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) [1] 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 [2] (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.

Due to the ongoing nature of Vizwiz VQA, the Test set answers are not available to the public. In this work, we address the lack of Test set answers in VQD. To reiterate, in reality, the VQD task is 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.

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 [3] technique to distill knowledge about the missing 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 [2].

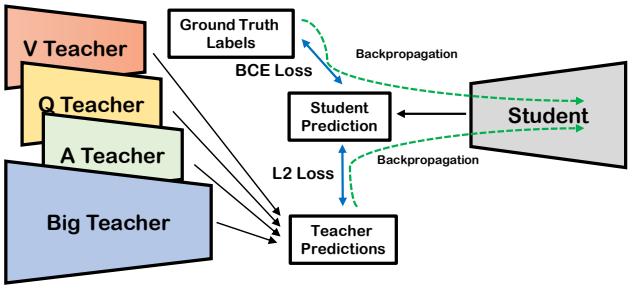

Figure 1. Simplified illustration of Knowledge Distillation using Individual Modalities and a Combined Modality "Big" Teacher. The outputs of each teacher are leveraged and loss is backpropagated using L2 Loss into the Student Model along with the ground truth loss.

2. Methodology

To deal with the missing modality, we leverage a *teacher-student* framework with knowledge distillation [3] to solve the current problem at hand. In this problem definition, we require several teacher models. From Hinton *et al*. [3] and intuitively, transferring more features and information generally leads to less overfitting and better performance compared to a model that lacks this extra information. As we propose that each modality gives important information to the student model when distilling knowledge based on what each modality learns, so we create individual modality teacher models as shown in Fig. 2. In addition, we create a "Big" Teacher, as shown in Fig. 3, uses all modalities present, including the ground truth answers.

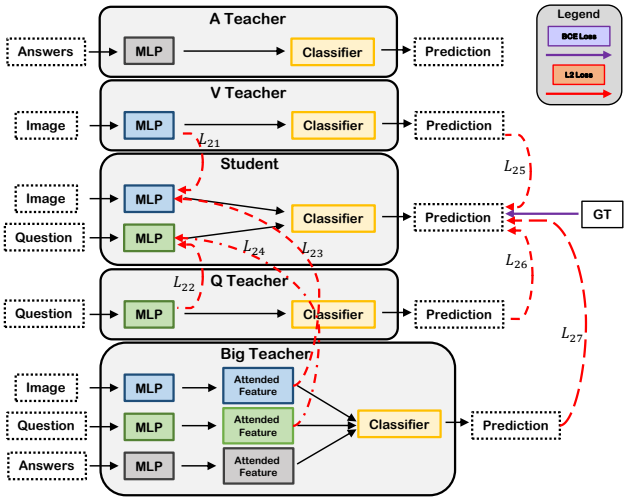

Figure 2. Detailed model of our distillation method. The L_{2i} listed are distillation losses. Although A Teacher is shown in the figure, we do not use A Teacher in our final model, so we do not draw the distillation loss arrows

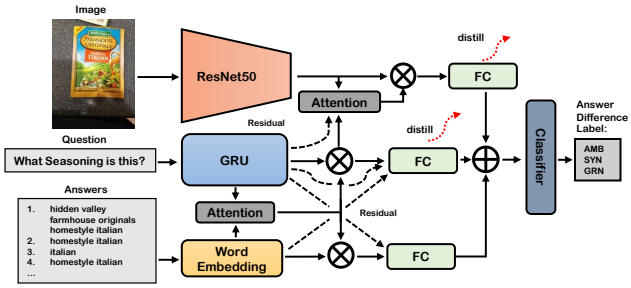

Figure 3. Architecture of our *"Big" Teacher* Model. We show with red dotted arrows the distilled features. Black dashed lines show residual connections.

We set our *Student Model* to be the same as that of the Q+I model in [2]. From the given teacher models, we distill the knowledge using knowledge distillation losses and train our student model. With our final loss equation shown below:

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

where $\{\lambda_i\}$ and λ_0 are weights for losses with L_2 being $L2$ losses and L_{BCE} being BCE loss.

3. Experiments

We show in the following tables the experimental results on the VizWiz VQD dataset. Table 1 shows the ineffectiveness of using single modal teachers. Table 2 shows that using all teacher models are not actually effective in improving the performance. Overall, we show that using the "Big", Visual, and Question Teachers with intermediary features shows the best performance.

Table 1. Student Model performance with individual modality teachers, Visual (V), Question (Q), and Answer(A). Single-modality Teacher Models are *ineffective*.

	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
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**

Table 2. Student Model performance with combinations Teacher Models. (w/I) means the intermediary features. Note that combining all the Teacher Models (denoted as All) is not helpful.

	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

4. Conclusion

We revisit the task of Answer-Difference Prediction with Visual Question Answering and study how to deal with the realistic problem of ground truth answers not being available at test time. We devise a method to improve the performance of a model with the given modalities by training teacher models. We show that using 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 VQD task and can inspire the future research on VQA or other multi-modal tasks, such as image captioning, with missing modalities [4].

References

[1] 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.

[2] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *Int. Conf. Comput. Vis.*, 2019.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[4] 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.