IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 20, 2020, accepted November 24, 2020, date of publication December 1, 2020,
date of current version December 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3041503

Fair-VQA: Fairness-Aware Visual Question Answering Through Sensitive Attribute Prediction

SUNGHO PARK, SUNHEE HWANG, (Graduate Student Member, IEEE), JONGKWANG HONG, (Member, IEEE), AND HYERAN BYUN, (Member, IEEE)
Department of Computer Science, Yonsei University, Seoul 03722, Republic of Korea

Corresponding author: Hyeran Byun (hrbyun@yonsei.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) under Grant 2019R1A2C2003760, and in part by the Institute for Information & Communications Technology Planning & Evaluation (IITP) Grant funded by the Korean Government (MSIT) (Development of framework for analyzing, detecting, mitigating of bias in AI model, and training data) (Artificial Intelligence Graduate School Program (Yonsei University)) under Grant 2019-0-01396 and Grant 2020-0-01361.

ABSTRACT Visual Question Answering (VQA) is a task that answers questions on given images. Although previous works achieve a great improvement in VQA performance, they do not consider the fairness of answers in terms of ethically sensitive attributes, such as gender. Therefore, we propose a Fair-VQA model that contains two modules: VQA module and SAP (Sensitive Attribute Prediction) module. On top of VQA module, which predicts various kinds of answers, SAP module predicts only sensitive attributes using the same inputs. The predictions of SAP module are utilized to rectify answers from VQA module to be fairer in terms of the sensitive attributes with graceful performance degradation. To validate the proposed method, we conduct extensive experiments on VQA, GQA, and our proposing VQA-Gender datasets. In all the experiments, our method shows the fairest results in various metrics for fairness. Moreover, we demonstrate that our method works interpretably through the analysis of visualized attention maps.

INDEX TERMS FAI, fairness, visual question answering, VQA.

I. INTRODUCTION

In recent years, Artificial Intelligence (AI) systems provide our society with an efficient and accurate decision-making process, but it often causes unfair decisions that give benefits to particular demographic groups, *i.e.,* ethnicity or gender. For example, COMPAS, which is an AI system exploited by courts in the United States, predicts that African-American are more likely to commit recidivism than Caucasians. In addition, the AI recruitment system used in Amazon gives higher acceptance scores for male applicants than female applicants [1]. These social issues raise criticism of existing AI systems and the need for fairness-aware AI systems in terms of sensitive attributes, such as gender, age, and ethnicity [2]–[6]. As a result, fairness studies are performed in variety of tasks, from a simple classification on structured data [7]–[9] to Natural Language Processing (NLP) [2], face recognition [3], and Generative Adversarial Network (GAN) [4].

Nonetheless, there are still some research fields where fairness on sensitive attributes is not addressed much yet. Particularly, in this paper, we pay attention to Visual Question Answering (VQA), which is a task to answer to a question by utilizing an relevant image [10]–[15]. Although it is one of the most popular research topics in computer vision and most VQA benchmark datasets [10], [13], [14] contain questions related to sensitive attributes, there are a few approaches to predict fair answers to the sensitive questions. In our experiments on VQA 2.0 dataset [13], a baseline VQA model [15] shows better performance in male data than female data. It demonstrates that there is a male-biased problem in the existing VQA model.

There are several previous works [10], [16], [17] which try to overcome the general bias problem in VQA tasks, but they are inefficient to mitigate unfairness in terms of sensitive attributes. Since sensitive questions are a small part of the whole question, the prior approaches, which removes the general bias in all kinds of questions, cause a large trade-off with accuracy.

The associate editor coordinating the review of this manuscript and approving it for publication was Seok-Bum Ko.

Therefore, we propose a Fair-VQA model to predict fair answers to sensitive questions while maintaining the overall performance. Our model consists of two modules: VQA module and Sensitive Attribute Prediction (SAP) module. Given input images and corresponding questions, VQA module predicts various kinds of but unfair answers, while SAP module focuses on predicting sensitive attributes of answers (*e.g.*, male or female). The sensitive attributes from SAP module are combined with the answers from VQA module if the answers are related to sensitive attributes. If not, SAP module is not utilized. This de-biasing fusion scheme encourages answers to be fairer in terms of sensitive attributes and prevents performance degradation in answers irrelevant to sensitive attributes.

To validate our method, we compare it with baselines [15] on two benchmark datasets, VQA 2.0 [13] and GQA [14]. Moreover, we construct a new evaluation dataset, VQA-Gender, to measure the degree of fairness in terms of gender. In all the datasets, our method achieves a great improvement in fairness scores measured by various metrics with graceful performance degradation. Besides, through qualitative analysis on visualized attention maps, we demonstrate the prediction of sensitive attributes is based on the right evidence in input images.

The Main Contributions: We summarize our main contributions as follows:

- We propose a Fair-VQA model that predicts fair answers to sensitive questions while maintaining the overall performance.
- We introduce a new VQA dataset, which is VQA-Gender, to evaluate the degree of fairness in terms of gender.
- Through extensive experiments, we validate that our method achieves a significant improvement in fairness with graceful performance degradation.

II. RELATED WORKS

A. FAIRNESS IN COMPUTER VISION

Several previous works [2], [3], [18]–[20] indicate that there are unfairness issues in existing image datasets. [3] shows that most benchmark datasets for face recognition are biased to Caucasians than other ethnicities and proposes Pilot Parliament Benchmark (PPB) dataset that is balanced for skin color and gender [21], [22]. Furthermore, [18]–[20] claims sensitive attributes, such as gender, age, and ethnicity, are correlated with specific facial attributes in CelebA and UTK Face datasets [23], [24]. Similarly, [25] finds out that some gender-stereotyped objects are frequently co-occurred with a specific gender (*i.e.*,male or female) in MS-COCO dataset [26].

In addition, [25], [27]–[30] try to solve the problem in various computer vision tasks. [27] shows that existing image captioning models generate unfair captions in terms of gender and proposes a fairness-aware image captioning model. By introducing a loss function that encourages the model to decide the gender of words using appropriate visual evidence, it performs fairer image captioning in terms of gender. In addition, [28] solves the problem that sensitive attributes are unintentionally translated in image-to-image translation tasks. To this end, they disentangle representation into two spaces related to sensitive attributes or target attributes. Likewise, [29], [30] utilize disentangled representation to mitigate unfairness in face recognition and face attribute classification tasks, respectively. Reference [29] separates data representation into gender, age, ethnicity, and identity spaces, which performs invariant face recognition with respect to sensitive attributes. Reference [30] disentangles data representation in terms of multiple sensitive attributes without target attribute labels. They perform various downstream classification tasks fairly by excluding spaces with corresponding sensitive attributes. Besides, [25] proposes a method that removes sensitive attribute information in the intermediate representation of CNN with adversarial training to perform fair object recognition tasks.

B. VISUAL QUESTION ANSWERING

Encoding informative representation from questions and images is important to improve Visual Question Answering (VQA) performances [15], [32]. In the early studies, input images are encoded with VGG [33] or ResNet [34] networks and visual attention is applied to the encoded visual features in the form of grid [15], [32]. However, the grid-formed attention has the disadvantage of not accurately reflecting the boundary of objects. To solve this problem, [15] uses Faster R-CNN [35] network to obtain image representation at the level of objects and proposes an object-based visual attention mechanism. On the other hand, for text encoding, many studies [15], [32], [36] first encode questions in word-level using GloVe [37] and re-encode them in sentence-level using Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) [38] networks.

On top of that, [32], [36], [39]–[42] propose methods to combine multimodal data efficiently. In the early studies, text and visual features are combined by simple concatenation, element-wise multiplication, and outer product. However, the concatenation and element-wise multiplication can not reflect sufficient relation between textual and visual features, and the outer product has a huge computational cost [39]. Therefore, [39]–[41] propose multimodal fusion methods that sufficiently reflect relation between the two features with an efficient computational cost. Moreover, [36] and [32], [42] propose methods that calculate relation between all words and visual objects utilizing bilinear attention and self-attention mechanism, respectively.

From a different perspective, [10], [16], [17] point out that the distribution of answers per question is extremely biased in most VQA datasets, which leads VQA models to find an answer by simply memorizing relation between questions and answers. They call this phenomenon language bias or unimodal bias and try to mitigate this problem. [10] constructs VQA-CP dataset by reorganizing the distribution of answers per question in VQA dataset. Since the distribution

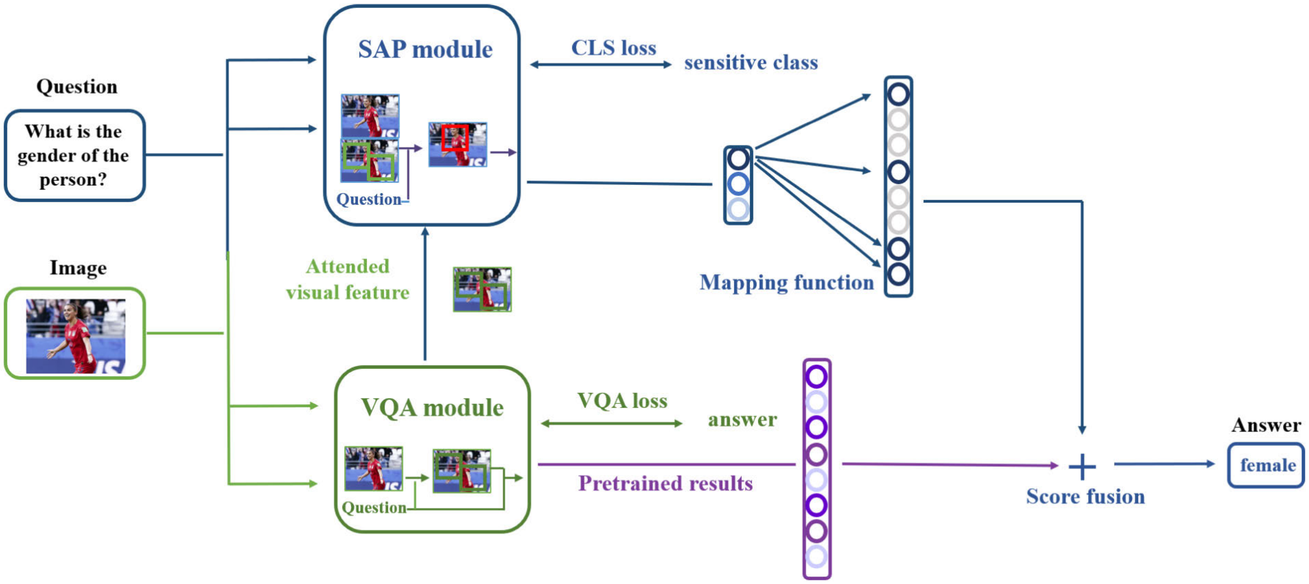


FIGURE 1. The overall architecture of our Fair-VQA, which consists of SAP module and VQA module. The outputs from SAP and VQA modules are combined into the final answer through the de-biasing fusion scheme.

is different in training and test sets, models over-fitted to the biased distribution show low performance in this dataset. In addition, they propose a VQA model divided into two sub-networks to avoid finding answers using only questions. One network extracts visual evidence from input images regardless of question types and the other network selects only the evidence related to the question types. Furthermore, [16] introduces a QA model sharing a question encoder with a VQA model. The question encoder is trained not to predict answers within QA model but to predict answers well within VQA model. It encourages the model to exploit visual information to find answers.

III. TERMINOLOGY AND DATA PRE-PROCESSING

Most VQA datasets [10], [13], [14] have questions on sensitive attributes, such as gender, age, or ethnicity. We define these questions as 'sensitive questions' and answers to the sensitive questions as 'sensitive answers'. For example, 'what gender is the basketball players?' is one of the sensitive questions on gender, and words such as 'woman', 'man', and 'female' can be sensitive answers to the questions. Although there are various types of sensitive attributes in VQA datasets, such as ethnicity and age, we only conduct experiments on gender since there is little data on ethnicity, age, or other sensitive attributes in VQA 2.0 and GQA datasets. This scarcity of data makes it quite difficult to train and evaluate VQA models.

Besides, since there is no label of sensitive attributes in the datasets, we find out sensitive questions as follows. Firstly, we cluster answer labels into 'female', 'male', and 'background' groups. For example, 'woman' and 'girl' are clustered into 'female' group and 'tennis' and 'red' are clustered into 'background' group. Then, the questions of which answers are in 'female' and 'male' groups are clustered into sensitive questions. The others are clustered into non-sensitive questions. Lastly, exceptions of 'female' and 'male' questions, which are not related to human beings, are replaced into the non-sensitive questions. Through this process, we find about 2,100 male and 1,600 female questions in VQA 2.0 dataset, and about 9,000 male and 6,000 female questions in GQA 2.0 dataset.

IV. BASELINE

We utilize a BUTD model [15] as our baseline. It encodes images and questions into visual vectors $v = [v_1 \cdots v_K]$, $v_k \subseteq \mathbb{R}^{d_v}$ using a pre-trained Faster R-CNN [35] and question vectors $q \subseteq \mathbb{R}^{d_q}$ using GRU, respectively. v and q are combined by Hadamard product [40] to calculate visual attention $Att_v \subseteq \mathbb{R}^K$. v are weighted by Att_v into attended visual features $\acute{v} \subseteq \mathbb{R}^{d_v}$, which are combined with q into multimodal features $m \subseteq \mathbb{R}^{d_m}$. Finally, m is fed into a MLP classifier to output answers. BUTD shows good performances without glimpses [39], stack structures [43], and question attention. In addition, it has low computational cost and good scalability due to its simple structure. For these reasons, we set this model as the baseline.

V. PROPOSED METHOD

A. OVERVIEW

The overall architecture of our Fair-VQA model is shown in Figure 1. It consists of two modules, SAP module and VQA module, and de-biasing fusion scheme. Instead of mitigating unfairness with respect to sensitive attributes by directly designing VQA module, we rectify unfair answers from VQA module using SAP module that predicts sensitive attributes of

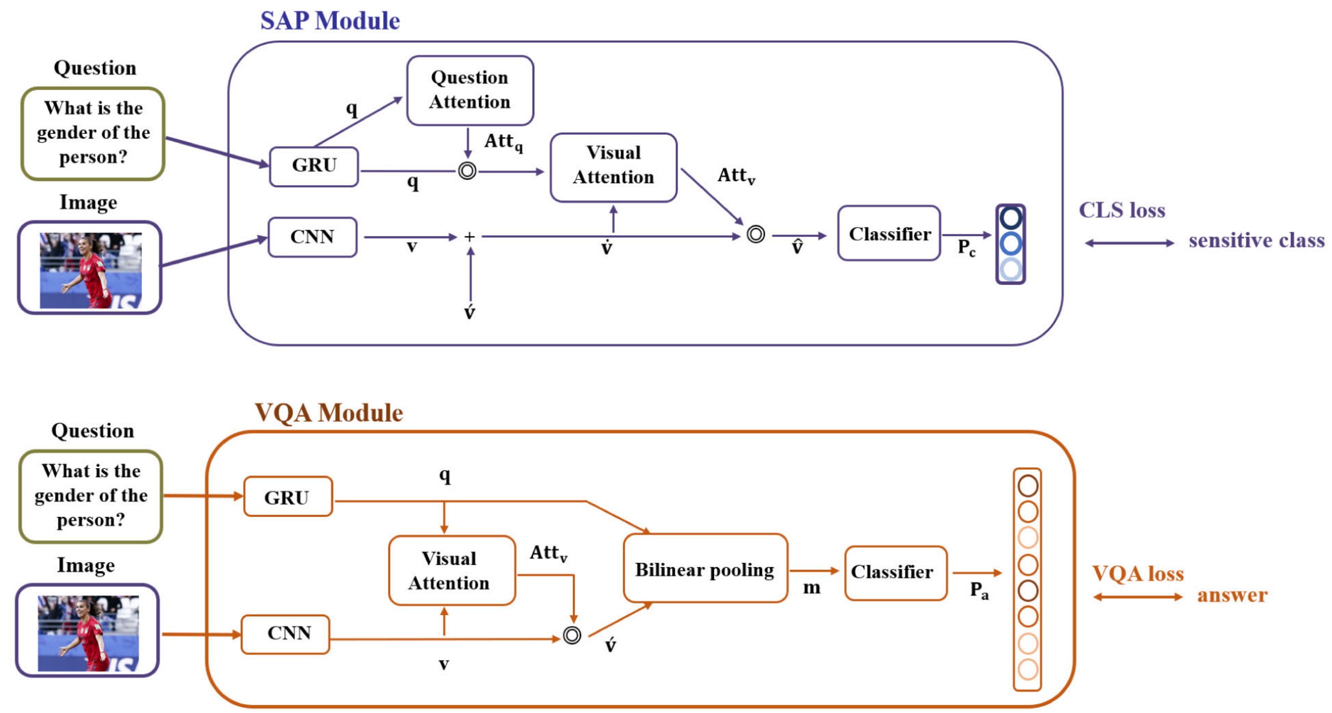


FIGURE 2. The structures of SAP module and VQA module. SAP module (top) predicts sensitive attributes and VQA module (bottom) predicts various kinds of answers.

answers. The outputs of the two modules are combined by debiasing fusion scheme into final answers in execution time.

B. SAP MODULE

The structure of SAP module is shown in Figure 2. This module receives images and questions to predict sensitive attributes $s \subseteq S$. Firstly, it encodes an image into visual vectors $v = [v_1 \cdots v_K]$, $v_k \subseteq \mathbb{R}^{d_v}$ with a pre-trained Faster R-CNN and a question into word-level question vectors $q = [q_1 \cdots q_T]$, $q_t \subseteq \mathbb{R}^{d_q}$ with GloVe word embedding and GRU. K is the number of objects in an image and T is the length of a question, which is fixed at 14 as in [32], [36], [44]. Using the encoded vectors q and v, it calculates question and image attention to weight important vectors in q and v, respectively. The calculation and applying process of question attention is as follows:

$$Att_q = softmax(W_q q), \tag{1}$$

$$\hat{q} = \sum_t Att_{q_t} q_t, \tag{2}$$

where $W_q \subseteq R^{d_q}$ is learnable parameters and $Att_q \subseteq R^n$ is visual attention. In addition, before it calculates visual attention Att_v, visual vectors v are combined with $\acute{v}$, which are attended visual features from VQA module, as follows:

$$\dot{v} = v \circledcirc \acute{v}, \tag{3}$$

$\circledcirc$ denotes element-wise product. Since SAP module is supervised by only sensitive attribute labels, $\acute{v}$ can provide plentiful visual guidance for finding more accurate visual attention. The visual attention is calculated as follows:

$$Att_v = softmax(W_h(\mathrm{ReLU}(W_{\hat{q}}\hat{q}) \circledcirc \mathrm{ReLU}(W_{\dot{v}}\dot{v}))), \tag{4}$$

where $W_{\dot{v}} \subseteq R^{h,d_v}$, $W_{\hat{q}} \subseteq R^{h,d_q}$, and $W_h \subseteq R^h$ are learnable parameters. Lastly, visual vectors are weighted summed by the visual attention Att_v and fed into classifier to predict sensitive attributes s^* as follows:

$$\hat{v} = \sum_i Att_{v_k} v_k, \tag{5}$$

$$P_c = softmax(W_c(\mathrm{ReLU}(W_{\hat{v}}(\hat{v})))), \tag{6}$$

$$s^* = argmax(P_c), \tag{7}$$

where $W_{\dot{v}} \subseteq R^{h,d_v}$ and $W_c \subseteq R^{c,h}$ are learnable parameters. P_c is the softmax score for sensitive attributes.

C. VQA MODULE

VQA module has the same structure as the baseline. In the training step, it is pre-trained on its own and jointly retrained with SAP module while transferring the attended visual features $\acute{v}$ into SAP module. Since the retraining process is only to transfer $\acute{v}$ into SAP module, we utilize the pre-trained VQA module in de-biasing fusion scheme.

D. DE-BIASING FUSION SCHEME

De-biasing fusion scheme is for combining the outputs of two modules fairly and with graceful performance degradation. When $VQA_{label} = [l_1, \cdots l_N]$ is a list of VQA answers and $VQA_{score} = [va_1, \cdots va_N]$ are output scores of VQA module, we map predicted sensitive attributes $s*$ into relevant VQA_{label}. The mapping function f is defined as:

$$\underset{l_n, s^* \rightarrow \{0,1\}}{f} : f(l_n, s^*) = \begin{cases} 1, & l_n \in s^* \\ 0, & l_n \notin s^*. \end{cases} \tag{8}$$

Using the mapping function f, the predicted score of $s*$ ($Max(P_c)$) is copied into corresponding VQA answers as follows:

$$SAP_{score} = Max(P_c) * M_{label}, \tag{9}$$

where M_{label} is the outputs of mapping function f for VQA_{label}. Lastly, we combine VQA_{score} with SAP_{score} into $Total_{score}$ as follow:

$$Total_{score} = \begin{cases} VQA_{score} + \alpha SAP_{score}, & s^* \neq \text{background} \\ VQA_{score}, & s^* = \text{background}, \end{cases} \tag{10}$$

where α is a hyperparameter which adjusts the mixing ratio of the outputs of SAP module and VQA module. Since SAP_{score} is not combined with VQA_{score} if s^* is background group, it causes little performance degradation for questions irrelevant to sensitive attributes. Finally, we obtain the final answer as follows:

$$Answer = argmax(Total_{score}). \tag{11}$$

VI. DATASET

A. BENCHMARK VQA DATASETS

VQA 2.0 dataset is composed of images from MS-COCO dataset [26] and question-answer pairs for the images. It is modified from previous VQA 1.0 [45] to balance answer distributions to each question. 3 questions are assigned per image, and answers to each question are annotated by 10 different people. We use only train and validation sets in our experiment because we cannot access to answers of test set.

GQA dataset mitigates unbalanced answer distributions in previous datasets and provides semantic information of both images and questions. The images are associated with scene graphs of their objects, attributes, and relations, and the questions are associated with the structured representation of its semantic information. In our work, we only leverage data whose answers are related to gender. As a result, we obtain about 9,000 data for men and about 6,000 data for women. We use this dataset only for evaluation.

B. VQA-GENDER DATASET

We construct VQA-Gender dataset for a more reliable evaluation of fairness in terms of gender since VQA 2.0 dataset includes only a small amount of gender data in validation set, about 500 for female and 700 for male. To gather images related to gender in VQA dataset, we first pick out the images of which questions include words about gender. Then, we remove images containing both words for female and male. In this way, we could obtain 19,836 images for male and 8,362 images for female.

In addition, since their questions do not necessarily ask their gender, we substitute them into new questions on gender, such as 'what is the gender of the person?'. All the questions are composed of simple questions and divided into two cases. One case requires a particular word as its answer, such as 'what is the gender of the person, female or male?', and the other does not have this limitation, such as 'what is the gender of the person'. In the former, a specific word included in the question is assigned to the answer (female or male in this example). Meanwhile, in the latter, the answer is randomly selected in the list of female or male answers. The details of the list are described in the implementation detail section.

VII. EXPERIMENTS

A. IMPLEMENTATION DETAIL

The dimensions of visual features d_v, question features d_q, and hidden features h are 2,048, 1,280, and 1,280, respectively. We use 10∼100 visual features per image depending on the threshold for selection of salient image regions and set the length of questions to 14 by either filling empty spaces with zero-padding or truncating extra words as in [15]. The number of possible answers is fixed to 3,129, which are the most popular answers. Since the ratio of female, male, and background groups is very unbalanced (about 2:3:1,000), we train SAP module using a balanced cross-entropy loss [46]. Specifically, the weights for this loss are set to 1.0, 0.5, and 0.0001 in female, male, and background groups, which are roughly decided by considering the data ratio. Besides, we train all the VQA networks in our experiment with a binary cross-entropy loss. The learning rate of VQA module follows the learning rate scheduler utilized in [36]. The learning rate of SAP module is 10^{-1} for the first 6 epochs and is decayed by 1/100 for the rest epochs. All results in our tables are obtained from fully converged models with the same batch size of 512 and random seeds. We use an Adamax optimizer [47] with a gradient clipping of 1/4. We set the lists of female and male answers to ['woman', 'women', 'female', 'girl', 'lady', 'girls'] and ['man', 'men', 'male', 'boy', 'boys'], respectively.

B. COMPARABLE MODELS

To better validate our contributions, we introduce the following comparable models.

Baseline + VQclassifier: On top of the baseline, we add a classifier (VQclassifier), which predicts sensitive attributes using multimodal features m. Unlike our method, the output scores of the baseline and VQclassifier are not combined. Instead, the multimodal features are learned to include more information about sensitive attributes through VQclassifier.

Baseline + Vclassifier: Similar to the above, this model has an additional classifier (Vclassifier). The difference is that Vclassifier predicts sensitive attributes using attended visual

TABLE 1. *VQA score* on VQA 2.0, GQA, and VQA-Gender datasets. Diff_I denotes the difference of VQA score between female and male answers. For all the datasets, Fair-VQA($\alpha = 3$) shows the fairest performances.

Dataset	Model	Overall (↑)	Female (↑)	Male (↑)	Diff_I (↓)
VQA 2.0	Baseline	64.71	56.32	66.23	9.91
	Baseline+Vclassifier	64.73	58.30	66.96	8.66
	Baseline+VQclassifier	**64.74**	59.88	67.10	7.22
VQA 2.0	Fair-VQA(α=1)	64.71	61.46	68.27	6.81
	Fair-VQA(α=2)	64.67	72.92	76.17	3.25
	Fair-VQA(α=3)	64.60	**73.52**	**76.75**	**3.23**
GQA	Baseline	-	16.24	67.25	51.01
	Baseline+Vclassifier	-	18.89	**70.07**	51.18
	Baseline+VQclassifier	-	21.06	65.91	44.85
GQA	Fair-VQA(α=3)	-	**31.79**	69.23	**37.44**
VQA-Gender	Baseline	-	76.25	78.72	2.47
	Baseline+Vclassifier	-	75.50	**79.67**	4.17
	Baseline+VQclassifier	-	76.14	79.59	3.45
VQA-Gender	Fair-VQA(α=3)	-	**77.43**	78.64	**1.21**

features $\acute{v}$. It encourages visual attention to weight visual features in consideration of sensitive attributes.

C. QUANTITATIVE ANALYSIS IN INSTANCE-LEVEL

We compare our model with the baseline and comparable models on VQA, GQA, and VQA-Gender datasets [13], [14]. In Table 1, we report *VQA score* (Overall), which is the most popular metric to measure VQA performances [45], and the difference of *VQA score* between female and male answers (Diff_I). The lower difference of *VQA score* indicates that the results are fairer in terms of the sensitive attribute, gender. In benchmark datasets, VQA 2.0 and GQA, the baseline shows highly unfair results in terms of gender, achieving Diff_I of 9.91% and 51.0%. In addition, Baseline + Vclassifier has little effect on the improvement of fairness in both datasets. Although Baseline + VQclassifier further improves fairness, it is still not significant. Our method ($\alpha = 3$) achieves the fairest results on all the benchmark datasets, overperforming the baseline by 6.68% and 13.57%, respectively. These results indicate our two-branch approach is more effective than comparable models which learn sensitive attribute information directly in VQA networks. Even though the overall performance of ours is degraded on VQA 2.0 dataset, it is very slight (0.11%).

Compared to the benchmark datasets, all the models show lower Diff_I scores on VQA-Gender dataset since it is only composed of simple questions on gender. Our method also records the fairest Diff_I of 1.21% on this dataset.

D. QUANTITATIVE ANALYSIS ON GROUP-LEVEL

We point out that the existing *VQA score* metric has a limitation in evaluating the fairness of results. For example, if a VQA network predicts an answer as 'woman' but the ground truth (GT) is 'female', the existing metric judges this prediction is wrong. However, in terms of fairness, this answer should be judged to be correct. Rather, it is more important to measure the proportion of answers mispredicted into the other sensitive attribute group or background group. Therefore, we measure group-level accuracy, which indicates whether the prediction is included in the same sensitive attribute group as GT, and its difference between male and female groups (Diff_G). Moreover, we measure the proportion of answers mispredicted into the other sensitive attribute group among all data: $abs_{F \to M}$ (female to male) and $abs_{M \to F}$ (male to female). $rel_{F \to M}$ and $rel_{M \to F}$ represent the proportion of answers mispredicted into male or female among their errors, respectively.

As shown in Table 2, the baseline shows the unfairest results on the benchmark datasets and both comparable models improve fairness in terms of all the metrics, Diff_G, Diff_A, and Diff_R. However, they slightly degrade the fairness scores on VQA-Gender dataset. In all the datasets, our method achieves the fairest results, particularly showing the improvement of 21.9% (Diff_G), 25.68% (Diff_A), and 64.43% (Diff_R) over the baseline on GQA dataset.

E. QUALITATIVE ANALYSIS

In this section, we qualitatively analyze visualized attention maps of the baseline and SAP module. In Figure 3, we show two examples of visualized attention maps, where brighter boxes indicate that the models pay more attention to the regions. Although both the baseline and SAP module focus on the right evidence, which are the regions related to women, the baseline outputs biased answers (female to male). Compared to this, SAP module predicts the correct sensitive attributes (female class) from the right evidence and encourages our model to predict the correct answers (female).

F. ABLATION STUDY

In Table 3, we analyze the effectiveness of transferring visual guidance ($\acute{v}$) from VQA module to SAP module. In all the datasets, visual guidance significantly improves fairness scores at the group-level. Specifically, our model with visual guidance overperforms the other by 30.28% (Diff_G), 30.42% (Diff_A), and 33.84% (Diff_R) on GQA dataset. This is because

TABLE 2. Group-level accuracy on VQA 2.0, GQA, and VQA-Gender datasets. $abs_{F\rightarrow M}$ and $abs_{M\rightarrow F}$ denote the proportion of answers mispredicted into the other sensitive attribute groups. $rel_{F\rightarrow M}$ and $rel_{M\rightarrow F}$ are their ratio to errors of female and male groups, respectively.

Dataset	Model	Female (↑)	Male (↑)	Diff_G (↓)	$abs_{F\rightarrow M}$	$abs_{M\rightarrow F}$	Diff_A (↓)	$rel_{F\rightarrow M}$	$rel_{M\rightarrow F}$	Diff_R (↓)
VQA 2.0	Baseline	63.83	70.61	6.78	21.54	10.96	10.58	59.55	37.29	22.26
	Baseline+Vclassifier	65.02	71.35	6.33	20.16	10.38	9.78	57.63	36.23	21.40
	Baseline+VQclassifier	67.19	71.05	3.86	17.79	10.38	7.41	54.22	35.85	18.37
VQA 2.0	Fair-VQA(α=1)	70.75	74.27	3.52	18.77	12.72	6.05	64.17	49.44	14.73
	Fair-VQA(α=2)	72.92	76.17	3.25	19.36	13.74	5.62	71.49	57.66	13.83
	Fair-VQA(α=3)	**73.51**	**76.75**	**3.24**	19.76	15.06	**4.70**	74.59	64.77	**9.82**
GQA	Baseline	29.35	85.90	56.55	60.67	0.24	60.43	85.87	1.70	84.17
	Baseline+Vclassifier	32.89	**89.25**	56.36	59.02	2.82	56.20	87.95	26.23	61.71
	Baseline+VQclassifier	37.36	84.24	46.88	53.91	4.26	49.65	86.06	27.03	59.03
GQA	Fair-VQA(α=3)	**54.16**	88.81	**34.65**	43.05	8.30	**34.75**	93.91	74.17	**19.74**
VQA-Gender	Baseline	80.73	82.30	1.57	15.68	13.90	1.78	81.37	78.53	2.84
	Baseline+Vclassifier	79.89	83.47	3.58	16.48	12.95	3.53	81.95	78.34	3.61
	Baseline+VQclassifier	80.54	83.40	2.86	15.71	12.93	2.78	80.73	77.89	2.84
VQA-Gender	Fair-VQA(α=3)	**83.37**	**83.99**	**0.62**	14.90	14.41	**0.49**	89.60	90.01	**0.41**

TABLE 3. Ablation study on transferring visual guidance. The performances are measured by group-level accuracy. Models with visual guidance show fairer results than models without it on all the datasets.

Dataset	Visual Guidance	Female (↑)	Male (↑)	Diff_G (↓)	$abs_{F\rightarrow M}$	$abs_{M\rightarrow F}$	Diff_A (↓)	$rel_{F\rightarrow M}$	$rel_{M\rightarrow F}$	Diff_R (↓)
VQA 2.0	X	70.76	**77.63**	6.87	19.17	13.16	6.01	65.56	58.83	6.73
	O	**75.49**	76.46	**0.97**	19.37	17.54	**1.83**	79.03	74.51	**4.52**
GQA	X	44.74	**91.38**	46.68	51.98	4.33	47.65	94.06	50.23	43.82
	O	**65.29**	81.69	**16.40**	32.60	15.37	**17.23**	93.92	83.94	**9.98**
VQA-Gender	X	81.32	82.40	1.08	14.87	13.33	1.54	79.60	75.73	3.87
	O	**83.89**	**82.95**	**1.04**	15.51	16.07	**0.56**	96.28	93.70	**2.57**

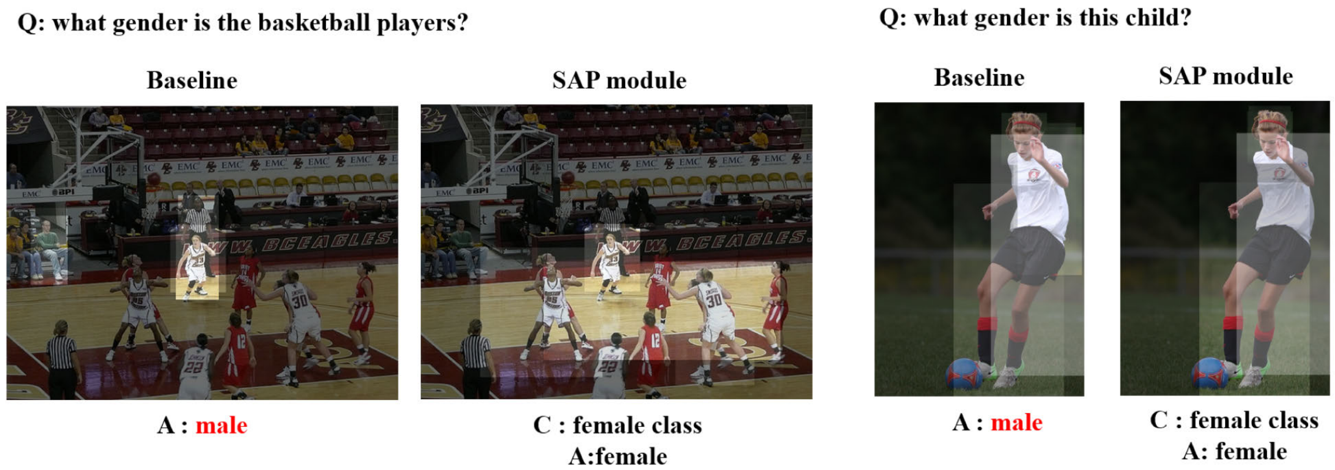


FIGURE 3. Visualized examples of attention maps on VQA 2.0 dataset. Each image contains five attended regions with the largest attention weights. The brighter boxes indicate greater attention weights. Q, A, and C denotes questions, answers, and sensitive classes, respectively.

SAP module could predict sensitive attributes more accurately with visual guidance.

In Figure 4, we demonstrate this phenomenon through visualized attention maps. In the examples, SAP module without visual guidance (VG) has difficulty to focus on accurate visual evidence, thus it mispredicts the female group to the background group. Meanwhile, SAP with visual guidance predicts the correct group by utilizing appropriate visual regions.

G. LIMITATIONS

In this section, we discuss the limitations of our study. Although our method could be potentially applied to various sensitive attributes in VQA tasks, we conduct experiments only in gender attributes due to the scarcity of ethnicity and age data in benchmark datasets. The scarcity makes it difficult not only to train and test our model stably but to construct a new dataset using our labor-saving method. Therefore, the generalizability of our method for various

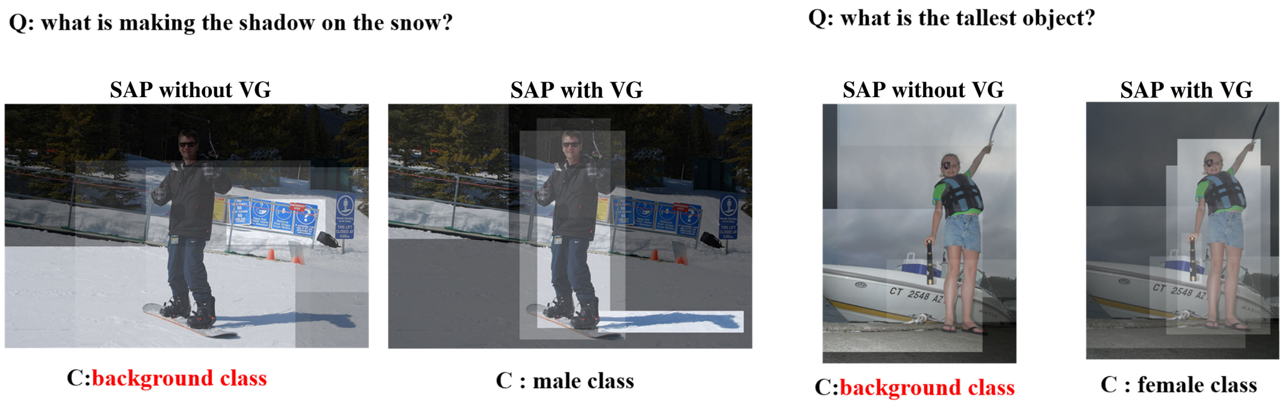


FIGURE 4. Visualized attention maps of SAP with visual guidance (VG) and without it. SAP without visual guidance does not find the right visual evidence well.

sensitive attributes is yet to be experimentally proven in this paper.

VIII. CONCLUSION

In this paper, we conducted a fairness study in VQA tasks. To mitigate unfairness in terms of sensitive attributes with graceful performance degradation, we proposed a fairness-aware VQA model, namely Fair-VQA, which is composed of two modules: VQA module and SAP module. This two-branch approach showed much fairer performances than the baseline and comparable models on VQA 2.0, GQA, and VQA-Gender datasets. We extensively evaluate the fairness of results with instance-level and group-level metrics. Moreover, through ablation study, we demonstrated that transferring visual guidance from VQA module to SAP module helps to focus on the right visual evidence and to predict correct sensitive attributes.

ACKNOWLEDGMENT

(Sungho Park and Sunhee Hwang contributed equally to this work.)

REFERENCES

[1] I. A. Hamilton. (2018). *Amazon Built an ai Tool to Hire People But Had to Shut it Down Because it Was Discriminating Against Women*. [Online]. Available: https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10

[2] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2016, pp. 4349–4357.

[3] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Conf. Fairness, Accountability Transparency*, 2018, pp. 77–91.

[4] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness GAN," 2018, *arXiv:1805.09910*. [Online]. Available: http://arxiv.org/abs/1805.09910

[5] T. Wang, J. Zhao, M. Yatskar, K. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, 2019, pp. 5309–5318, doi: 10.1109/ICCV.2019.00541.

[6] T. Yoon, J. Lee, and W. Lee, "Joint transfer of model knowledge and fairness over domains using wasserstein distance," *IEEE Access*, vol. 8, pp. 123783–123798, 2020.

[7] J. Larson. (2016). *Introduction to Bayesian Statistics*. [Online]. Available: https://github.com/propublica/compas-analysis

[8] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[9] D. Dheeru and E. K. Taniskidou. (2017). *UCI Repository of Machine Learning Databases*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Adult

[10] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4971–4980.

[11] C. Chen, D. Han, and J. Wang, "Multimodal encoder-decoder attention networks for visual question answering," *IEEE Access*, vol. 8, pp. 35662–35671, 2020.

[12] M. Lao, Y. Guo, H. Wang, and X. Zhang, "Cross-modal multistep fusion network with co-attention for visual question answering," *IEEE Access*, vol. 6, pp. 31516–31524, 2018.

[13] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6904–6913.

[14] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6700–6709.

[15] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[16] S. Ramakrishnan, A. Agrawal, and S. Lee, "Overcoming language priors in visual question answering with adversarial regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1541–1551.

[17] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, "RUBi: Reducing unimodal biases in visual question answering," 2019, *arXiv:1906.10169*. [Online]. Available: http://arxiv.org/abs/1906.10169

[18] K. Kärkkäinen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age," 2019, *arXiv:1908.04913*. [Online]. Available: http://arxiv.org/abs/1908.04913

[19] L. Celona, S. Bianco, and R. Schettini, "Fine-grained face annotation using deep multi-task CNN," *Sensors*, vol. 18, no. 8, p. 2666, Aug. 2018.

[20] R. Torfason, E. Agustsson, R. Rothe, and R. Timofte, "From face images and attributes to attributes," in *Proc. Asian Conf. Comput. Vis.* Taipei, Taiwan: Springer, 2016, pp. 313–329.

[21] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.

[22] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1931–1939.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[24] Z. Zhang, Y. Song, and H. Qi, "Age Progression/Regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5810–5818.

[25] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5309–5318.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, May 2014, pp. 740–755.

[27] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *Proc. Eur. Conf. Comput. Vis.* Munich, Germany: Springer, 2018, pp. 793–811.

[28] S. Hwang, S. Park, D. Kim, M. Do, and H. Byun, "FairfaceGAN: Fairness-aware facial image-to-image translation," in *Proc. BMVC*, 2020, pp. 1–14.

[29] S. Gong, X. Liu, and A. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 330–347.

[30] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," in *Proc. 36th Int. Conf. Mach. Learn. Proceedings of Machine Learning Research*, vol. 97. K. Chaudhuri and R. Salakhutdinov, Eds. Long Beach, California, USA: PMLR, 09–15 Jun. 2019, pp. 1436–1445. [Online]. Available: http://proceedings.mlr.press/v97/creager19a.html

[31] S. Park, D. Kim, S. Hwang, and H. Byun, "README: REpresentation learning by fairness-aware disentangling MEthod," 2020, *arXiv:2007.03775*. [Online]. Available: http://arxiv.org/abs/2007.03775

[32] G. Peng, Z. Jiang, H. You, P. Lu, S. Hoi, X. Wang, and H. Li, "Dynamic fusion with Intra- and Inter- modality attention flow for visual question answering," 2018, *arXiv:1812.05252*. [Online]. Available: http://arxiv.org/abs/1812.05252

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[36] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1564–1574.

[37] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[38] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555

[39] A. Fukui, D. Huk Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016, *arXiv:1606.01847*. [Online]. Available: http://arxiv.org/abs/1606.01847

[40] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," 2016, *arXiv:1610.04325*. [Online]. Available: http://arxiv.org/abs/1610.04325

[41] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2612–2620.

[42] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6281–6290.

[43] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

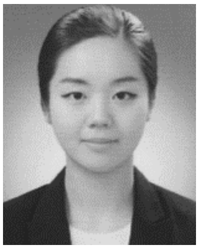
[44] D. Teney, P. Anderson, X. He, and A. V. D. Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4223–4232.

[45] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

SUNGHO PARK received the B.S. degree in computer science from Yonsei University, Seoul, Republic of Korea, in 2018, where he is currently pursuing the Ph.D. degree in computer science. His research interests include computer vision, deep learning, fairness AI, representation learning, and visual question answering.



SUNHEE HWANG (Graduate Student Member, IEEE) received the B.S. degree in information and telecommunication engineering from Korea Aerospace University, Republic of Korea, in 2014. She is currently pursuing the Ph.D. degree in computer science with Yonsei University, Seoul, Republic of Korea. Her research interests include computer vision, deep learning, fairness AI, and generative model.

JONGKWANG HONG (Member, IEEE) received the B.S. degree in computer science from Sangmyung University, Seoul, Republic of Korea, in 2012, and the Ph.D. degree in computer science from Yonsei University. His research interests include computer vision, deep learning, visual question answering, and object detection.

HYERAN BYUN (Member, IEEE) received the B.S. and M.S. degrees in mathematics from Yonsei University, Seoul, Republic of Korea, and the Ph.D. degree in computer science from Purdue University, West Lafayette, IN, USA. She is currently a Professor of computer science with Yonsei University. Her research interests include computer vision, artificial intelligence, deep learning, and pattern recognition.