modalities and further enhance the applicability of the fusion method for Med-VQA.

- A novel methodology achieves state-of-the-art performance. Extensive experiments on VQA-RAD dataset [14] demonstrate the effectiveness of **F**rom **I**mage type point **T**o **S**entence (**FITS**) over the previous state-of-the-art baselines.

2. METHODOLOGY

An overview of our model architecture is illustrated in Fig 2.

2.1. Problem formulation

Given a question $q \in \mathcal{Q}$ grounded in an image $v \in \mathcal{I}$, the goal is to predict a correct answer $a \in \mathcal{A}$. F is our model. $\hat{t}$ is the type point of our prediction by TI module. Thus, the predicted answer $\hat{a}$ is formulated as follows:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} F(a \mid v, q, \hat{t}). \tag{1}$$

2.2. Feature Encoder

2.2.1. TI Module

Contextual Transformer (CoT) block [15] is a transformer-style module for vision tasks which can fully capitalize on the contextual information. Considering the global characteristics of medical images, we use CoTNet152 [15] as the backbone in TI module. We design a compatible module with double-branch to achieve effective representation of medical images in cross-modal fusion.

The first branch of TI module is a separate pre-trained classification task which predicts the type point $\hat{t}$. The $\hat{t}$ can be formulated as:

$$\hat{t} = \arg\max_{t \in \mathcal{T}} F_c(t \mid v), \tag{2}$$

where $t \in \mathcal{T}$ denotes the candidate type point (chest, head, abdomen), $v \in \mathcal{I}$ denotes a image and F_c denotes a classification model. To pretrain the classification model, we crawled 727 medical images of abdomen, chest, and head from PEIR Digital Library[1]. We employ these images to train the classification model, and test its performance on all 315 images in VQA-RAD [14] dataset. Although the number of training sets is small, the task still achieves a 96% accuracy.

The another branch loads the weights of the first branch to fine-tune for medical image feature extraction. In this branch, we do not employ multi-scale image feature maps to avoid introducing more parameters and reduced computational effort. Hence we obtain a 2048-D feature vectors V_I of the image v through the CoTNet backbone CoT and downsample the vectors of the image to the hidden size 768-D by a 1x1 convolution kernel. The reduced-dimensional features E_I are used as image features embeddings. The E_I can be formulated as:

$$V_I = CoT(v) \tag{3}$$

$$E_I = Pooling(Conv(V_I)), \tag{4}$$

In addition, in both branches, we add a global average pooling layer to reduce the number of FC (Fully Connected) parameters, incidentally use adequate global information, and prevent overfitting.

2.2.2. *TQ Module*

BioM-Electra [16] is a biomedical language model, which is pre-trained on biomedical domain corpora based on ELECTRA architecture [17]. Since most of the question sentences in Med-VQA are related to biomedical language, we utilize BioM-Electra-base Embeddings layer to generate separate representations for type point and questions. We figure these representations to generate TQ embeddings. The embeddings E_{TQ} are calculated by:

$$E_{TQ} = ME(E_{tp}) \oplus E_q, \tag{5}$$

where $E_{tp} \in \mathbb{R}^{n_{tp} \times 768}$ denotes Type Point Embeddings and $E_q \in \mathbb{R}^{n_q \times 768}$ denotes Question Embeddings. n_{tp} and n_q represent the length for type point token and question token, respectively. ME denotes calculating the mean of n_{tp}-D and expanding the mean in n_q-D dimensions. In other words, we employ the representation of type point as a guidance to make question embeddings offset in the embedding space. Finally, the converted E_{tp} are added to the E_q to generate the new embeddings E_{TQ}.

2.3. Cross Modality Fusion

The architecture of fusion includes 6 layers of transformer blocks with 12 attention heads and a hidden size of 768. Inspired by Pre-LN Transformer layer [18], we perform layer normalization before the multi-headed attention layer and the fully connected feedforward network. This normalization prevents overfitting and resolves the problem of slow convergence of the optimization process. The answer $\hat{a}$ can be formulated as:

$$E = \{E_{CLS}, E_I, E_{SEP}, E_{TQ}, E_{SEP}\} \tag{6}$$

$$\hat{a} = Classifier(Transformer(E)), \tag{7}$$

where E denotes the joint embeddings after fusion of image feature embeddings E_I and TQ embeddings E_{TQ}. As shown in the upper left dotted box of Fig 2, we encode [CLS] token and [SEP] token into E_{CLS} and E_{SEP}, and we use E_{CLS} for classification at the beginning and separate E_I and E_{TQ} with E_{SEP}. Finally, E is input as joint representation into transformer block and the answer $\hat{a}$ is predicted by the classifier.

3. DATASET AND EXPERIMENTS

3.1. Dataset

We conduct experiments on VQA-RAD [14] dataset which is used by most Med-VQA works [1, 2, 3, 4, 5]. VQA-RAD contains 315 radiology images and 3515 question-answer pairs provided by clinicians (3064 for train and 451 for test). Questions are classified into two types: closed-ended (2093) and open-ended (1422). Closed-ended questions have limited choices such as “yes/no” while the open-ended do not. We illustrate 3 examples in Fig 1 (a).

3.2. Experimental Setup

We use PyTorch [19] as our deep learning framework. In both models, cross-entropy is applied as loss function. In the pre-trained task, we use the SGD optimizer [20] with an initial learning rate at 2e-2. We set the weight decay to 2e-4 and set the momentum to 0.9. The batch size is 64 and the number of epochs is 15. In addition, the Med-VQA model is trained with Adam optimizer [21] with an initial learning rate at 7e-5. The batch size is 64 and the number of epochs is 100. The learning rate reduces by a factor of 0.1 if the loss does not decrease for consecutive 10 epochs.

3.3. Evaluation and Comparison with other methods

Followed by the previous works [1, 2, 3, 4, 5], accuracy is used as our evaluation metric and we evaluate our method in two kinds of answer combination settings. Table 1 shows the accuracy of our complete model on VQA-RAD test set.

[1] `https://peir.path.uab.edu/library/index.php?/category/106`

Table 1. Accuracy comparison with other methods on VQA-RAD dataset. "free" denotes using only "freeform" answer type in test set and "free+para" denotes using both. "Dedicated Model" denotes whether the method treats open-ended and closed-ended questions as separate models when training, and our method uses only a dedicated model for all kinds of questions.

Method	Answer Type	Dedicated Model	Open	Closed	Overall
MEVF-SAN [1]	free	√	40.7	74.1	60.8
MEVF-BAN [1]	free	√	43.9	75.1	62.7
MMQ-BAN [2]	free	√	53.7	75.8	67.0
MTPT-CMSA [4]	free	√	**56.1**	**77.3**	**68.8**
FITS (ours)	free	√	**61.0**	**80.0**	**72.4**
MEVF-BAN-CR [3]	free+para	×	60.0	79.3	71.6
MMBERT [12]	free+para	√	**63.1**	77.9	72.0
MTPT-CMSA-CR [4]	free+para	×	61.5	**80.9**	**73.2**
FITS (ours)	free+para	√	**68.2**	**82.0**	**76.5**

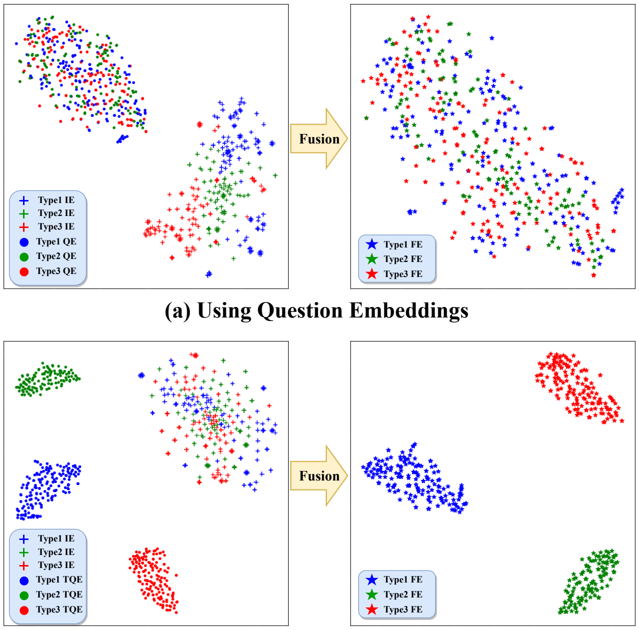
Table 2. Ablation study of FITS on VQA-RAD dataset. "w/o Type Point" denotes that we use question embedding in the fusion section without using TQ embedding.

Method	Answer Type	Open	Closed	Overall
FITS	free	**61.0**	**80.0**	**72.4**
w/o Type Point	free	57.7	78.4	70.1
FITS	free+para	**68.2**	**82.0**	**76.5**
w/o Type Point	free+para	63.1	81.6	74.3

Our method outperforms all the baselines in each combination setting. For one setting where the test set only contains "freeform" answer types of samples, our method exceeds the previous state-of-the-art model MTPT-CMSA [4] by 4.9%, 2.7%, and 3.6% for closed-ended, open-ended, and overall questions, respectively. Both MMBERT [12] and our method are trained in the dedicated model setting. For another answer combination setting, we compare with MMBERT for a fair comparison. Our method improves accuracy by 5.1%, 4.1%, and 4.5% on open-ended, closed-ended, and overall questions. In additon, although MTPT-CMSA-CR [4] is trained in separate model settings, our model outperforms them by 6.7%, 1.1%, and 3.3% in the open-ended, close-ended, and overall questions, respectively. The experimental results prove that our method is effective for both "freeform" and "para" answer types. In conclusion, our method outperforms all of the previous baselines in accuracy and has a substantial performance improvement on open-ended questions.

3.4. Ablation Study and Analysis

In our ablation study, Table 2 shows a consistent increase of accuracy compared to the model without type point in all types of questions, especially in open-ended questions. To further verify the effectiveness of each module in our method, we use t-SNE algorithm [22] for reducing dimension to observe the relative positions of image embeddings and text embeddings in the embedding space. As shown in Fig 3, encoding only the questions as text representation makes the embeddings distribute irregularly before and after fusion. Instead, combining the encoding of both the questions and the type point as text representation makes the embeddings distribute more regularly. More importantly, the joint embeddings after the fusion also have a regular distribution, which make the model learn implicit knowledge of the image type in advance. Thus, the novel representation provides excellent assistance for predicting the final answer. Note that the design of our method is not limited to the number of image types and the three types are chosen to experiment because of the dataset.



(a) Using Question Embeddings

(b) Using TQ Embeddings

Fig. 3. Illustration of embedding distribution in 2-D space. IE, QE, TQE, and FE represent Image Embeddings, Question Embeddings, TQ Embeddings, and the joint Embeddings after fusion. Type represents image type. The right part represents the distribution after fusion.

4. CONCLUSION

In this paper, we propose a new method for Med-VQA, FITS, which fully exploits semantic features of medical data and enhances the ability of cross-modal semantic alignment. The experimental results show that our method performs significantly and consistently better than other baselines on open-ended, closed-ended, and overall questions. In the future, we plan to improve the compatibility of FITS to resolve the semantic alignment problem in other medical vision-language tasks.

5. REFERENCES

[1] Binh D. Nguyen, Thanh-Toan Do, Binh X. Nguyen, Tuong Do, Erman Tjiputra, and Quang D. Tran, "Overcoming data limitation in medical visual question answering," in *MICCAI (4)*. 2019, vol. 11767 of *Lecture Notes in Computer Science*, pp. 522–530, Springer.

[2] Tuong Do, Binh X. Nguyen, Erman Tjiputra, Minh Tran, Quang D. Tran, and Anh Nguyen, "Multiple meta-model quantifying for medical visual question answering," in *MICCAI (5)*. 2021, vol. 12905 of *Lecture Notes in Computer Science*, pp. 64–74, Springer.

[3] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu, "Medical visual question answering via conditional reasoning," in *ACM Multimedia*. 2020, pp. 2345–2354, ACM.

[4] Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li, "Cross-modal self-attention with multi-task pre-training for medical visual question answering," in *ICMR*. 2021, pp. 456–460, ACM.

[5] Haiwei Pan, Shuning He, Kejia Zhang, Bo Qu, Chunling Chen, and Kun Shi, "Muvam: A multi-view attention-based model for medical visual question answering," *arXiv Preprint*, 2021.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP*. 2014, pp. 1724–1734, ACL.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*. 2016, pp. 770–778, IEEE Computer Society.

[9] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola, "Stacked attention networks for image question answering," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 2016, pp. 21–29, IEEE Computer Society.

[10] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang, "Bilinear attention networks," in *NeurIPS*, 2018, pp. 1571–1581.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[12] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U. Deva Priyakumar, and C. V. Jawahar, "MMBERT: multimodal BERT pretraining for improved medical VQA," in *ISBI*. 2021, pp. 1033–1036, IEEE.

[13] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.

[14] J. J. Lau, S. Gayen, A. B. Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific Data*, vol. 5, pp. 180251, 2018.

[15] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei, "Contextual transformer networks for visual recognition," *arXiv Preprint*, 2021.

[16] Sultan Alrowili and Vijay Shanker, "Biomtransformers: Building large biomedical language models with bert, ALBERT and ELECTRA," in *BioNLP@NAACL-HLT*. 2021, pp. 221–227, Association for Computational Linguistics.

[17] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning, "ELECTRA: pre-training text encoders as discriminators rather than generators," in *ICLR*. 2020, OpenReview.net.

[18] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu, "On layer normalization in the transformer architecture," in *ICML*. 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533, PMLR.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019, pp. 8024–8035.

[20] H. Robbins and S. Monro, *A Stochastic Approximation Method*, Herbert Robbins Selected Papers, 1985.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[22] Hinton Laurens, Van Der Maaten and Geoffrey, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2605, pp. 2579–2605, 2008.