MEDICAL VQA: MIXUP HELPS KEEP IT SIMPLE

Preprint · March 2022

DOI: 10.13140/RG.2.2.29705.11362

3 authors:

Jitender Singh
Indian Institute of Technology Ropar
10 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Dwarikanath Mahapatra
IBM Research
143 PUBLICATIONS **3,229** CITATIONS
SEE PROFILE

Deepti R. Bathula
Indian Institute of Technology Ropar
15 PUBLICATIONS **31** CITATIONS
SEE PROFILE

MEDICAL VQA: MIXUP HELPS KEEP IT SIMPLE

Jitender Singh
Department of Computer Science and Engineering
Indian Institute of Technology Ropar
Rupnagar, Punjab, India, 140001
jitender.virk@iitrpr.ac.in

Dwarikanath Mahapatra
Inception Institute of Artificial Intelligence
Abu Dhabi, UAE
dwarikanath.mahapatra@inceptioniai.org

Deepti R. Bathula
Department of Computer Science and Engineering
Indian Institute of Technology Ropar
Rupnagar, Punjab, India, 140001
bathula@iitrpr.ac.in

March 16, 2022

ABSTRACT

Recently, Medical VQA became an active area of research with the induction of several publicly available benchmark datasets and the organization of challenges. Like many competitions, the quest for success has driven the use of increasingly complex neural networks. Winning strategies generally leverage multi-scale architectures and model ensembling to achieve state-of-the-art performance. However, several studies have established the capability of simpler architectures in learning more meaningful features and avoiding over parameterization. Specifically, the use of MixUp based image augmentation with a simple VGG16 network helped achieve significant improvement in performance for medical VQA. Inspired by this finding, we propose a modified version VQAMixUp that leverages both images and questions for augmenting VQA datasets. VQAMixUp combined with a few enhanced training strategies help simple models achieve state-of-the-performance on benchmark ImageCLEF-VQA-MED validation datasets.

1 Introduction

Recently, Medical VQA became an active area of research with the induction of several publicly available benchmark datasets and the organization of challenges. A sophisticated Medical VQA system can provide invaluable assistance to overburdened and under-resourced healthcare systems worldwide. However, training a reliable VQA model for the medical domain is difficult due to limited annotated data and domain-specific characteristics. In 2018, ImageCLEF [5] released a radiological dataset (VQA-MED) and organized the first VQA grand challenge in medicine. With increasing interest in utilizing medical artificial intelligence for improved patient care, several editions of ImageCLEF-VQA-MED along with other datasets have been introduced. Concurrently, several researchers employed state-of-the-art machine learning algorithms to address this challenge.

For a better understanding of the inherent challenges, we investigated the approaches adopted by some of the top-ranking participants on the leaderboards. For ImageCLEF-MED-VQA 2020, six of the top ten teams published their methods. While the top-ranking team leveraged an ensemble of multiple network architectures, most others used either VGG16 or variants of ResNet as baseline architectures for extracting image features. Three of these teams considered the VQA task as a multiclass image classification problem without modelling the QA strings. The other groups used either BioBERT or GRU to extract QA related features. With ImageCLEF-MED-VQA 2021, seven of the top ten teams published their techniques. Continuing the trend from the previous year's challenge, three teams simplified the VQA task into image classification without any text models. The remaining groups used either BioBERT or LSTMs for QA feature extraction. While four participants used VGG16 as a backbone, others used ResNet, ResNest, DenseNet121 or

Bilateral-Branch Networks (BBNs) for image-based features. Most of the high-ranking participants used multi-scale architectures and model ensemble to achieve their best performance.

Models used for VQA tasks have also grown increasingly complex to achieve state-of-the-art performance. Specifically for Medical VQA that suffers from low sample space, training such large models is very difficult. Consequently, several recent studies have questioned the use of such complex architectures and established the effectiveness of simpler models when trained systematically. For instance, SYSU-HCP [3] used a simple VGG16 network with MixUp [1] based data augmentation and some simple training mechanisms to achieve one of the best performances in ImageCLEF-VQA-MED-2021. However, as *MixUp* can only be used with images, they posed the VQA task as a simple image classification task and omitted questions information while treating answers as target classes.

Inspired by these findings and observations, this work attempts to leverage multi-modal information and simple training mechanisms that enable simple models to achieve competitive performance. The main contributions of this work are as follows: Firstly, we propose *VQAMixUp* – a data augmentation technique that extends the conventional *MixUp* to the VQA task. It uses text-based questions along with images to leverage multimodal information for augmenting VQA datasets. Additionally, we explore the effectiveness of several training schemes and their combinations. Experimental results from two medical VQA datasets establish the efficacy of *VQAMixUp* and training mechanisms in boosting the performance of a simple model to match or exceed that of complex model ensembles. Code is available at *https://github.com/VirkSaab/VQAMixUp*.

Table 1: ImageCLEF-VQA-MED datasets details. * Combination of 2020's training and validation datasets.

Year	Set Type	*Images*	*Questions*	*Answers*
2019	Train	3200	247	1552
2019	Val	500	186	470
2019	Test	500	138	166
2020	Train	4000	38	332
2020	Val	500	26	232
2020	Test	500	40	NA
2021	Train*	4500	40	332
2021	Val	500	16	236
2021	Test	500	24	NA

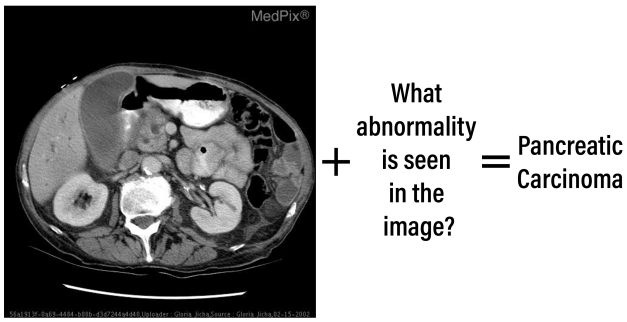

Figure 1: VQA Sample – CT Image with corresponding free-form question and answer pair.

2 Datasets

We used the ImageCLEF-VQA-MED datasets for our work. Specifically, 2020 and 2021 versions of the challenge that focused on abnormality related questions. They contain MRI, X-Ray, CT, Angiogram, Mammogram, Ultrasound, and PET scans. The datasets are highly imbalanced on both image modalities and ground truths. Figure 1 depicts a VQA sample from the ImageCLEF dataset.

Following the official suggestion and for consistency with other methods, a subset of VQA-Med 2019 related to abnormality is used to extend the VQA-Med 2020 training dataset. Hence, the training set contains a total of 6583 images with 57 unique questions and 332 unique answers. Furthermore, for a fair comparison, we used the SYSU-HCP team's publicly available curated training dataset for 2021 experiments. In addition to the VQA-MED 2021 training set,

Table 2: Results from ImageCLEF-VQA-MED 2020 Validation Set: (✓) indicates the techniques used in a particular method. Performance is reported as best accuracy followed by mean (μ) and standard deviation (σ) across 4 runs. ($\star$) indicates that μ and σ were not reported.

Method	IM	TM	QG	SAN	ENS	BBN	SSM	KA	LS	SL	MP	VMX	Acc ($\mu \pm \sigma$) (%)
Our	VGG16	GRU		✓									53.0 (52.55±0.35)
Our	VGG16	GRU	✓										55.2 (54.20±0.78)
AIML[8]	3xR50, R101, R152	-			5		✓						55.2*
Our	VGG16	GRU	✓						✓	✓	✓	✓	56.2 (55.40±0.51)
Our	VGG16	GRU		✓					✓		✓	✓	56.6 (55.65±0.59)
Our	VGG16	GRU	✓						✓			✓	57.2 (57.10±0.10)
HCP-MIC[9]	RS50	BB				✓		✓					57.2*
Our	VGG16	GRU	✓	✓					✓			✓	58.4 (57.85±0.43)
Our	VGG16	GRU							✓			✓	58.6 (58.35±0.17)
Our	VGG16	GRU											59.6 (58.80±0.58)
AIML[8]	3xR, 2xRX, 2xD, M	-			8		✓						59.6*
Our	VGG16	GRU							✓		✓	✓	**60.4 (60.05±0.38)**

ACRONYMS: IM – Image Model, TM – Text Model, QG – Question Generation, SAN - Stacked Attention Network, ENS – Ensemble, BBN – Bilateral Branch Network , SSM – Skeleton-based Sentence Mapping, KA – KL Divergence with Answer Selection, LS – Label Smoothing, SL – Super-Loss, MP – Mixed Pooling, VMX – VQAMixUp

it includes the abnormality subset of the VQA-MED 2019 dataset and the test set of the VQA-MED 2020 dataset. The yes-no type QA samples are removed from the training set as they are not part of validation or test sets. This generates a training set with 5683 images, 24 unique questions and 330 unique answers. Dataset details from different editions are given in Table 1.

3 Methodology

To develop a simple yet powerful VQA model, we evaluated top-performing models to help select the best baseline models. Consequently, we chose VGG16 (pre-trained on ImageNet) as the visual/image model (IM) and a single layer GRU with word embedding as the question/text model (TM). Furthermore, based on preliminary experiments, simple multiplication based fusion is utilized to combine the image and text features. Subsequently, we explored the effectiveness of several strategies to boost the performance of our baseline model. These strategies include:

MixUp (MX) – Generates new image V_{mix} with ground truth labels A_{mix} as a weighted combination of original images V_x and V_y and their corresponding ground truth labels A_x and A_y using the following equations:

$$V_{mix} = \lambda V_x + (1 - \lambda) V_y \tag{1}$$

$$A_{mix} = \lambda A_x + (1 - \lambda) A_y \tag{2}$$

where the value of mixing parameter $\lambda \in [0, 1]$ is chosen randomly from a $Beta$ distribution with α as the hyper-parameter. While Eqn. 2 works for one-hot encoded labels, a Loss *MixUp* is used for other types of targets as shown below:

$$loss = \lambda \mathcal{L}(\hat{P}_x, A_x) + (1 - \lambda) \mathcal{L}(\hat{P}_y, A_y) \tag{3}$$

where $\mathcal{L}$ can be any loss function that takes predictions ($\hat{P}_x$ and $\hat{P}_y$) and target values or ground truth labels (A_x and A_y).

VQAMixUp (VMX) – We propose a modified version of MX data augmentation technique for the VQA task by leveraging both images and questions. The technique is described in Algorithm 1. Similar to conventional MX, a new image sample ($\hat{V}$) is generated as a weighted combination of sample images as shown in Step 3. However, generating a combination of two questions is not as straightforward as questions are represented as vectors of tokens. Hence, we avoid generating a mixed question and use LossMixUp to generate the target answer for the mixed image. To this end, the influence of questions is accounted for by generating two target answer predictions for the new mixed image linked with two sample questions. This mechanism is depicted in Steps 4 and 5, where the VQA model ($\mathcal{M}$) is used to generate predictions $\hat{P}_x$ and $\hat{P}_y$ for a new image ($\hat{V}$) corresponding to questions Q_x and Q_y, respectively. These

individual QA specific predictions are further combined using the mixing parameter (λ) to generate the final prediction as shown in Step 6. Finally, LossMixUp is used to calculate the task-specific loss using predicted and ground-truth answers as in Step 7.

Algorithm 1: *VQAMixUp* algorithm.

Data: VQA samples – (V_x, Q_x, A_x), (V_y, Q_y, A_y)
Model: VQA model – $\mathcal{M}$
Loss: Loss function ($\mathcal{L}$) - cross-entropy in our case.
Function `VQAMixup`(V_x,Q_x,A_x,V_y,Q_y,A_y,$\mathcal{M}$,α)**:**
  $\lambda \leftarrow Beta(\alpha, \alpha)$ where $\alpha \in (0, \infty)$
  $\hat{V} \leftarrow \lambda V_x + (1 - \lambda)V_y$ `/*mixed image*/`
  $\hat{P}_x \leftarrow \mathcal{M}(\hat{V}, Q_x)$
  $\hat{P}_y \leftarrow \mathcal{M}(\hat{V}, Q_y)$
  $\hat{P}_{vqa} \leftarrow \lambda \hat{P}_x + (1 - \lambda)\hat{P}_y$ `/*predictions*/`
  $loss \leftarrow \lambda \mathcal{L}(\hat{P}_x, A_x) + (1 - \lambda)\mathcal{L}(\hat{P}_y, A_y)$
  return $loss$, $\hat{P}_{vqa}$

Table 3: Results from ImageCLEF-VQA-MED 2021 Validation Set: (✓) indicates the techniques used in a particular method. Performance is reported as best accuracy followed by mean (μ) and standard deviation (σ) across 4 runs. ($\star$) indicates that μ and σ were not reported.

Method	IM	TM	Q	BBN	GAP	SAN	LS	SL	QG	MP	MX	VMX	Acc ($\mu \pm \sigma$) (%)
TeamS[7]	RS	-		✓									61.3*
Our	VGG16	GRU	✓			✓							63.2 (63.00±0.14)
Our	VGG16	GRU	✓			✓	✓		✓	✓		✓	63.8 (63.34±0.45)
Our	VGG16	GRU	✓			✓						✓	65.2 (64.95±0.32)
Our	VGG16	GRU	✓										66.2 (65.85±0.29)
SYSU-HCP[3]	VGG16	-											66.6 (66.25±0.60)
Our	VGG16	GRU	✓									✓	68.2 (68.00±0.14)
Our	VGG16	GRU	✓				✓			✓		✓	68.4 (67.95±0.29)
Our	VGG16	GRU	✓				✓	✓		✓		✓	68.6 (67.95±0.43)
SYSU-HCP[3]	VGG16	-			✓		✓	✓			✓		69.2 (68.35±0.85)
Our	VGG16	GRU					✓	✓				✓	69.4 (68.35±0.60)
Our	VGG16	GRU	✓			✓	✓		✓			✓	69.4 (69.10±0.22)
Our	VGG16	GRU	✓				✓					✓	**70.2 (69.70±0.30)**

ACRONYMS: IM – Image Model, TM – Text Model, Q - Used questions, BBN – Bilateral Branch Network, GAP – Global Average Pooling, SAN – Stacked Attention Network, LS – Label Smoothing, SL – SuperLoss, QG – Question Generation, MP – Mixed Pooling, MX – MixUp, VMX – VQAMixUp

MixPool (MP) – Uses a weighted combination of average and max pooling, as shown below, to boost invariance to data transformations and provide rich latent features

$$MixPool(x) = \frac{1}{\theta}\, AvgPool(x) + \frac{1}{\phi}\, MaxPool(x) \tag{4}$$

where θ and ϕ are trainable parameters initialized to 2.

Question Generation (QG) – Used for manually augmenting the dataset with slightly different versions of original questions associated with images, keeping the answers constant. For example, given an original question *What abnormality is seen in the image?*, sample generated questions include *What is most alarming about this image?*, *What is the primary abnormality in this image?*, etc.

SuperLoss (SL) – A curriculum learning approach where a model is presented with easy samples before difficult ones during training. It is used in conjunction with an existing task based loss function. We use SL with cross-entropy loss.

Label Smoothing (LS) – A regularization method that introduces noise (ϵ) to the ground truth values to avoid direct likelihood maximization of $log\ p(y|x)$ assuming there are mistakes in the data labels.

Stacked Attention Network (SAN) – A multi-step reasoning method for natural language VQA task. Multilayer attention helps SAN to pinpoint the salient regions and filter out noisy features.

4 Experimental Results

As ImageCLEF-VQA-MED challenges for 2020 and 2021 are no longer accepting submissions, performance comparison on test sets is not possible. However, as several top-ranking teams have published their results on validation sets, we use them for evaluation. For 2020, teams AIML [8] and HCP-MIC[9] ranked 1st and 4th respectively. Similarly, for 2021, SYSU-HCP[3] ranked 1st and TeamS[7] was placed 3rd. Some of the features used by these winning teams include multi-scale networks, multi-model ensembles, Skeleton-based Sentence Mapping (SSM) for extracting relevant information from questions, Bilateral Branch Network (BBN), Hierarchical Adaptive Global Average Pooling (GAP), MixUp data augmentation, Label smoothing (LS) and SuperLoss curriculum learning (SL). It was not practically feasible to try all possible combinations of strategies due to limited computational resources. Nevertheless, we tried numerous combinations and report the most significant ones due to space limitations.

Model Training: All models were trained using the PyTorch framework with cross-entropy loss and SGD optimizer, 0.001 learning rate, 0.0005 weight decay, 0.9 momentum, 32 batch size, and 60 epochs. Step-wise learning rate decay was used with a factor of 0.6 for every 20 epochs. Both α and ϵ were set to 0.1 for *VQAMixUp* and Label Smoothing, respectively. SL hyper-parameters are the same as SYSU-HCP. The SAN fusion method uses single layer attention because of the simplicity of questions. QG generates an average of 7 questions per sample.

Results: We compare the performance of our proposed approach with these winning groups on 2020 and 2021 validation sets. These results are shown in Tables 2 and 3. Specific strategies employed by each method is indicated with a (✓) in the corresponding column. For ensembles, the number of models used is specified. It can be observed that our simple model can achieve state-of-the-art performance with the proposed VQAMixUp and a few simple strategies. Improvement of our model is statistically significant as $p < 0.02$ and $p < 0.05$ compared to second-best methods, SYSU-HCP (2021) and AIML (2020) respectively, on validation sets.

5 Conclusion

In the context of Medical VQA, we demonstrate the effectiveness of simple training mechanisms in boosting the performance of a simple model to match or exceed that of large and complex models like ensembles. Remarkably, a simple VGG16 (for images) combined with a single layer GRU (for text) aided by our proposed VQAMixUp and generic training methods achieved state-of-the-art performance on benchmark ImageCLEF-VQA-MED validation sets. In future, a systematic and comprehensive ablation study that determines the relative contribution from each training strategy can further help establish their efficacy for VQA in general.

References

[1] Zhang, H., Cisse, M., Dauphin, Y., N., Lopez-Paz, D., "mixup: Beyond Empirical Risk Minimization," CoRR. abs/1710.09412, 2018.

[2] Castells, T., Weinzaepfel, P., Revaud, J., "SuperLoss: A Generic Loss for Robust Curriculum Learning," Curran Associates, Inc., 33, pp. 4308-4319, 2020.

[3] Gong, H., Huang, R., Chen, G., and Li, G., "SYSU-HCP at VQA-Med 2021: A Data-centric Model with Efficient Training Methodology for Medical Visual Question Answering," CLEF, 2021.

[4] Virk, J.S., Bathula, D.R., "Domain-Specific, Semi-Supervised Transfer Learning for Medical Imaging, " CODS COMAD 2021, pp 145-153, 2021.

[5] ImageCLEF, https://www.imageclef.org, last accessed 23 Feb, 2022.

[6] Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J., "Stacked Attention Networks for Image Question Answering," CoRR, abs/1511.02274, 2015.

[7] Eslami, S., de Melo, G., Meinel, C., "TeamS at VQA-Med 2021: BBN-Orchestra for long-tailed medical visual question answering," CEUR, vol. 2936, pp. 1211-1217, 2021.

[8] Liao, Z., Wu, Q., Shen, C., Hengel, A.V., Verjans, J.W., "AIML at VQA-Med 2020: Knowledge Inference via a Skeleton-based Sentence Mapping Approach for Medical Domain Visual Question Answering," CEUR, vol. 2696, pp. 1-14, 2020.

[9] Chen, G., Gong, H., Li, G., "HCP-MIC at VQA-Med 2020: Effective Visual Representation for Medical Visual Question Answering," CEUR, vol. 2696, 2020.