

Single-Modal Entropy based Active Learning for Visual Question Answering

1 Supplementary

The contents of this supplementary material include implementation details of our model, additional qualitative results, and additional analysis which were not included in the main paper due to space limitations.

1.1 Implementation Details

The VQA 2.0 Dataset can be found at this link (<https://visualqa.org/>). Following the model from [1], we make use of the pre-extracted image features from Mask-RCNN [2] for the visual input and Glove6B [3] embedding with a GRU [4] to process the question features. Then, the features are attended using the modified version of the attention proposed in [1]¹. We train our model with a batch size of 512 and use the Adamax optimizer from the PyTorch library. The hyperparameters used are: Learning Rate: 0.002, Betas: (0.9, 0.999), eps: $1e^{-0.8}$, Weight Decay: 0. Model hyperparameters are listed next. The vocabulary size used is 19902, the output of the Mask-RCNN is 2048, as for the Attention, the input is set at 2048 for the visual input, but all hidden states, question input, and output values are set to 1024. The word embedding size used is 300, and the GRU hidden state is set to 1024. The classifier, which can be seen as all 3 classifiers as stated in the method section of our paper, has an input of 1024 and an output of 3129, which is the number of answer candidates.

The VQA accuracy measure is from [2] and can be summarised as follows:

$$Accuracy(A) = \min \left\{ \frac{\# \text{ of humans that said } A}{3}, 1 \right\}, \quad (1)$$

where A is the Answer prediction from the model.

For our Active Learning setup, we set 10 stages with 40,000 question image pairs for each step, and increase 40,000 pairs each time, ending up at 400,000 at the final stage. We do not reinitialize the model each time but continue using the model that was previously trained. At Stage 0, we train the model for 20 epochs. For the next Stages from 1 to 9, as the model is already pre-trained in the previous stages, we train for only 10 epochs each. At the end of each epoch, the model is evaluated on the validation set that stays static and the

© 2021. The copyright of this document resides with its authors.

It may be distributed unchanged freely in print or electronic forms.

¹The modified version of the attention proposed in [1] is found on this github link: <https://github.com/hengyuan-hu/bottom-up-attention-vqa>

Algorithm 1 Training Our Model with Self-distillation

input Labeled pool \mathcal{D}_L , Initialized models $g_v(\cdot), g_q(\cdot)$, and $F(\cdot, \cdot)$ with parameters θ_v, θ_q , and θ respectively

input Hyper-parameters : learning rate η , maximum number of epoches $MaxEpoch$

output $g_v(\cdot), g_q(\cdot)$, and $F(\cdot, \cdot)$ with trained θ_v, θ_q , and θ

for $e = 1$ to $MaxEpoch$ **do**

sample $(X, Y) \in \mathcal{D}_L$

Compute \mathcal{L}_{main}

Compute \mathcal{L}_v

Compute \mathcal{L}_q

Update the model parameters:

$\theta \leftarrow \theta - \eta \nabla \mathcal{L}_{main}$

$\theta_v \leftarrow \theta_v - \eta \nabla \mathcal{L}_v$

$\theta_q \leftarrow \theta_q - \eta \nabla \mathcal{L}_q$

end for

Algorithm 2 Our Proposed Sampling Strategy with Single-Modal Entropic Measure (SMEM)

input Labeled pool \mathcal{D}_L , Unlabeled pool \mathcal{D}_U , Number of samples collecting for each stage b , Models $g_v(\cdot), g_q(\cdot)$, and $F(\cdot, \cdot)$

output Updated \mathcal{D}_L and \mathcal{D}_U

From \mathcal{D}_U , collect a set of samples $\{X^s\}$ with

$\max_b S(V, Q) = \max_b \alpha H(Y_q)$

$+ (1 - \alpha) H(Y_v) + \beta JSD(\mathbf{y}_v || \mathbf{y}_q) + \gamma H(\hat{Y})$

$\{(X^s, Y^s)\} = ORACLE(\{X^s\})$

$\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \{(X^s, Y^s)\}$

$\mathcal{D}_U \leftarrow \mathcal{D}_U - \{X^s\}$

evaluation score of the best is chosen at the end of each stage and reported as the accuracy of that stage including Stage 0.

Our model takes around 10 hours to train all 10 stages on a single Nvidia Titan Xp (12GB) GPU.

Description of the Training and Sampling Procedure. The detailed description of the algorithms for training our model and active sampling are shown in Alg. 1 and Alg. 2 respectively.

1.2 Additional Analysis

Other Approaches of Active Learning. As the comparison target of our approach, we also test different measures for distribution discrepancy. In the following paragraphs, we introduce several possible candidates to measure the distances between single-modal and multi-modal answer representations.

(1) **KL-Divergence (KLD)** is defined as follows:

$$D_{KL}(\mathbf{y} || \mathbf{y}_q) = \sum_{i=1}^{|\mathcal{A}|} \mathbf{y}^i \log(\mathbf{y}^i / \mathbf{y}_q^i), \quad (2)$$

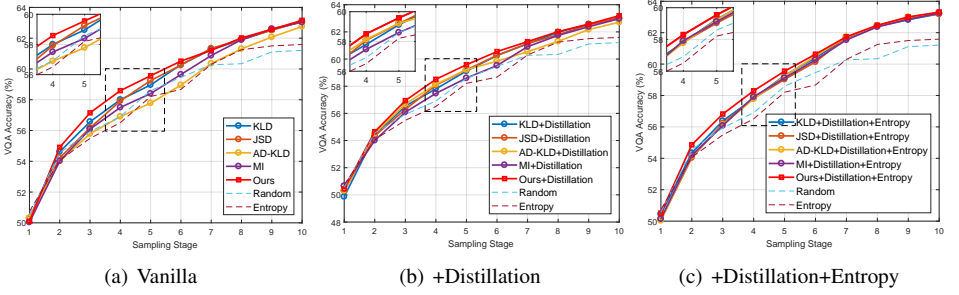


Figure 1: Comparison between different measuring methods for our method. Ours is the SMEM + JSD in all graphs above. (a):without anything, (b):with distillation loss, and (c):with both distillation loss and entropy.

$$D_{KL}(\mathbf{y}||\mathbf{y}_v) = \sum_{i=1}^{|\mathcal{A}|} \mathbf{y}^i \log(\mathbf{y}^i / \mathbf{y}_v^i). \quad (3)$$

Similar to our approach, by combining the two distance metrics via weighted sum, we have the KL-divergences:

$$S(V, Q) = \alpha D_{KL}(\mathbf{y}||\mathbf{y}_q) + (1 - \alpha) D_{KL}(\mathbf{y}||\mathbf{y}_v), \quad (4)$$

where α is a hyper-parameter that weighs between visual and question scores. Note that $S(V, Q) \geq 0$, because KL-divergence is always non-negative.

(2) **Absolute Difference of KL-Divergence (AD-KLD)** can also be used:

$$S(V, Q) = |D_{KL}(\mathbf{y}||\mathbf{y}_q) - D_{KL}(\mathbf{y}||\mathbf{y}_v)|. \quad (5)$$

Although these methods have strong theoretical backing, we empirically show that for our problem, our proposed method shows the best performance.

Ablation study for other approaches. Here, we introduce variants of our sampling methods. **KL-Divergence (KLD)** is described in Eq. (4). **Jensen-Shannon Divergence only (JSD)** is the baseline that is sampled by only using Jensen-Shannon divergence described as:

$$JSD(\mathbf{y}_v||\mathbf{y}_q) = (D_{KL}(\mathbf{y}_v||M) + D_{KL}(\mathbf{y}_q||M))/2, \quad (6)$$

where $M = (\mathbf{y}_v + \mathbf{y}_q)/2$. **Absolute Difference of KL-Divergence (AD-KLD)** is described in Eq. (5). **Mutual Information (MI)** is the baseline that is sampled by only using mutual information: $S(V, Q) = \alpha I(A; V|Q) + (1 - \alpha) I(A; Q|V)$. **Single-Modal Entropic Measure (SMEM) or (Ours)** is our proposed method:

$$S(V, Q) = \alpha H(Y_q) + (1 - \alpha) H(Y_v) + \beta JSD(\mathbf{y}_v||\mathbf{y}_q) + \gamma H(\hat{Y}). \quad (7)$$

The ablation study results are illustrated in Fig. 1. We compare the AL performance of our model with other design choices for distribution discrepancy measures. We draw the training curves for three different setups: (a) without distillation or entropy, (b) adding distillation loss only, and (c) adding distillation loss and entropy. Each setup shows its strengths in different areas of the AL stages. In (a), ours shows the best performance in the early

stages with less data while **KLD** shows better performance in the later stages as more data is accumulated. In (b), ours shows the best performance with less data while **MI** shows better performance as more data is accumulated. In (c), we show that by taking entropy into account, our model shows the best performance compared to all the other design choices throughout all given stages.

1.3 Qualitative Results

Fig. 2 shows the qualitative results of our model for every AL stage. For each stage, we show the Top-1 predicted answer colored in blue if correct and red if not correct. As the AL stages progress, the model is trained to generate increasingly accurate answers.

References

- [1] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [3] 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.
- [4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [5] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

				
What color are the gym shoes? what color is the flower? Is the catcher wearing safety gear? Is this going to be a feast?				
STAGE 0	brown	red	no	no
STAGE 1	gray	green	no	no
STAGE 2	gray	yellow	no	no
STAGE 3	blue	red	yes	no
STAGE 4	blue	green	yes	no
STAGE 5	black	pink	yes	yes
STAGE 6	black	pink	yes	yes
STAGE 7	black	pink	yes	yes
STAGE 8	black	pink	yes	yes
STAGE 9	white	pink	yes	yes
				
How many women are in the photo? What position does this person play? What color is the room? What color are the walls?				
STAGE 0	1	umpire	green	orange
STAGE 1	1	umpire	blue	orange
STAGE 2	3	catcher	white	orange
STAGE 3	1	batter	white	orange
STAGE 4	3	batter	blue and white	orange
STAGE 5	3	batter	blue and white	orange
STAGE 6	3	batter	green	red
STAGE 7	3	batter	white	orange
STAGE 8	3	batter	white	pink
STAGE 9	3	batter	white	red
				
What are the men sitting on? What is the leafy substance? Are they in a good location to catch fish? What room is this?				
STAGE 0	chairs	yes	no	hotel
STAGE 1	yes	wine	no	hotel
STAGE 2	chair	basil	no	hotel room
STAGE 3	chairs	peppers	no	hotel
STAGE 4	chairs	peppers	no	hotel
STAGE 5	chairs	parsley	no	bed
STAGE 6	chairs	peppers	no	living room
STAGE 7	chair	spinach	no	living room
STAGE 8	chair	basil	no	living room
STAGE 9	bench	basil	yes	bedroom

Figure 2: Qualitative result of the Top-1 prediction of our model for every Active Learning stage.