

Dealing with Missing Modalities in the Visual Question Answer-Difference Prediction Task through Knowledge Distillation

Jae Won Cho Dong-Jin Kim Jinsoo Choi Yunjae Jung In So Kweon
KAIST, South Korea.

chojw@kaist.ac.kr djnjusa@kaist.ac.kr jinsc37@kaist.ac.kr
yun9298a@gmail.com iskweon77@kaist.ac.kr

Abstract

In this work, we address the issues of the missing modalities that have arisen from the Visual Question Answer-Difference prediction task and find a novel method to solve the task at hand. We address the missing modality—the ground truth answers—that are not present at test time and use a privileged knowledge distillation scheme to deal with the issue of the missing modality. In order to efficiently do so, we first introduce a model, the “Big” Teacher, that takes the image/question/answer triplet as its input and outperforms the baseline, then use a combination of models to distill knowledge to a target network (student) that only takes the image/question pair as its inputs. We experiment our models on the VizWiz and VQA-V2 Answer Difference datasets and show through extensive experimentation and ablation the performance of our method and a diverse possibility for future research.

1. Introduction

With the advancements of the Visual Question Answering (VQA) [4] task, where a model learns to generate a correct answer to a *visual query* about a given image, a new task called Visual Question Answer-Difference [6] (we call VQD for short) has been proposed where a model is required to take VQA one step further and try to understand why or how the answers of a VQA model may differ.

The nature of the VQA task allows for an important real-world application of blind people using this framework to ask questions to an AI model to help them in their everyday lives [17]. However, with the limitations of current models, the designers of the VQD task [6] try to tackle a fundamental issue of the existing framework of models being trained to output the most probable answer without taking into consideration the subtle differences of answers or even the subjective nature of answers.

The VQD task is defined as follows: given a triplet of

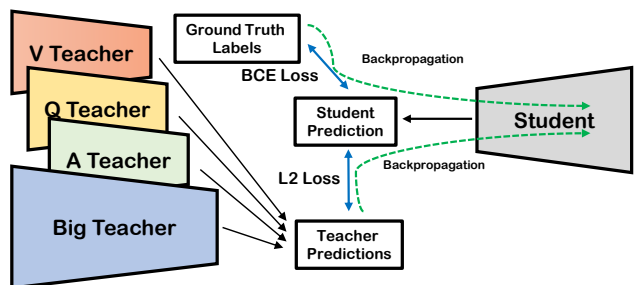


Figure 1. Simplified illustration of Knowledge Distillation using Individual Modalities and a Combined Modality “Big” Teacher. The outputs of each teacher are leveraged and loss is backpropagated using L2 Loss into the Student Model along with the ground truth loss.

Image, Question, and the 10 possible Answers gathered through crowdsourcing, an AI model is required to answer out of 10 answer difference categories why the 10 possible answers might be different given the Image and Question. Although this setting seems plausible, the creators of the task has not made the Answers of the Test set publicly available due to an ongoing challenge. This means that although we are free to use the Answers during training and validation, at test time, we are to use a different set of inputs. To reiterate, in reality, the VQD task should be defined as follows: given a pair (*not triplet*) of Image and Question, the model should be able to understand the intricacies of the Image and Question in order to output not only the correct answer but also the reasons as to why the answers could possibly be different. This makes the problem a much more challenging problem and also makes it much more real-world like.

Given this setting, we devise a method to tackle the challenge of the trying to understand the possibilities of the answers given only the Image and Question. We first propose a network that uses all 3 modalities and outperforms previous networks. Then, we propose to use a knowledge distillation [19] technique to distill knowledge about the miss-

ing modality to train another model that only has the Image and Question available to it. We show through our extensive experimentation and analysis the performance gains of this method on Vizwiz and VQA-V2 VQD dataset [6]. The main contributions of this work are summarized as follows:

1. We revisit the Visual Question Difference (VQD) task [6] to show a novel method to deal with the realistic issue of the missing modality at test time.
2. In order to effectively learn to predict VQD, we propose a new powerful network, “Big” Teacher, that attends to all three modalities and outperforms previous baselines. We believe the proposed model starts a new baseline for future research on VQD prediction task.
3. We perform extensive analysis and experiments to show our novel method’s intuitiveness and direction for possible future research in this task.

2. Related Work

Visual Question Answering Recently, by virtue of the advancements in deep learning, there has been significant improvements in VQA models [28, 32, 35], with recent VQA models mostly focusing on designing better visual/question bilinear fusion [5, 10, 11, 26, 37] or better attention mechanisms [2, 7, 25, 28, 35, 36]. However, the VQA task has shown several issues regarding its datasets such as (1) Low dependency on visual cues [14], (2) unimodal bias [1], (3) robustness [34], and (4) answer differences [4]. In this work, we specifically focus on the issue of answer difference in VQA datasets.

Answer Difference in VQA Datasets As multiple annotators are involved in the labeling process for VQA datasets, visual questions often lead to different answers from different people [4, 16, 30]. Even though we generally want to generate an answer that gives us the highest score for each visual question, a model trained on a datasets with an abundance of unique answers might generate ambiguous answers. Previously, this problem has been addressed via a consensus based performance metric [4, 30]. Recently, [6] introduced a dataset to explicitly learn why different answers occur. The work by Bhattacharya *et al.* extends prior works [4, 16, 30] which has suggested reasons why answers can differ such as visual questions are difficult, subjective, ambiguous, or containing synonymous answers. In particular, they explicitly label each visual question indicating which among nine options are the reasons for the observed answer differences for two popular VQA datasets, VizWiz [17] and VQA 2.0 [14]. The VizWiz VQD dataset, however, comes with an added challenge of the ground truth answers being missing at test time. We tackle this problem by using our novel training scheme with our new baseline.

Generalized Knowledge Distillation Our proposed learning scheme is based on the concept of generalized distillation [27] which combines the two popular “machine-teaching-machines” frameworks: knowledge distillation [19] and privileged information [33]. Vapnik and Vashist [33] first introduced a student-teacher analogy where they exploit a “teacher” model (model pre-trained with optical flow) that provides additional information about the training examples to a “student” model (model trained on RGB). The teacher model is trained with additional information (*e.g.* depth map or optical flow) available only in the training phase and not at test time [33], and they exploit the information of the teacher to better train the student. In our scenario, the ground truth answers are the privileged information available for training, along with visual/question, but only visual/question is available at test time. On the other hand, Hinton *et al.* [19] introduced the concept of knowledge distillation. The motivation stems from the desire to transfer knowledge learned from a teacher model (ensemble of models) to a student model (small model) by matching the outputs representation of the student to the outputs of the teacher in order to improve the performance of the student model. In our case, we have several models that act as teacher models. More specifically, we have models for each individual modality (visual, question, and answer) and also a model that takes the visual/question/answer triplet as inputs while our target or student model only takes visual/question pair as inputs. The generalized distillation approach is adapted by various applications that require cross-modality knowledge transfer including object detection [15, 20] and action recognition [9, 12, 13, 23, 24, 29]. To the best of our knowledge, we are the first to apply the generalized distillation scheme for VQD prediction task.

Table 1. The labels of reasons why answers differ defined by [6].

LQI	Low Quality Image
IVE	Insufficient Visual Evidence
INV	Invalid Question
DFD	Difficult Question
AMB	Ambiguous Question
SBJ	Subjective Question
SYN	Synonymous Answers
GRN	Granular (Answers present the same idea)
SPM	Spam Answers
OTH	Other

3. Methodology

In this section, we introduce the architecture of the teacher models utilized (including “Big” Teacher model) and the distillation framework we propose.

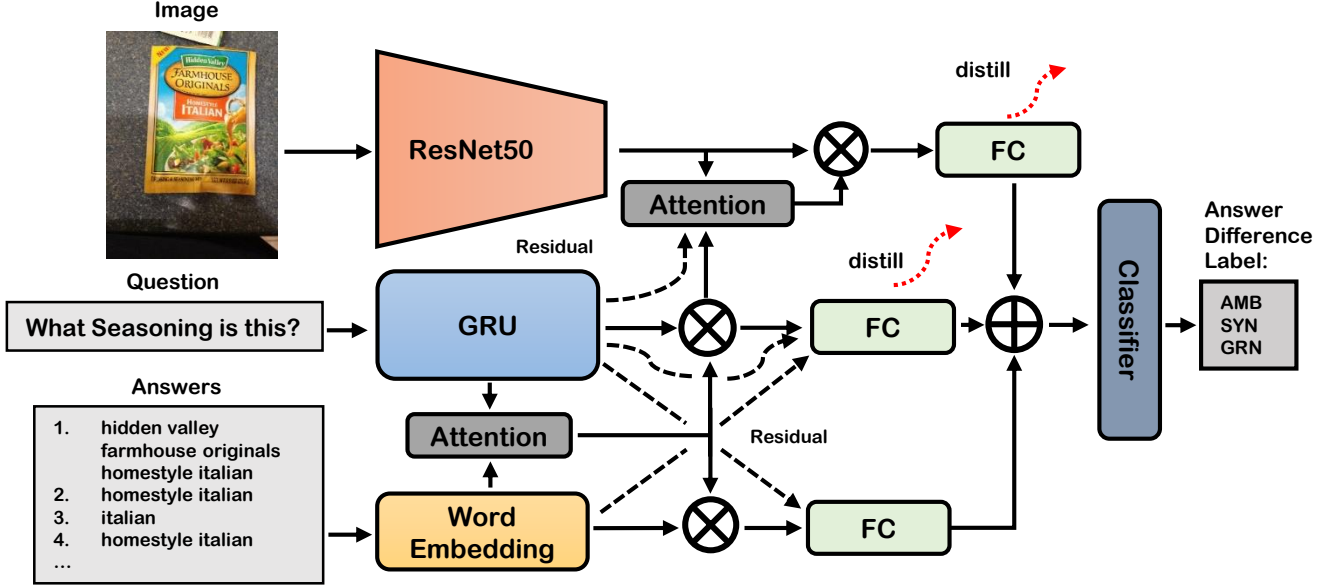


Figure 2. Architecture of our “Big” Teacher Model. We show with red dotted arrows the distilled features. Black dashed lines show residual connections.

3.1. Problem Definition

The Visual Question Answering (VQA) [4] task is required to generate a correct answer \hat{a} for a given visual question (x, q) . The question q is usually embedded into a vector with RNNs and the image x is represented by fixed-size features from CNNs. The traditional VQA model is trained with cross-entropy loss by comparing the model’s predictions with ground truth answers a .

According to [6], predicting visual question answer difference is designed as a multi-label classification problem. Given an image and question pair (x, q) , the goal is to learn from the given binary ground truth vector $Y = \{y_i\}_{i=1}^N$ for each of the N reasons ($N = 10$ including “Other” class) why the answer might be different (the reasons are listed in Table 1). For the ground truth labels, each class is labeled, either 1 (present) or 0 (not present), by five crowd workers.

While [6] utilizes the image, question, and ground truth answers as input cues for their final model, as we are not provided with the ground truth answers for the testing phase, we can only rely on the image and question. For this reason, our student model, which is our final model for answer difference predictions, only has image and question as inputs. Our final model utilizes ResNet50 [18] for image features x , and a 300 dimensional pre-trained Glove word vector [31] followed by a single-layer GRU [8].

Just like the given baseline for the model that makes use of the Question and Image (Q+I Model for short) as its input, in [6], all the input cues (image representation x and question representation q) are fed into our answer difference prediction model comprised of several fully connected

layers and outputs the class logit $Z_p = \{\hat{z}_i\}_{i=1}^N$. The final answer difference prediction $\hat{Y} = \{\hat{y}_i\}_{i=1}^N$ is computed via a Sigmoid non-linearity, and the model is trained with the Binary Cross-Entropy (BCE) loss.

3.2. Proposed Training Method

As previously stated, for the VizWiz VQD dataset, the ground truth answers for the test split datasets are not available to the public. The task deals with two unknowns since both the ground truth answers, which are important inputs for the VQD task, and the answer difference labels are missing. This is due to both the VizWiz VQA task and VizWiz VQD task being separate challenges that are concurrently open on the Eval AI test server.

Because of this issue, we propose a new method so that we can use only the questions and the image to output the difference. [6, 16] claims that even without the answers, a model should be able to anticipate the different kinds of answers a question and image pair should have. In practice, however, this model does not perform as accurately as a model that has the ground truth answer as its input as this input contains significantly more information to the model. Intuitively, compared to a model that has all three modalities present, a model that has to rely on only the question and the image to guess why the answers *could* be different would perform much worse as it is much more difficult. In addition, since we do not have the implementation details of the Question, Image, and Answer (Q+I+A) Model given by [6], we set our baseline model as the Q+I Model.

In light of this, we leverage a *teacher-student* framework

Table 2. The notations of the distillation losses. We make use of 7 distillation losses in total by exploiting three teachers (Visual, Question, and “Big” Teacher) intermediary features and the final predictions.

\mathcal{L}_{21}	\mathcal{L}_{22}	\mathcal{L}_{23}	\mathcal{L}_{24}	\mathcal{L}_{25}	\mathcal{L}_{26}	\mathcal{L}_{27}
$\mathcal{L}_{L2}(Z_{vi}^v, Z_{vi}^s)$	$\mathcal{L}_{L2}(Z_{qi}^q, Z_{qi}^s)$	$\mathcal{L}_{L2}(Z_{vi}^b, Z_{vi}^s)$	$\mathcal{L}_{L2}(Z_{qi}^b, Z_{qi}^s)$	$\mathcal{L}_{L2}(Z_p^v, Z_p^s)$	$\mathcal{L}_{L2}(Z_p^q, Z_p^s)$	$\mathcal{L}_{L2}(Z_p^b, Z_p^s)$

with knowledge distillation [19] to solve the current problem at hand. In this problem definition, we require several teacher models. From Hinton *et al.* [19] and intuitively, transferring more features and information generally leads to less overfitting and better performance compared to a model that lacks this extra information. Although the best performing models would have all modalities, with the current limitations of this task, we believe that using different combinations of teacher models is the best solution.

3.3. Individual Modality Teacher Models

As we propose that each modality gives important information to the student model when distilling knowledge based on what each modality learns, we describe in detail the teacher models. The *Visual Teacher* consists of a ResNet50 [18] backbone to generate Z_{vi}^v , and we use Z_{vi}^v through a simple Multi-Layer Perception (MLP) to generate class logit Z_p^v . The performance of this model is similar to that reported in [6] as it follows the same architecture. The *Question Teacher* consists of a Glove [31] embedding and a GRU [8] backbone. Similar to the visual teacher, the embedded question Z_{qi}^q is passed through a MLP to generate class logit Z_p^q . The performance of this model is also similar to that reported in [6] for similar reasons. We then propose a novel “*Big*” *Teacher*, a teacher model that outperforms the given baseline model with ground truth answer inputs and use this model to ultimately guide the student to make the correct predictions. This model uses all modalities and is the most complex model among all our models.

We also consider the *Answer Teacher* that uses the diverse ground truth answers as input and tries to understand the reasons for this diversity. This is slightly different from the other teacher models in that it directly takes the answers and tries to guess why the answers would be different without any other cues. This is a naive approach and could be seen as counter-intuitive to use on its own, but we still test it to show its potential impact. The official repository uses a vocabulary of 6250 dimensions and this value is achieved only by taking answers that have appeared more than 5 times within the entire train/val/test dataset. The total number of unique answers that are given due to the issues of the open-ended crowd sourcing are given as 58,789 [17] and using all the unique answers might seem inefficient since there are too many values. However, if the threshold is set at 5, too many unique answers are disregarded and the model may not be able to pick up on the minute differences that occur from either ambiguous or granular reasons. To

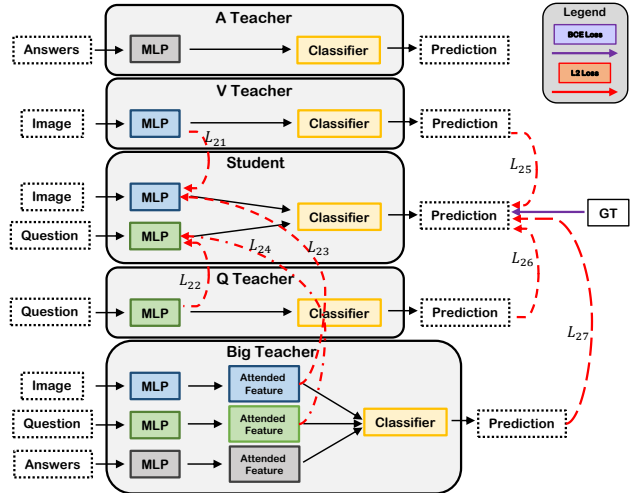


Figure 3. Detailed model of our distillation method. The L_{2i} listed correspond to the losses found in Table 2. Although A Teacher is shown in the figure, we do not use A Teacher in our final model, so we do not draw the distillation loss arrows

this end, we recreate the answer vocabulary to include all available answers, from the train/val set, without the test set due to our lack of access, and set the vocabulary size to be 45,304. We also use this same vocabulary space for our “*Big*” *Teacher* model’s answer input for the same reasons.

3.4. “Big” Teacher Model

Our “*Big*” *Teacher*, as shown in Fig. 2, uses all modalities present, including the ground truth answers. The “*Big*” *Teacher* uses a ResNet50 [18] backbone for image features and Glove [31] embeddings with a GRU [8] for question features and uses the [3]’s attention to attend to all three modalities. Traditional attention simply attends the question to the image to see where the question needs to attend in the image. However, since we are dealing with three different modalities, we devise a different architecture suitable for attending three modalities.

First, we can attend the question to the answer, and thereby we can see whether or not the answer to the question is present. At the same time, we can also attend the answer to the question to see if there is a part of the question that is important to the question. Attention goes both ways, so we attend it both ways as shown in Fig. 2. For the question that has been attended by the answer, since it holds information about which part of the question is important to the answer, we attend this attended question to the image to see if there

are parts of the image that correspond to this attended question. Additionally, in order to stop error propagation, we apply residual connections across the model. We show that this model outperforms the given baseline significantly. As a result, three feature vectors from three modalities are generated: Z_{vi}^b, Z_{qi}^b , and Z_{ai}^b for visual, question, and answer features respectively. The three feature vectors are summed up and pass through additional MLP to generate class logit Z_p^b similar to multimodal fusion methods [25, 21].

3.5. Student Model

The *Student Model* follows the same architecture of the Q+I model in [6] and consists of a ResNet50 [18] backbone with Glove [31] embeddings and a GRU [8] for its respective inputs. The features are passed through a FC layer to generate intermediary feature vectors (Z_{vi}^s and Z_{qi}^s) then multiplies into a FC classifier for the output class logit Z_p^s . The final answer prediction from the student model \hat{Y}^s is computed by feeding Z_p^s into Sigmoid non-linearity. A simplified *Student Model* is shown in Fig. 3.

3.6. Teaching the Student

Teacher Models need to be pre-trained in order to teach the student, so we train the teacher models with their respective inputs with BCE loss, then we use each of the Teacher Models as a guide in training the *Student Model*. From each Teacher Model t ($t \in \{v, q, b\}$), we make use of $L2$ loss to transfer the knowledge to the Student Model by matching the output:

$$\mathcal{L}_{L2}(Z^t, Z^s) = \|Z^t - Z^s\|_F^2. \quad (1)$$

In addition to applying $L2$ loss on the predictions of the Teacher and Student Models, we also leverage the intermediary features of the Teacher Models, specifically from the *Visual*, *Question* and *“Big” Teacher*. We use the intermediary visual and question features from the respective teachers and leverage $L2$ loss. These features are the features before the classifier and can be shown in Fig. 3.

As single modality Teachers are trying to learn from their given modalities, each Teacher Model tries to leverage as much knowledge as it can from its given modality as there is no other modality to depend on. Due to this, if we force our Student Model to learn something from this feature representation, the Student Model can learn something that it normally would not due to the presence and dependence on another modality.

As for the *“Big” Teacher* Model, since the answer attends to the visual and question features, intermediary visual and question features hold information about the missing modality. By distilling knowledge from these representation, the student model can learn to mimic the kind of representation it needs to follow. By this logic, the Student

Model learns to anticipate the kind of diverse answers that can exist from the question and image pair.

For our final model, we apply 7 $L2$ losses on the Student Model with one BCE loss as shown below. We use the intermediary features and predictions of the *Visual*, *Question*, and *“Big” Teacher* to distill the loss to the *Student Model*. The final loss equation is shown below:

$$\mathcal{L}_{total} = \sum_{i=1}^7 \lambda_i \mathcal{L}_{2i} + \lambda_0 \mathcal{L}_{BCE}(Y, \hat{Y}^s), \quad (2)$$

where $\{\lambda_i\}$ and λ_0 are weights for losses with L_2 being $L2$ losses and L_{BCE} being BCE loss. The seven losses are a summation of individual losses and the detailed description of the seven distillation losses are shown in Table 2.

In applying the loss in this form, the Student Model is able to understand diverse feature representations from each of the given modalities, allowing the Student to perform better than it could without this guide. A visual description is shown in Fig. 3 to aid in understanding the method at hand.

4. Experiments

In this section, we describe our experimental setups, experimental results, and implementation details.

4.1. Implementation Details

As previously mentioned, we use a CNN backbone of ResNet 50 [18] with a single layer GRU, and a modified answer space of 45,304. We use a word embedding to embed the answers into features instead of an RNN as the answers are not sentences. We set all hidden sizes to be 1024 throughout for all models present in our system including the attention layers. Each MLP used in the models consists of 2 FC layers with a ReLU activation in between. We use an Adam optimizer for all models with $\gamma = 0.1$ with a learning rate of $1e^{-2}$. All teacher models are pretrained for 5 epochs while student models are trained for 20 epochs.

4.2. VizWiz VQD Dataset Setup

We first test our method on the VizWiz VQD dataset [17], which consists of 19,176/3,063 (image, question, answers) triplets in the train/val splits respectively with additional 7,668 triplets for the test split. We only use the train set to train and validation set to test due to the test set not being publicly available. Each question is independently annotated with 10 answers and reasons for why answers differ, which has 10 classes with independent number of people annotating this separately. As mentioned above, we reconstruct a new vocabulary space for the answers from the train/val split.

Table 3. The comparison between our “Big” Teacher Model (Ours) with the Baseline Models. Our “Big” Teacher Model shows the best performance among the baselines.

	Overall	LQI	IVE	INV	DFE	AMB	SBJ	SYN	GRN	SPM	OTH
Q	39.45	33.75	51.93	35.23	11.04	83.41	11.99	79.06	84.50	2.88	0.81
I	40.50	56.69	50.44	26.98	7.58	83.18	9.44	80.27	86.40	2.12	1.91
A	50.10	65.28	77.07	55.92	6.96	88.16	11.97	89.89	94.06	2.50	8.67
Q+I	44.91	57.49	60.46	41.45	10.96	86.02	12.73	85.38	91.26	1.98	1.34
Q+I+A	49.72	66.65	75.51	53.86	10.18	89.11	11.86	89.84	95.47	2.33	2.47
Ours	51.60	66.63	77.98	57.74	9.7	89.14	12.89	90.63	95.79	2.06	13.46

Table 4. The performance of a Student Model with different combinations of Teacher Models with individual modalities, Visual (V), Question (Q), and Answer(A). Simply using single-modality Teacher Models is *ineffective* in improving the Student Model.

	Overall	LQI	IVE	INV	DFE	AMB	SBJ	SYN	GRN	SPM	OTH
Baseline	44.91	57.49	60.46	41.45	10.96	86.02	12.73	85.38	91.26	1.98	1.34
A	44.05	57.15	61.67	39.8	6.07	85.79	10.31	85.74	91.58	2.03	0.41
Q	40.58	39.01	54.14	36.88	9.9	84.19	11.2	80.84	86.61	2.27	0.79
V	40.96	55.68	52.13	28.89	7.76	83.77	9.32	81.72	87.15	2.02	1.15
V&Q	43.95	56.14	58.03	37.91	10.88	85.81	11.83	84.82	90.49	2.48	1.14
Q&A	43.99	53.12	59.65	41.1	8.31	86.47	11.85	85.66	90.88	2.32	0.59
V&A	44.45	58.92	60.85	38.84	7.81	85.99	10.55	85.45	90.82	2.29	0.99
V&Q&A	44.68	56.83	60.06	39.27	10.02	86.92	12.02	86.14	94.41	2.46	1.69

4.3. Answer Difference Baselines

Since we use a different split from [6], we set up our own baselines. [6] does not give clear implementation details for the Question, Image, Answer (Q+I+A) Model, so we assume the Q+I+A.GT¹ Model as the Q+I+A Model from here on for the purposes of this paper and show our comparisons in Table 3.

The Q, I, and A Models are models trained on individual modalities and are the models that we use as our teachers. The Q+I+A Model is the baseline model that [6] proposes trained on our hyperparameters with the changed answer space, and the “Big” Teacher Model that we propose is denoted as *Ours*. As shown in Table 3, Q Model shows high performance in question related classes (*i.e.* INV, DFE, SBJ), and I Model shows high performance in image related class (*i.e.* LQI). As answers hold essential information to reason the answer difference, the A Model shows higher scores compared to Q Model and I Model in almost all classes and especially so in the Other category. Combining question and image information leads to performance improvements (Q+I Model), and adding answer information further improves the performance. Finally, we show from our “Big” Teacher Model that attending all three modalities with our method further boosts the performance. This more powerful model is shown to be crucial in distilling knowledge to the student model.

¹Q+I+A.GT is the model that takes the ground truth answers as input

4.4. Results of Knowledge Distillation

In light of the baseline models and the performance of our “Big” Teacher Model, we conduct extensive experimentation to see how each distillation method would affect the performance of the student model. “Baseline” is the Q+I Model without any knowledge distillation loss. We evaluate all possible teacher combinations of A Model, Q Model, and V Model including intermediate feature losses denoted as “w/I”. We list our findings in Table 4.

Table 4 show that distilling individual modalities into the baseline does not necessarily aid in the student model, but shows that it can be counter productive to the model. For example, in Table 4, we can see that not having any visual ques from the teachers would actually harm the model in understanding of the image such as LQI (Low Quality Image) as shown in red. This clearly shows the degradation that can occur from the individual modality teachers. Generally for the criterions that are easy regardless, *AMB*, *SYN*, *GRN*, it shows that the difference is minuscule. Although the Answer Teacher performs significantly better than the Question and Visual Teachers, the Student Model does not seem to glean too much knowledge from the Answer Teacher. This could be due to the fact that the Answer Teacher’s predictions are based on a different modality that is not present in the Student Model. Here, we show that the individual modality teachers are not helpful for the Student Model.

To combat this, we further experiment with distilling knowledge from the “Big” Teacher Model and show our ablations in Table 5, and we find the most powerful combination is to “Big”, Visual, and Question Teacher with all

Table 5. The performance of a Student Model with combinations of Teacher Models that includes our “Big” Teacher. (w/I) means the intermediary features. Note that combining all the Teacher Models (denoted as All) is not helpful. The Student Model trained using the “Big”, Visual, and Question Teacher with all their intermediary features shows the best performance.

	Overall	LQI	IVE	INV	DFF	AMB	SBJ	SYN	GRN	SPM	OTH
Baseline	44.91	57.49	60.46	41.45	10.96	86.02	12.73	85.38	91.26	1.98	1.34
Big	43.74	56.95	60.94	41.16	8.27	86.18	11.61	86.25	91.7	2.14	0.69
Big (w/I)	44.61	57.51	50.88	40.2	9.97	86.4	11.34	85.71	91.14	2.38	0.55
Big&Q	44.23	51.56	60.61	41.74	10.65	85.78	12.15	85.8	91.03	2.02	0.9
Big&Q (w/I)	44.48	51.56	61.51	41.01	9.98	86.27	13.75	85.87	91.34	2.78	0.72
Big&V	45.36	58.01	61.09	39.12	9.71	86.26	12.2	85.54	91.1	2.06	8.53
Big&V (w/I)	44.67	58.27	60.42	38.02	10.5	86.67	10.37	85.38	90.98	2.32	1.74
Big&A	44.46	56.51	61.99	39.84	7.95	86.48	11.82	85.84	91.38	2.25	0.55
Big&A (w/I)	44.18	57.54	60.51	38.51	6.7	86.3	11.35	86.19	91.67	2.15	0.88
Big&Q&A	44.25	54.9	61.98	40.77	6.81	86.63	11.39	85.91	91.11	2.45	0.53
Big&Q&A (w/I)	44.05	55.07	60.12	39.6	7.49	86.23	11.0	86.08	91.63	2.3	0.97
Big&V&A	44.23	57.98	61.19	38.06	7.85	86.6	11.02	85.6	91.17	2.07	0.75
Big&V&A (w/I)	44.07	58.27	60.35	36.72	7.14	86.5	11.03	85.8	91.34	2.5	1.11
Big&V&Q	45.41	59.09	61.49	39.57	11.91	86.47	14.08	86.23	91.66	2.08	1.5
Big&V&Q (w/I)	45.75	58.46	62.39	39.87	12.71	86.52	11.52	86.31	91.85	2.32	5.52
All	44.85	56.14	60.27	39.4	10.6	87.09	13.24	86.29	91.24	2.26	1.95
All (w/I)	44.73	57.12	61.22	40.07	9.25	86.55	11.77	86.13	91.52	2.42	1.28

Table 6. The Baseline Models for the VQA 2.0 Answer Difference Dataset.

	Overall	LQI	IVE	INV	DFF	AMB	SBJ	SYN	GRN	SPM	OTH
Q	43.61	8.04	58.62	44.25	28.67	96.74	23.99	89.1	85.02	1.33	0.36
I	31.52	3.93	29.24	8.55	16.96	92.52	17.29	73.81	72.12	0.55	0.2
A	42.57	9.62	59.76	46.86	18.94	96.49	17.3	89.63	85.81	0.91	0.42
Q+I	43.19	8.32	57.27	41.32	28.98	96.35	24.87	88.31	83.49	2.72	0.34
Q+I+A	43.2	8.57	57.52	40.33	27.88	96.73	26.12	89	84.67	0.87	0.33
Ours	44.32	10.39	59.73	43.22	31.5	96.89	24.19	89.73	85.42	1.75	0.39

their intermediary features. The intermediary features from the “Big” Teacher gives a guide to the Student Model’s features as it has rich representations embedded in it from attention. In addition, although the “Big” Teacher’s visual and question features are rich, the “Big” Teacher is prone to being overly dependent on the answers, and we can see this through the *Big* metric that it still does not outperform the best model. To combat the “Big” Teacher’s reliance on the ground truth answers, we apply the losses of the Visual and Question Teacher and show that the Visual and Question Teachers’ intermediary features aids in boosting performance significantly.

4.5. Testing on the VQA 2.0 Dataset

The VQA 2.0 Answer Difference Dataset has a smaller dataset size of 15034, (train/val/test: 9735/1511/3788) compared to the 29,907 in the VizWiz Answer Difference Dataset. The VQA 2.0 Dataset has the ground truth answers available for the test split, thus we use the trainval split to train the models and the test split for evaluation.

Generally, the trend of the classes that are easily wrong are different from that of the VizWiz Dataset. Table 6 shows that the VQA dataset shows to have a much lower LQI (Low Quality Image) score as there are few number of such images within the dataset unlike the VizWiz Dataset. For the answers that are easy on the VizWiz set, *i.e.* AMB, SBJ, and GRN (Ambiguous, Synonym, and Granular), we see that the model’s performance is not hugely affected. On the other hand, it seems that the classes such as DFF and SBJ (Difficult and Subjective) show a relatively higher score.

By looking at the Visual Teacher, or the standalone Visual Model, on the VQA dataset, it is very difficult to find any answer differences just by looking at the image. This idea is transferred on to the classifications of the classes as we see the clear performance drop in the class LQI. Surprisingly, in the ablation study on VQA 2.0 dataset as shown in Table 4, the individual modalities noticeably improve the model’s performance. However, the best performing model still requires the “Big” Teacher Model as shown in Table 5.

Fig. 4 shows qualitative results of how the Teacher Mod-


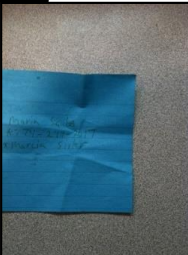

V	Q	 <p>'What toppings are on this pizza? Is it a pepperoni pizza, is it pepperoni and sausage, is it a supreme?'</p>	A	Predictions:										
					LQI	IVE	INV	DFF	AMB	SBJ	SYN	GRN	SPM	OTH
		pepperoni pepperoni pepperoni pepperoni chest pepperoni pepperoni pepperoni pepperoni pepperoni pepperoni pizza		Baseline	0.58	0.70	0.35	0.08	0.46	0.03	0.30	0.33	0.02	0.00
				V Teacher	0.27	0.31	0.13	0.05	0.72	0.04	0.65	0.74	0.01	0.01
				Q Teacher	0.40	0.64	0.29	0.11	0.55	0.03	0.37	0.49	0.02	0.01
				Big Teacher	0.03	0.03	0.01	0.01	0.89	0.03	0.90	0.97	0.04	0.01
				Student	0.31	0.47	0.16	0.05	0.64	0.02	0.51	0.66	0.01	0.00
				GT	0	0	0	0	0	0	1	1	0	0
V	Q	 <p>'Can you read this phone number? Thank you'</p>	A	Predictions:										
					LQI	IVE	INV	DFF	AMB	SBJ	SYN	GRN	SPM	OTH
		unsuitable 774 244 2017 774 244 2617 *774 244 2017 774 244 2017 774 244 2017 yes 7742442617 insufficient image quality 774 244 2017		Baseline	0.46	0.62	0.56	0.11	0.49	0.02	0.35	0.38	0.02	0.00
				V Teacher	0.28	0.32	0.13	0.05	0.72	0.05	0.65	0.73	0.01	0.01
				Q Teacher	0.35	0.57	0.39	0.10	0.61	0.04	0.44	0.54	0.02	0.01
				Big Teacher	0.19	0.30	0.12	0.07	0.89	0.09	0.78	0.87	0.03	0.01
				Student	0.33	0.46	0.25	0.07	0.68	0.05	0.51	0.62	0.02	0.01
				GT	0	0	0	0	1	0	1	0	0	0
V	Q	 <p>'Id like to know whats contained in this can'</p>	A	Predictions:										
					LQI	IVE	INV	DFF	AMB	SBJ	SYN	GRN	SPM	OTH
		baked beans baked beans beans baked beans in tomato sauce farmgirl baked beans baked beans baked beans beans baked beans beans		Baseline	0.54	0.58	0.28	0.08	0.55	0.04	0.44	0.46	0.02	0.01
				V Teacher	0.25	0.31	0.12	0.05	0.73	0.05	0.66	0.75	0.01	0.01
				Q Teacher	0.35	0.50	0.25	0.07	0.65	0.03	0.54	0.63	0.02	0.01
				Big Teacher	0.03	0.05	0.01	0.01	0.93	0.03	0.93	0.98	0.03	0.00
				Student	0.21	0.30	0.11	0.05	0.75	0.04	0.68	0.80	0.02	0.01
				GT	0	0	0	0	1	0	1	1	0	0

Figure 4. Qualitative examples showing how the Teacher Models affect the Baseline Model, and how it changes its answers after distillation. The output probability predictions are computed via Sigmoid function, and we hold the probability threshold of 0.5 to be a positive prediction. The numbers in red show wrong answers and green show correct answers. While the Baseline Model has wrong predictions, by virtue of the Teacher Models, our Student Model is able to generate correct predictions.

els affect the Baseline (Q+I) Model’s performance. The figure shows the changes in predictions that the Student Model makes after having knowledge distilled from the Teacher Models. The model named “Baseline” is the student model without knowledge distillation, and the model named “Student” is our proposed model which learns through knowledge distillation. In most of the cases, although the Baseline generates wrong predictions, by learning from teachers that can generate correct outputs, the student model is also able to generate correct predictions. One drawback is that if the Teachers are all wrong, the Student can also be affected by this and change from a correct answer to a wrong answer (e.g. AMB class in the first example and the GRN class in the second example). This shows the importance of a plausible Teacher Model when distilling the knowledge.

5. Conclusion

In this paper, we revisit the task of Answer-Difference Prediction with Visual Question Answering. We study how

to deal with the realistic problem of ground truth answers not being available at test time. First, we devise a method to improve the performance of a model with the given modalities. Through the use of all available modalities, we train individual teacher models and a plausible “Big” Teacher model. Our “Big” Teacher model shows favorable performance against the current baselines and sets a new bar for future works to come. Ultimately, we use the given teacher models to distill the information into a Student model with the privileged information they are given as our solution for our given problem. We believe this work can be the new first step towards solving VQA task and can inspire the future research on VQA or other multi-modal tasks, such as image captioning, with missing modalities [22].

Acknowledgements. This work was supported by the Institute for Information & Communications Technology Promotion (2017-0-01772) grant funded by the South Korean government.

References

- [1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Anirudha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.
- [3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.
- [4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Int. Conf. Comput. Vis.*, 2015.
- [5] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Int. Conf. Comput. Vis.*, 2017.
- [6] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *Int. Conf. Comput. Vis.*, 2019.
- [7] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
- [8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [9] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
- [10] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
- [11] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.
- [12] Nuno C Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Dmcl: Distillation multiple choice learning for multimodal action recognition. *arXiv preprint arXiv:1912.10982*, 2019.
- [13] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Eur. Conf. Comput. Vis.*, 2018.
- [14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.
- [15] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.
- [16] Danna Gurari and Kristen Grauman. Crowdverge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.
- [17] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.
- [19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [20] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.
- [21] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.
- [22] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.
- [23] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, Youngjin Yoon, and In So Kweon. Disjoint multi-task learning between heterogeneous human-centric tasks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.
- [24] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *Eur. Conf. Comput. Vis.*, 2020.
- [25] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Adv. Neural Inform. Process. Syst.*, 2018.
- [26] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016.
- [27] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.
- [28] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Adv. Neural Inform. Process. Syst.*, 2016.
- [29] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *Eur. Conf. Comput. Vis.*, 2018.
- [30] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Int. Conf. Comput. Vis.*, 2015.

- [31] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [32] Idan Schwartz, Alexander Schwing, and Tamir Hazan. High-order attention models for visual question answering. In *Adv. Neural Inform. Process. Syst.*, 2017.
- [33] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.
- [34] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4951–4961, 2018.
- [35] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.
- [36] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019.
- [37] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018.