Single-Modal Entropy based Active Learning for Visual Question Answering

Dong-Jin Kim*
djnjusa@gmail.com
Jae Won Cho*
chojw@kaist.ac.kr
Jinsoo Choi
jinsc37@kaist.ac.kr
Yunjae Jung
yun9298a@gmail.com
In So Kweon
iskweon77@kaist.ac.kr

Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea
(* indicates equal contribution)

Abstract

Constructing a large-scale labeled dataset in the real world, especially for high-level tasks (*e.g.*, Visual Question Answering), can be expensive and time-consuming. In addition, with the ever-growing amounts of data and architecture complexity, Active Learning has become an important aspect of computer vision research. In this work, we address Active Learning in the multi-modal setting of Visual Question Answering (VQA). In light of the multi-modal inputs, image and question, we propose a novel method for effective sample acquisition through the use of ad hoc single-modal branches for each input to leverage its information. Our mutual information based sample acquisition strategy Single-Modal Entropic Measure (SMEM) in addition to our self-distillation technique enables the sample acquisitor to exploit all present modalities and find the most informative samples. Our novel idea is simple to implement, cost-efficient, and readily adaptable to other multi-modal tasks. We confirm our findings on various VQA datasets through state-of-the-art performance by comparing to existing Active Learning baselines.

1 Introduction

Recent successes of learning-based computer vision methods rely heavily on abundant annotated training examples, which may be prohibitively costly to label or impossible to obtain at large scale [10]. In order to mitigate this drawback, Active Learning (AL) [7] has been introduced to minimize the number of expensive labels for the supervised training of deep neural networks by selecting (or *sampling*) a subset of informative samples from a large unlabeled collection of data [47, 55]. AL has been shown to be a potential solution [55] for vision & language tasks like image captioning [50] that rely extensively on large, expensive, curated datasets [2, 15, 34, 50]. Nonetheless, we find through experimentation that existing state-of-the-art AL strategies to these tasks result in a similar performance to each other

without much distinction and without significant improvements when compared to random sampling [11, 41, 48]. A key distinction in vision & language tasks is the presence of *multi-modal* inputs. Previous AL works have focused more on single-modal inputs; hence, studies of muli-modal AL is less prevalent. As multi-modal tasks also suffer the same data issues, we turn our focus to the Visual Question Answering (VQA) task as it provides a simple multi-modal input, single output classification framework.

As studies have shown that VQA models can be reliant on single modalities [3, 14], we define a novel pool-based AL strategy for VQA by leveraging mutual information of the multiple inputs, image and text, individually through a *multi-branch* model. In this multi-branch model, we denote a "main" branch that relies on all modalities. If a single-modal prediction $P(A|V)$ or $P(A|Q)$ is different from the "main" prediction $P(A|V,Q)$, it may signify that the missing modality is informative, where A, V, and Q are the answer output, visual input, and question input respectively. Hence, we propose a method called Single-Modal Entropic Measure (SMEM) that selects instances for labeling from the unlabeled data pool based on the differences among the predictions from the single-modal and "main" predictions. We also show that we can take the differences among the predictions from the single-modal and "main" predictions into account by simply computing the *entropy of the single-modal* predictions. In effect, we propose an uncertainty based sampling paradigm that enables simple inference without any ad hoc steps to measure uncertainty by relying on multi-modal uncertainty. In addition, we use self-distillation, *not as a regularizer*, but to directly aid in sample acquisition by forcing the single-modal branches to generate outputs similar to the "main" branch. If the discrepancy between the single- and multi-modal outputs of a sample are still high despite self-distillation, that sample is highly likely to have large information gain.

Through our experimentation, we show the effectiveness of SMEM on multiple VQA datasets: VQA v2 [7], VizWiz VQA [15], and VQA-CP2 [14]. In addition, although we design our model with multi-modal inputs in mind (*e.g.*, vision-language tasks), we also evaluate our model on NTU RGB+D dataset [49] in order to show the extensibility of our approach. Our solution achieves favorable performance in VQA metrics [7] compared to existing sampling strategies [9, 45, 51, 53, 58, 59], while being cost-efficient. In particular, the proposed method is able to decrease labeling efforts by about 30% on VQA-v2.

Our main contributions are summarized as follows: (1) We propose a novel sampling method, that we call Single-Modal Entropic Measure (SMEM), based on mutual information for Active Learning for VQA. (2) To better sample from the single-modal branches, we employ self-distillation between the main branch and ad hoc single-modal branches to better train the single-modal branches for sampling. (3) Through extensive experimentation, we show the effectiveness of our method on multiple VQA datasets compared to various competing Active Learning methods. (4) We show that our method is task and architecture agnostic by testing our method on the NTU RGB+D action recognition task [49].

2 Related Work

Active Learning. Knowing that labeling data is costly and time consuming, the task is to minimize the number of samples to annotate from a set of unlabeled data while maximizing performance on a given task [55]. AL has been widely studied in image recognition [6, 23, 46, 58], information extraction [9, 22, 45], and text categorization [18, 35], using uncertainty-based sampling [8, 23], information gain [19], or theoretical dropout-based frameworks [13, 24, 38]. Recent works leverage latent space representations using VAEs

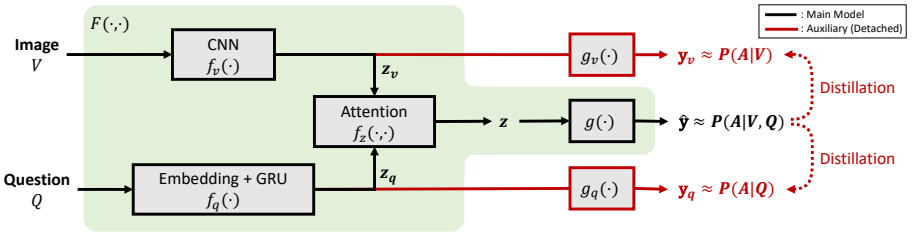

Figure 1: Illustration of our model architecture. Our model can be applied to any off-the-shelf VQA models (area in green) as our goal is to design ad hoc branches solely for the purpose of *measuring uncertainty*. We add two auxiliary branches $g_v(\cdot)$ and $g_q(\cdot)$ that process the visual or question features exclusively to generate answers. Along with the cross-entropy loss from ground truth answers, the single-modal branches are trained with knowledge distillation loss (Eq. (8)) to follow the output from the multi-modal branch $g(\cdot)$.

and adversarial training for improved sampling [53, 56, 59]. We do not compare to methods that require multiple steps such as Monte-Carlo dropout or model ensemble (Query-by-Committee) as they are hard to apply to recent large models for higher-level tasks. We also do not compare with the Core-Set Approach [46] as this method's computational time is directly proportional to the number of classes and the VQA v2 dataset over 3,000 classes, making this difficult to recreate. In addition, recent approaches such as [53, 58, 59] have shown to outperform this approach, so we do not compare this method in our experiments. In addition, AL has been explored in vision & language tasks with varying degrees of success [4, 38, 42, 50]. Although various semi-supervised learning methods for vision & language tasks have been explored [27, 39], active learning has been relatively less explored.

Visual Question Answering. VQA has received a considerable amount of attention due to its real-world applicability through development of several benchmark datasets [14, 15, 21, 34, 51] and state-of-the-art models [12, 32, 40, 57]. Works such as [54] and [42] have explored VQA in a setting similar, however [42] achieves performance improvements by augmenting data with an additional network and [54] focuses on meta learning tasks such as few-shot and zero-shot learning with VQA. [58] explores AL using existing dropout-based method without considering the multi-modality; thus, we do not experimentally compare their work. To the best of our knowledge, we are one of the first to develop an efficient method of leveraging individual modalities to measure uncertainty for AL in VQA.

3 Proposed Method

Single-Modal Entropic Measure (SMEM) consists of two key components: (1) devising an efficient and effective sample acquisition function which takes the mutual information of the individual modalities into account. (2) introducing a novel usage for the self-distillation technique to effectively train the auxiliary branches for more effective sampling.

3.1 Visual Question Answering Baseline

Given an image V and question Q pair, a VQA task tries to correctly predict an answer from the answer set $A \in \mathcal{A}$ (in a vectorized form, $\hat{\mathbf{y}} = F(V,Q) \in [0,1]^{|\mathcal{A}|}$). We adapt one of the

famous VQA state-of-the-art model as our base architecture [1] where $F(\cdot,\cdot)$ is implemented via a combination of deep neural networks. Images and questions are mapped to feature vectors through convolutional ($f_v(\cdot)$) and recurrent ($f_q(\cdot)$) neural networks respectively (*i.e.*, $\mathbf{z}_v = f_v(V), \mathbf{z}_q = f_q(Q)$). Then following [1], $\mathbf{z}_v$ and $\mathbf{z}_q$ are combined via an additional sub-network $f_z(\cdot,\cdot)$ which uses attention to generate joint representation (*i.e.*, $\mathbf{z} = f_z(\mathbf{z}_v, \mathbf{z}_q)$). Finally, $\mathbf{z}$ is fed through a fully-connected layer $g(\cdot)$ to predict $\hat{\mathbf{y}}$ (*i.e.*, $\hat{\mathbf{y}} = g(\mathbf{z})$). The goal for training the VQA model is to follow the conditional distribution of the target dataset, $\hat{\mathbf{y}} = F(V,Q) \approx P(A|V,Q)$, where $F(\cdot,\cdot)$ is the VQA model, and $P(\cdot)$ is the probability of the answer A given image V and question Q that we want $F(\cdot,\cdot)$ to follow. Thus, the model is trained with Binary Cross-Entropy loss compared with ground truth $\mathbf{y}$, $\mathcal{L}_{main} = BCE(\mathbf{y}, \hat{\mathbf{y}})$.

3.2 Active Learning with SMEM

The goal of AL is to find the best sample acquisition function $S(V,Q)$ that assigns a high score to *informative* samples. In the AL setup, a labeled set $\mathcal{D}_\mathrm{L}$ is used to train a model, then after the model converges (end of a *stage*), the model is used to sample from the unlabeled set $\mathcal{D}_\mathrm{U}$, then the chosen samples are given labels and added into the labeled set $\mathcal{D}_\mathrm{L}$ and removed from the unlabeled set $\mathcal{D}_\mathrm{U}$. This step is repeated several times depending on the number of stages. There are several ways to define the informativeness of a sample. For example, entropy defined as $H(\hat{Y}) = -\sum_i \hat{\mathbf{y}}^i \log \hat{\mathbf{y}}^i$, defines uncertainty as informativeness. However, in this paper, we define a novel acquisition function that takes *mutual information* into account. Our approach exploits the multi-modality of this task and directly measures the mutual information $I(A;V|Q)$ and $I(A;Q|V)$ individually to aid the sampling criteria.

Intuitively, if the value of mutual information, for example $I(A;V|Q)$, is high, the variable V plays an important role in predicting the answer A. In order to compute mutual information of individual modalities, we add auxiliary classification branches, as shown in Fig. 1, that predict the answer exclusively from individual visual and question features similar to [6, 52]. The branches are detached from the model and only utilized to "sample" the unlabeled data instead of training the model; thus, the *presence of the branches have no effect on the main model performance* and *does not act as a regularizer* of any kind. The auxiliary branches are also trained with BCE loss so that the output answer probability can approximate the conditional distribution of answers given cues such that $\mathbf{y}_v = g_v(\mathbf{z}_v) \approx P(A|V)$ and $\mathbf{y}_q = g_q(\mathbf{z}_q) \approx P(A|Q)$. Note however that the answer predictions from the single-modal classifiers do not have to be accurate as their purpose is only to measure uncertainty. As we try to measure the informativeness of a given input visual question pair (V,Q) through a given acquisition function $S(V,Q)$, we can intuitively expect that the input cue V or Q may contain a considerable amount of information if the single-modal outputs $\mathbf{y}_v$ and $\mathbf{y}_q$ are considerably different from $\hat{\mathbf{y}}$. Here, we devise a novel acquisition function based on entropy that takes into account the relation between the single-modal and multi-modal answer representations.

Single-Modal Entropic Measure (SMEM). As our final goal is to find the most informative samples, one of our objectives is to find the training data points that have high mutual information $I(A;V|Q)$ and $I(A;Q|V)$. According to the property of mutual information, it can be decomposed into the difference of entropies:

$$I(A;V|Q) = H(A|Q) - H(A|V,Q) \approx H(Y_q) - H(\hat{Y}), \tag{1}$$

$$I(A;Q|V) = H(A|V) - H(A|V,Q) \approx H(Y_v) - H(\hat{Y}), \tag{2}$$

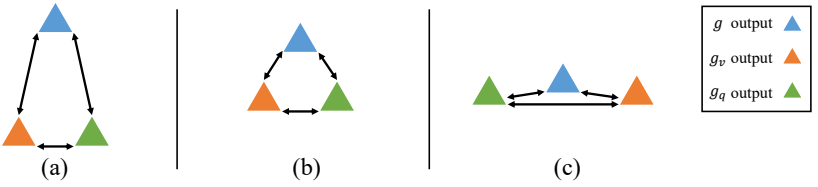

Figure 2: Simple illustration of the $\mathbf{y}_v$, $\mathbf{y}_v$, and $\mathbf{y}_q$. SMEM gives a higher score to (a) than (b) in the sampling stage. Also, since we can intuitively expect (c) to be more informative than (b) when labeled, we propose to add Jensen-Shannon divergence between the two single-modal outputs in order to account for the discrepancies between single-modal outputs.

where $H(\hat{Y}) = -\sum_i \hat{\mathbf{y}}^i \log \hat{\mathbf{y}}^i$ is the entropy of a given distribution. As shown in the equations above, we see that mutual information is easily computed from the entropy values of the single-modal and multi-modal outputs. This is from the assumption that $H(A|V,Q)$, $H(A|Q)$, and $H(A|V)$ are modeled with $H(\hat{Y})$, $H(Y_q)$, and $H(Y_v)$ respectively (*i.e.*, $H(A|V,Q) \approx H(\hat{Y})$, $H(A|Q) \approx H(Y_q)$, $H(A|V) \approx H(Y_v)$). Similarly, rearranging Eq. (1) and Eq. (2), we get:

$$H(Y_q) \approx I(A;V|Q) + H(\hat{Y}), \tag{3}$$

$$H(Y_v) \approx I(A;Q|V) + H(\hat{Y}), \tag{4}$$

where $H(Y_q) = -\sum_i \mathbf{y}_q^i \log \mathbf{y}_q^i$ and $H(Y_v) = -\sum_i \mathbf{y}_v^i \log \mathbf{y}_v^i$. From Eq. (3) and Eq. (4), we can see that the entropy of the single-modal output is the sum of the mutual information and the entropy of the multi-modal output or "main" entropy (that we call *main entropy* from here on out to avoid confusion), and we want both values to be high when sampling. In other words, both uncertainty and informativeness can be measured by simply computing the *entropy of the single-modal output*. The empirical advantage of leveraging the single-modal entropy instead of only using the mutual information is shown in the supplementary material.

In this regard, we propose an informativeness value to be the weighted sum of the single-modal entropies for image and question:

$$S(V,Q) = \alpha H(Y_q) + (1-\alpha) H(Y_v), \tag{5}$$

where $0 \leq \alpha \leq 1$ is a hyper-parameter that weights the relative importance of visual and question scores. Note that $0 \leq S(V,Q) \leq \log|\mathcal{A}|$, because $0 \leq H(\cdot) \leq \log|\mathcal{A}|$.

In addition, we also conjecture that as the information of the relationships between all single- and multi-modal outputs are important, the relationship between the single-modal outputs is just as important. As in Fig. 2 (b) and (c), even though the distance between single-modal outputs and the multi-modal output is fixed, the relationship between the three distributions can differ due to the distance between single-modal outputs. If we only leverage the distance between the single-modal and multi-modal outputs, (b) and (c) would give the same score, however, we can intuitively expect that (c) is more uncertain than (b) as the distance between the single-modal outputs is much greater. In order to leverage this, we also include **Jensen-Shannon divergence** (JSD) as an additional sampling criterion:

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

where $M = (\mathbf{y}_v + \mathbf{y}_q)/2$. The addition of JSD in the sampling criterion forces the sampler to find samples with a large divergence between single-modal outputs, ultimately increasing the maximum distance among all three points. Therefore, the final function of **SMEM**:

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

where we introduce an additional weight hyper-parameter $\beta \geq 0$. The introduction and comparison with other possible approaches, including pure mutual information or distance between single-modal and multi-modal branches, can be found in the supplementary material. We empirically determine the hyper-parameters via extensive greed search. We also empirically found that adding the main entropy further improved the performance: $S(V,Q) + \gamma H(\hat{Y})$, where $\gamma \geq 0$ is the hyper-parameter that weights between $S(V,Q)$ and the main entropy. Note that $S(V,Q) + \gamma H(\hat{Y}) \geq 0$, as $S(V,Q)$ and $H(\hat{Y})$ are always non-negative.

3.3 Training Auxiliary Branches with Self-Distillation

Eq. (3) and Eq. (4) are set on the basis that the single-modal classifiers' output distribution will be similar to that of the multi-modal classifier's especially if the visual input is conditionally independent on answers given a question (*i.e.*, $P(A|VQ) \approx P(A|Q)$) or vice versa (*i.e.*, $P(A|VQ) \approx P(A|V)$). In order to satisfy this assumption, $g_v(\cdot)$ and $g_q(\cdot)$'s outputs distribution should be trained to resemble $g(\cdot)$'s output distribution as closely as possible. To do so, we additionally utilize the self-distillation technique to explicitly train the auxiliary single-modal classifiers to mimic the "main" classifier's distribution [44, 60]. Note that as the single-modal classifiers are detached, this self-distillation technique has no effect on the performance of the model and *does not act as a regularizer*. Also they are *not used to for any predictions*, hence having no effect on the performance as well. Our insight is, our acquisition function $S(V,Q)$ tries to find the sample that has a large difference between the multi-modal output and the single-modal output, and the knowledge distillation loss tries to directly minimize this difference. In other words, if a sample gives a high $S(V,Q)$ value even after the explicit minimization of the $S(V,Q)$ score meaning the single-modal classifiers have difficulty following the output distribution of the multi-modal classifier, the sample can be thought of as a sample that is difficult or affected by both modalities.

Formally, Knowledge Distillation [16] is to train a "student" (in our case, $\mathbf{y}_v$ and $\mathbf{y}_q$) to follow a "teacher" model or representation (in our case, $\hat{\mathbf{y}}$), which has been utilized for various proposes [5, 17, 25, 28, 30, 57]. To this end, we add the self-distillation loss in addition to the classification loss in training $g_v(\cdot)$ and $g_q(\cdot)$ to improve our acquisition function directly. The auxiliary single-modal branches are trained with the following loss functions with weight hyper-parameter $\lambda \geq 0$:

$$\mathcal{L}_v = BCE(\mathbf{y},\mathbf{y}_v) + \lambda BCE(\hat{\mathbf{y}},\mathbf{y}_v), \quad \mathcal{L}_q = BCE(\mathbf{y},\mathbf{y}_q) + \lambda BCE(\hat{\mathbf{y}},\mathbf{y}_q), \tag{8}$$

4 Experiments

In this section, we describe the experimental setups, competing methods, and provide performance evaluations of the proposed methods.

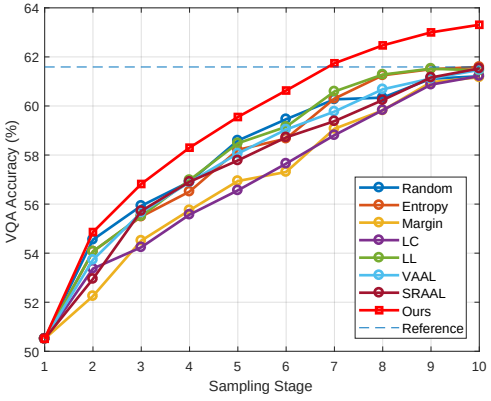

Figure 3: Comparison with existing approaches on the VQA v2 dataset. Reference is the best performance of the existing methods. Our model shows the best performance among all the other methods by a large margin. Full dataset training at 440K samples shows 63.52% accuracy, while ours at *stage 10* at 400K samples shows 63.31% accuracy.

4.1 Experiments on VQA v2 Dataset

Dataset and metric. We evaluate our method on VQA v2 dataset [2], which consists of 1.1M data points (visual-question-answer triplets), with 444K/214K/448K for train/val/test respectively. Each image has at least 3 questions and 10 answers per question, thus contains 204K images, 1.1M questions, 11.1M answers, and we utilize the VQA accuracy metric [2].

Active Learning setup. For AL, labeled set $\mathcal{D}_{\mathrm{L}}$ with a size of 40,000 out of 443,757 training data points, roughly 10% of the full dataset, is initialized randomly, and our model is trained until the training loss converges. After converging, the end of the first stage, we use this model to collect samples from the unlabeled dataset $\mathcal{D}_{\mathrm{U}}$ according to an acquisition function $S(V,Q)$. For every stage, we sample the next 40,000 data points from the unlabeled dataset and add them to the labeled dataset. After increasing the labeled set, we train a newly initialized model using the resulting labeled set, and repeat the previous steps (marking the subsequent stages), increasing the dataset and training models for 10 iterations, resulting in a final $\mathcal{D}_{\mathrm{L}}$ size of 400,000 data points.

Baseline approaches. In order to demonstrate the effectiveness of our method, we compare the performance of our model to widely used sample acquisition approaches [9, 45, 51] as well as the recent deep AL methods [53, 58, 59]. However, we do not compare to ensemble based methods such as BALD, MC-dropout as they are inefficient and show performance worse than current state-of-the-arts such as VAAL or SRAAL. We also do not compare to Core-Set Approach [46] as the computation cost is directly correlated to the class number, which VQA v2 has a class number of over 3,000 and they show worse performance than current state-of-the-arts such as LL4AL, VAAL, or SRAAL. Our comparison baselines are Random (which is a uniform random distribution, sometimes called passive learning), Entropy [51] (which is the same as *main entropy* from Sec. 3.2), Margin [45], Least Confident (LC) [9], LL [58], VAAL [53], and SRAAL [59]. For VAAL and SRAAL, as our inputs are multi-modal, we use a joint representation for the Variational Auto-Encoder [33].

Comparison with existing approaches. In Fig. 3, among the existing methods, ***Margin*** and ***Least Confident(LC)*** show similarly poor performance, while ***Random*** and ***Entropy*** show

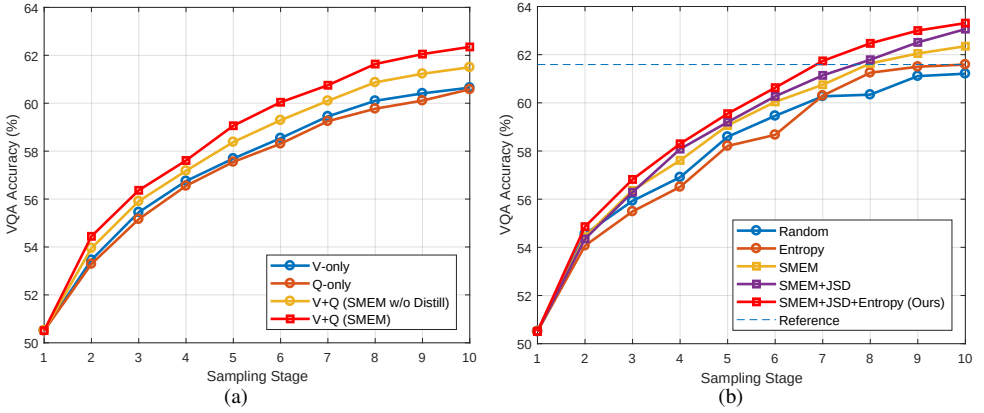

Figure 4: Ablation study of our method. (a) Using both V and Q modalities along with distillation (SMEM) shows the best performance in all the Active Learning stages. (b) SMEM already shows favorable performance against traditional Active Learning methods, Random and Entropy. Moreover, Our final method (SMEM+JSD+Entropy) shows the best performance among the variants of our methods in all the Active Learning stages. Note that as Entropy here is the same as *main entropy*, the Entropy curve is identical to that in Fig. 3.

better performance in the early stage and the late stage respectively. We conjecture that the poor performance of ***Margin*** and ***LC*** is due to the large number of classes. VQA v2 has 3,129 candidate classes, causing a long-tailed distribution problem, and makes conventional AL approaches difficult to select informative samples. In particular, the model predictions might be dominated to major classes; regarding only the most confident class (***LC***) or second (***Margin***) might lead to biased acquisition to major classes. Note that recent state-of-the-art Active Learning methods, ***LL***, ***VAAL***, ***VRAAL***, show similar performance to ***Random*** and ***Entropy*** in the VQA task. Moreover, our proposed method shows the best performance compared to all the counterparts by a large margin in all stages. Note that the best among the existing approaches achieved 61.59% accuracy when using 400K training samples, which is similar performance of our model trained with 280K training samples, signifying that SMEM can save about 30% of the human labeling effort compared to existing approaches.

Ablation study. We also perform an ablation study on each of the components of our final method. ***+Distillation*** is for the proposed self-distillation loss in Eq. (8). ***+Jensen-Shannon Divergence (+JSD)*** is the additional JSD term which changes our acquisition function from Eq. (5) to Eq. (7). ***+Entropy*** indicates the additional entropy term of the "main" branch output $H(\hat{Y})$. In Fig. 4 (b), we include traditional AL approaches, ***Random*** and ***Entropy*** for comparison. In Fig. 4 (a), compared to the baselines with only single-modal entropy ($\alpha = 0$ for V-only and $\alpha = 1$ for Q-only in Eq. (5)), using both V and Q modalities (Eq. (5)) along with ***+Distillation*** shows the best performance in all the AL stages. In Fig. 4 (b), SMEM without any auxiliary terms already shows favorable performance against traditional Active Learning methods, ***Random*** and ***Entropy***. Moreover, ***+JSD*** noticeably improves the performance of SMEM, and our final method with both ***+JSD*** and ***+Entropy*** shows the best performance among all the variants of our methods in all AL stages.

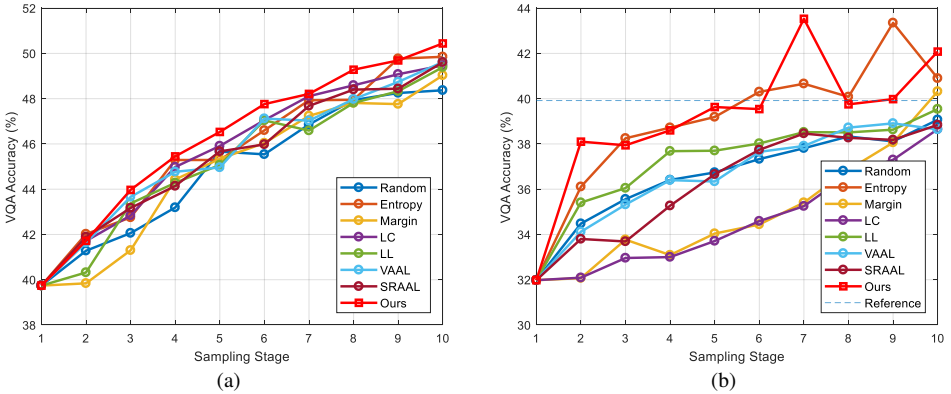

Figure 5: Comparison with existing approaches on the (a) VizWiz and (b) VQA-CP2 datasets. Our proposed method shows favorable performance on both datasets. The Reference line for (b) denotes the performance of the VQA model with the full 100% of the dataset at 39.91% accuracy.

4.2 Experiments on Additional VQA Datasets: VizWiz and VQA-CP2

To demonstrate the extensibility of our method, we extend our method on other VQA datasets. **VizWiz VQA** [15], an arguably more difficult dataset, contains 19,425 train image/question pairs. We sample 1,900 image/question pairs per stage with the final stage ending at 19,000 image/question pairs. For architecture, we use a publicly available code of [57] , and show that our method is model agnostic. We compare to the same baselines as in the previous subsection, and experimental results are shown in Fig. 5 (a). As VizWiz is a more challenging dataset than VQA v2 dataset, all the baselines show similar performance. Even the recent state-of-the-art such as ***VAAL*** and ***SRAAL*** show performance similar to that of all other baselines with no baseline being clearly above another. Among all other baselines, our proposed method shows the best performance in most of the sampling stages.

VQA-CP2 [14] is a re-ordering of VQA v2 and the dataset size almost the same, containing 438K train image/question pairs. We sample 40K image/question pairs per stage with the final stage ending at 400K image/question pairs. We use the same network architecture as VQA v2 and compare to the same baselines. The experimental results are shown in Fig. 5(b). Among the existing approaches, ***Entropy*** shows the best performance, and the recent state-of-the-art methods of ***VAAL*** and ***SRAAL*** show mediocre performance, performing worse than ***Random*** in several stages. Note that when tested on all 438K training data, the model performance is 39.91% (denoted as Reference in Fig. 5(b)) which is lower than the performance of ***Entropy*** and ***Ours*** *trained with the subset of the training data*. We find that by elaborately sampling some meaningful data, we are able to surpass the performance of the full dataset. As the semi-supervised learning for bias data started to be explored [31, 43], we believe this finding also opens up interesting possible future research ideas.

4.3 Experiments on NTU Action Recognition Dataset

Although we designed our model specifically for multi-modal task of VQA, we also evaluate our model on one of the popular multi-modal action recognition datasets, NTU RGB+D

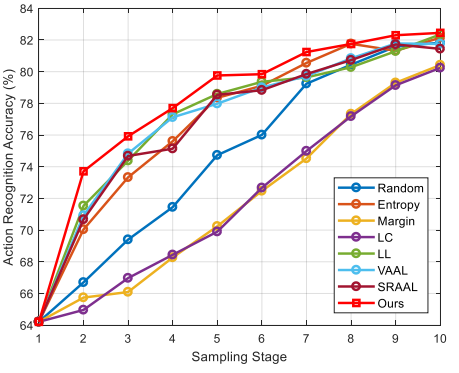

Figure 6: Comparison with existing approaches on the NTU RGB+D Action Recognition Dataset. Although our method is designed specifically for VQA task, our method consistently outperforms all the existing approaches in every sampling stage.

dataset [49], to show the extensibility of our approach. We use the model from [36] with the NTU RGB+D dataset [49] which leverages multiple views or subjects with the same action class and trains the network. The dataset processing used in the implementation of [36] makes use of 37,646 pairs of inputs and we sample 3,000 at each stage, ending at 30,000 pairs at the last stage. We use cross-subject evaluation [49] that splits the 40 subjects into training and testing sets. We apply our method on the subject dataset and treat each subject as a single modality and train additional classifiers similar to the VQA settings.

We use the same baselines as the previous subsection, and show experimental results in Fig. 6. Note that the gap between ***Margin*** or ***LC*** and the other baselines is larger than that of the previous experiments. Among ***Random***, ***Entropy***, ***LL***, ***VAAL***, and ***SRAAL***, ***LL*** and ***VAAL*** shows the best performance in the early stages, and ***Entropy*** shows the best performance in the later stages. Moreover, compared to baseline approaches, SMEM shows favorable performance for all sampling stages in the NTU RGB+D dataset as well. The effectiveness of SMEM opens up the possibilities of applying our method on other datasets and tasks where multiple inputs maybe used whether they are of the same modality of not.

5 Conclusion

In this work, we introduced a novel Active Learning framework specifically tailored for the multi-modal task of Visual Question Answering. We showed through our empirical evidence the performance gains of our novel method in the Active Learning setup compared to prior state-of-the-art Active Learning methods. We also empirically showed that our method is extendable to other multi-modal applications such as Action Recognition with ease and promise, which opens a window for further research in other tasks. In addition, we also show through our testing on the VQA-CP2 dataset that by using our method to sample a subset of the data surpasses using the full dataset, showing that a higher performance can be achieved with less data. We believe this shows the potential possibilities of our research and shed new light on this field and other multi-modal research fields. We hope to explore Active Learning for other multi-modal tasks [8, 20, 26, 29] and encourage future researchers to study Active Learning for other multi-modal tasks in the future as well.

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

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] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. In *Advances in Neural Information Processing Systems (NIPS)*, 2019.

[4] Ta-Chung Chi, Mihail Eric, Seokhwan Kim, Minmin Shen, and Dilek Hakkani-tur. Just ask:an interactive learning framework for vision and language navigation, 2019.

[5] Jae Won Cho, Dong-Jin Kim, Jinsoo Choi, Yunjae Jung, and In So Kweon. Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] Jae Won Cho, Dong-Jin Kim, Yunjae Jung, and In So Kweon. Mcdal: Maximum classifier discrepancy for active learning. *arXiv preprint arXiv:2107.11049*, 2021.

[7] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[8] Brendan Collins, Jia Deng, Kai Li, and Li Fei-Fei. Towards scalable dataset construction: An active learning approach. In *European Conference on Computer Vision (ECCV)*, 2008.

[9] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2005.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[11] Rosa L Figueroa, Qing Zeng-Treitler, Long H Ngo, Sergey Goryachev, and Eduardo P Wiechmann. Active learning for clinical text classification: is it better than random sampling? *Journal of the American Medical Informatics Association*, 19(5):809–816, 2012.

[12] 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.

[13] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning (ICML)*, 2017.

[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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[17] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Batch mode active learning and its application to medical image classification. In *International Conference on Machine Learning (ICML)*, 2006.

[19] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[20] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Rosie Jones, Rayid Ghani, Tom Mitchell, and Ellen Riloff. Active learning for information extraction with multiple view feature sets. *Proc. of Adaptive Text Extraction and Mining, EMCL/PKDD-03, Cavtat-Dubrovnik, Croatia*, 2003.

[23] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[24] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[25] 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)*, 2018.

[26] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] 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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[28] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision (ECCV)*, 2020.

[29] Dong-Jin Kim, Tae-Hyun Oh, Jinsoo Choi, and In So Kweon. Dense relational image captioning via multi-task triple-stream networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.

[30] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Acp++: Action co-occurrence priors for human-object interaction detection. *IEEE Transactions on Image Processing (TIP)*, 2021.

[31] Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2020.

[32] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[33] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017.

[35] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, 1994.

[36] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[37] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision (ECCV)*, 2016.

[38] Xiao Lin and Devi Parikh. Active learning for visual question answering: An empirical study, 2017.

[39] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *European Conference on Computer Vision (ECCV)*, 2018.

[40] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[41] François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. Phrase-based statistical language generation using graphical models and active learning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.

[42] Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, and Laurens Van Der Maaten. Learning by asking questions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[43] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. *arXiv preprint arXiv:2106.05682*, 2021.

[44] Mary Phuong and Christoph H Lampert. Distillation-based training for multi-exit architectures. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[45] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, 2001.

[46] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[47] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[48] Burr Settles, Mark Craven, and Lewis Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, 2008.

[49] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[50] Tingke Shen, Amlan Kar, and Sanja Fidler. Learning to caption images through a lifetime by asking questions. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[51] Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*, 2017.

[52] Inkyu Shin, Dong-Jin Kim, Jae Won Cho, Sanghyun Woo, KwanYong Park, and In So Kweon. Labor: Labeling only if required for domain adaptive semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[53] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[54] Damien Teney and Anton van den Hengel. Visual question answering as a meta learning task. In *European Conference on Computer Vision (ECCV)*, 2018.

[55] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research (JMLR)*, 2(Nov): 45–66, 2001.

[56] Shuo Wang, Yuexiang Li, Kai Ma, Ruhui Ma, Haibing Guan, and Yefeng Zheng. Dual adversarial network for deep active learning. In *European Conference on Computer Vision (ECCV)*, 2020.

[57] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[58] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[59] Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[60] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[61] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.