UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

EaSe: A Diagnostic Tool for VQA Based on Answer Diversity

Jolly, S.; Pezzelle, S.; Nabi, M.

DOI
10.18653/v1/2021.naacl-main.192

Publication date
2021

Document Version
Final published version

Published in
The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

License
CC BY

Link to publication

Citation for published version (APA):
Jolly, S., Pezzelle, S., & Nabi, M. (2021). EaSe: A Diagnostic Tool for VQA Based on Answer Diversity. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: NAACL-HLT 2021 : proceedings of the conference : June 6-11, 2021* (pp. 2407-2414). The Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.192

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Oct 2026

EaSe: A Diagnostic Tool for VQA Based on Answer Diversity

Shailza Jolly
TU Kaiserslautern, Germany
DFKI GmbH, Germany
shailza.jolly@dfki.de

Sandro Pezzelle
ILLC, University of Amsterdam, Netherlands
s.pezzelle@uva.nl

Moin Nabi
SAP AI Research, Germany
m.nabi@sap.com

Abstract

We propose EASE, a simple diagnostic tool for Visual Question Answering (VQA) which quantifies the difficulty of an $image, question$ sample. EASE is based on the pattern of answers provided by multiple annotators to a given question. In particular, it considers two aspects of the answers: (i) their Entropy; (ii) their Semantic content. First, we prove the validity of our diagnostic to identify samples that are easy/hard for state-of-art VQA models. Second, we show that EASE can be successfully used to select the most-informative samples for training/fine-tuning. Crucially, only information that is readily available in any VQA dataset is used to compute its scores.[1]

1 Introduction

Visual Question Answering (VQA; Antol et al., 2015) requires models to jointly understand an image and a natural language question. This is a challenging task; despite massive training data and recent pre-training strategies (Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2020) models still struggle to close the gap with oracle performance.

VQA datasets (e.g., Goyal et al., 2017; Gurari et al., 2018) consist of $\langle image, question \rangle$ pairs for which N human annotators have provided an answer in natural language. When trained on these samples, VQA models are fed with the most frequently chosen answer in the pattern. During inference, the answer with the highest probability is evaluated against the pattern of N ground-truth answers. According to the standard VQA metric (Antol et al., 2015), a model's prediction is considered as perfectly correct if it matches an answer that was frequent in the pattern; less accurate if matching an underrepresented one. This metric implies that, for the majority of cases, several annotators agree on the same exact answer—and a model can thus achieve 100% accuracy in the task. On the other hand, this suggests that various $\langle image, question \rangle$ pairs can have different patterns of answers; i.e., they can be more or less scattered depending on the features of the question, the image, or both. In Fig. 1, the annotators did not converge on the same answer for either of the two questions. However, while in the top question the 10 annotators provided semantically similar answers (e.g., *plaid*, *plaid and floral*, etc.), in the bottom one very different answers were given (e.g., *road*, *sweden*).

Q: What is the pattern of the little girl's dress?
GT: plaid: 4, checks and flowers: 1, checkered with flowers: 1, polka dots, squares, plaid: 1, squares and flowers: 1, flowers: 1, plaid and floral: 1
EaSe: 1.0

Q: Where is this?
GT: road: 4, outside: 2, pakistan: 1, outdoors: 1, sidewalk: 1, sweden: 1
EaSe: 0.30

Figure 1: One image from VQA2.0 with two questions and the answers by 10 annotators. Frequency of each unique answer (e.g., $plaid : 4$) and EASE values of the samples (the higher, the easier) are reported.

In line with recent work aimed at predicting the agreement between annotators (Gurari and Grauman, 2017), the distribution of answers for a given $\langle image, question \rangle$ pair (Yang et al., 2018), or the difficulty of visual questions (Terao et al., 2020), in this paper we introduce EASE, a diagnostic tool for VQA which is based on the answers provided to a given question. We propose that two main features of the answer pattern, **E**ntropy **a**nd **Se**mantic content, are informative of the degree of difficulty of a sample. In particular, we conjecture that the more scattered an answer pattern, the more difficult the sample (Fig. 1, down)—unless some or all of those answers are semantically similar (Fig. 1, top).

[1] Code at: github.com/shailzajolly/EaSe

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2407–2414
June 6–11, 2021. ©2021 Association for Computational Linguistics

By experimenting with various VQA datasets and models, we first assess the effectiveness of our diagnostic to identify the samples that are easy/difficult for a model. Second, we use EASE to select increasingly difficult subsets of data that we use to train/fine-tune our models, based on the hypothesis that difficult cases are also more informative during training. In both cases, we show that our simple method is very effective: (1) models are shown to struggle with the most difficult samples according to it; (2) training/fine-tuning models with only a little fraction of samples—the most difficult ones—makes them achieve very high results, which are comparable to models trained/fine-tuned with the whole training data. Finally, EASE is shown to correlate with the confidence scores provided by human annotators along with their answers, which reveals that it captures a notion of difficulty in line with that by human speakers.

2 Approach

We focus on $\langle image, question \rangle$ VQA samples and aim to quantify their *difficulty*, i.e., how challenging it is for a model to answer them correctly. We propose that the difficulty of a sample can be quantified based on the (readily available) characteristics of the pattern of answers provided by the annotators, and devise a *diagnostic* tool that builds on this assumption. In particular, we focus on two aspects of the pattern: 1) its **Entropy**, i.e., how scattered it is in terms of the number of unique answer strings; 2) its **Semantics**, i.e., how (dis)similar are the answers in it with respect to their overall semantic representation. We name our diagnostic tool **EASE** and describe it in detail below.

Entropy (E) We consider all the answers provided by the annotators for a given sample. Similar to Yang et al. (2018), we measure the Entropy of a pattern using Eq. 1:

$$E(p_f) = \frac{-1}{\eta} \sum_{k=1}^{M} p_k * log(p_k) \qquad (1)$$

where p_f is the distribution of the M unique answers based on their frequency, and η is the highest possible Entropy value[2] that is used to normalize E in $[0, 1]$. High E values (close to 1) are assigned to highly scattered distributions; *vice versa*, low values of E (close to 0) are assigned to highly consistent distributions, e.g., when all annotators agree on the same answer.

Semantics (SE) E is based on the frequency of unique answer strings in a given pattern. As such, it treats various strings as different, regardless of whether strings are semantically similar. This, however, is crucial: answers to a given question that are semantically different reveal inconsistencies among annotators, which in turn is indicative of the difficulty of a sample. In contrast, semantically similar answers are a proxy for the ease of the sample, though these answers are different in their surface realization (see, e.g., *a couple* vs. *a pair*).

We use a simple method based on pre-trained word embeddings (Mikolov et al., 2018) to operationalize SE. In particular, given a pattern of answers, we perform the following steps to reorganize it by aggregating semantically similar answers and their corresponding frequencies: (1) We compute a representation of each answer in the pattern by averaging its words embeddings, similar to Chao et al. (2018); (2) We build an answer's centroid, i.e., an average representation of all the unique answers that encodes the overall semantics in the pattern; (3) We compute the pairwise cosine similarity (cos) between the centroid and each unique answer in the pattern (negative values are clamped to 0 to have similarity in $[0, 1]$); (4) We group together all the answers whose cos with the centroid embedding exceeds a certain threshold. The threshold τ is dynamically set. It is computed at the *datum*-level to adapt to the features of each datapoint, and is defined by:

$$\tau = cos\ (\text{MAX},\ centroid) - \varepsilon \qquad (2)$$

where ε is a small positive number close to 0 (here we experiment with $\varepsilon = 0.0001$), and MAX is the answer with the maximum frequency in the pattern. In case more than one MAX is present, the lowest τ is used. Finally, we obtain a new distribution where the answers that are semantically consistent with the pattern's overall content (the centroid) are put together, and their frequencies are summed up.

EASE diagnostic We take the new distribution of answers after applying SE, p_{se}, and compute EASE, a single value in $[0, 1]$ which quantifies the ease of a VQA sample. We obtain it as follows:

$$\text{EASE}(p_{se}) = 1 - E(p_{se}) \qquad (3)$$

[2] In our data, the maximum Entropy value is equal to 2.302.

Method	Split	VQA2.0		VizWiz	
		T	V	T	V
EaSe	TH	40522 (9%)	19805 (9%)	3201 (16%)	522 (16%)
	BH	189281 (43%)	92606 (43%)	10443 (52%)	1646 (52%)
	E	213954 (48%)	101943 (48%)	6356 (32%)	1005 (32%)
Entropy	TH	108457 (25%)	53230 (25%)	11903 (60%)	1897 (60%)
	BH	187287 (42%)	90896 (42%)	7337 (36%)	1165 (37%)
	E	148013 (33%)	70228 (33%)	760 (4%)	111 (3%)
	Total	443757 (100%)	214354 (100%)	20000 (100%)	3173 (100%)

Table 1: Top: Number of samples in the TH, BH, and E splits of VQA2.0 and VizWiz based on EaSe. Bottom: number of samples based on Entropy. In brackets: percentage in the corresponding T(rain)/V(al) partition.

Dataset / Split		BUTD	LXM	LXM-S
VQA2.0	*all*	63.43	71.48	63.18
	TH	29.82	36.56	30.52
	BH	63.97	71.26	64.06
	E	69.47	78.46	68.73
VW	*all*	50.35	53.75	45.79
	TH	29.48	31.84	26.61
	BH	49.08	52.82	44.38
	E	63.27	66.65	58.08

Table 2: Accuracy by BUTD, LXM, and LXM-S on the entire validation set (*all*) of VQA2.0 and VizWiz (VW) and the 3 splits defined by EaSe. For all models in both datasets, accuracy consistently increases from TH to E.

where the second term quantifies the Entropy of p_{se} (see Eq. 1), and the first term is introduced to make EASE values increase with the ease of a sample.

3 Method

3.1 Models

We experiment with two models: BUTD (Anderson et al., 2018) and LXMERT (Tan and Bansal, 2019) (LXM). **BUTD** uses a GRU to encode the input questions and to attend the image RoI features, enabling region-based attention to generate the answer. **LXM** is a transformer-based architecture pretrained on several language and vision tasks. We use it with the default hyper-parameters set in the original implementation. The models are trained (BUTD) or fine-tuned (LXM), and then evaluated, on the datasets described below.

3.2 Datasets

We experiment with VQA2.0 (Goyal et al., 2017) and VizWiz (VW; Gurari et al., 2018). We choose these two datasets since they are very different from each other, both in terms of the images (object-centered vs. everyday-life) and the type and purpose of the questions (written, crowdsourced vs. spoken, goal-oriented) they contain. This fundamental diversity is confirmed by a preliminary analysis[3] on the answers to the questions contained in the validation split. In VQA2.0, 33% of the questions are assigned the same answer string by all annotators; as for VizWiz, this percentage drops to only 3%. We take this low agreement as a proxy for the difficulty of the samples in this (and any) dataset: the more disegreement, the harder.

3.3 Proof-of-Concept Analysis

To preliminarly test our hypothesis, we compute the EASE value for each sample in the train/val partitions of the two datasets and assign the samples into 3 splits based on their EASE value (number of samples per split in Tab. 1, top): (1) **EASY (E)**: EASE $= 1.0$; (2) **BOTTOM-HARD (BH)**: $0.5 <=$ EASE < 1.0; (3) **TOP-HARD (TH)**: EASE < 0.5. We then test our trained models on each of our validation splits. If our hypothesis is correct, models should struggle with the harder splits selected by our tool. Tab. 2 shows that all models—BUTD, LXM and LXM-S, a version of LXM trained from scratch on the task—indeed achieve much lower performance on the hard splits; in TH, their accuracy is halved compared to the entire (*all*) data. Moreover, it is interesting to note that, for LXM, pretraining appears to be overall beneficial, with the pretrained version outperforming the non-pretrained one in both datasets and all splits, with a margin of around 8 points on the entire data.

For comparison, we run the same analysis using Entropy (specifically, $1 - Entropy$) instead of EASE. As can be seen in Table 1 (bottom), the two methods give rise to very different data dis-

[3]Further details in Appendix B. See also Jolly et al. (2018).

Model	TD	VQA2.0				VizWiz			
		all	TH	BH	E	*all*	TH	BH	E
BUTD	TH(R)*	50.14	20.46	53.34	53.0	42.75	24.91	40.57	55.58
BUTD	TH	44.13	26.1	51.3	41.13	42.46	25.1	39.69	56.02
BUTD	TH+BH	56.6	29.73	61.2	57.64	48.58	29.58	47.57	60.1
BUTD	TH+BH+E	61.43	29.61	62.81	66.36	50.12	29.56	48.95	62.73
LXM	TH(R)*	69.61	34.76	69.44	76.55	46.42	26.03	45.78	58.06
LXM	TH	67.24	35.64	67.58	73.02	46.65	26.13	45.79	58.73
LXM	TH+BH	69.85	37.05	70.63	75.52	51.65	30.29	50.07	65.36
LXM	TH+BH+E	70.57	35.51	70.26	77.65	53.40	32.82	52.26	65.97

Table 3: Accuracy on each split of VQA2.0 and VizWiz obtained by gradually training models first on TH, then adding BH and finally adding E samples. TD refers to type of training data used for training. TH(R) refers to the setting in which we use a split randomly sampled from the training data with the same size of TH. *The random sampling was performed 10 times; as such, the reported accuracy is the average over 10 accuracy values.

tributions. For example, in the train partition of VQA2.0, Entropy assigns much more cases than EASE to the TH split (in proportion, 25% cases for Entropy vs. 9% for EASE) and much less to the E one (33% Entropy vs. 48% EASE). On the one hand, this confirms the crucial role of our semantic component in determining EASE scores. On the other hand, we notice that the results obtained by the three models on the splits defined by Entropy follow a less clear pattern compared to the EASE ones (see Tab. 4 in Appendix). For example, in VizWiz, both BUTD and LXM-S achieve higher results in BH compared to E, which indicates that Entropy is not as effective as our tool in measuring the difficulty of a sample. Finally, for sanity check, we also tested model performance on splits having the same size of EASE's TH, BH and E but including random samples (see Tab. 5 in Appendix). The sampling was performed 10 times and results averaged. As expected, no difference in performance between the three splits was observed.

Overall, this proof-of-concept analysis reveals that current SOTA models—including the extensively pretrained LXM—suffer with samples that are deemed hard by EASE. This suggests that our diagnostic tool genuinely selects the most challenging samples of a dataset. An intuitive question is whether training a model with these hard samples can make models more robust. This is based on the intuition that challenging samples could be more informative during training compared to easy ones.

We test this hypothesis in the next section, where we use the splits defined by EASE to train models in a HardFirst (HF) approach.

4 Experiments

In HF, we train our VQA models incrementally, first using TH samples only, then adding BH samples, and finally using all training samples. The weights for the first stage are initialized randomly; we load the model's weights from previous stages for each incremental stage. For VQA2.0, the percentage of samples for each stage is 9.13% (TH), 51.79% (TH+BH), and 100% (ALL), and for VizWiz is 16%, 68.22%, and 100%. We hypothesize that harder splits, i.e., with low EASE scores, contain richer multimodal information that could be more informative during a model's learning. For comparison, we also evaluate models in the TH(R) condition: we train/fine-tune models with a set of data (with the same size as TH) randomly sampled from the training set. We repeat the sampling 10 times, and report the average accuracy.

5 Results

Results in Tab. 3 support our hypotheses. (1) With only 52% of the training data (TH+BH), BUTD obtains 90% of *all* validation accuracy (VA) in VQA2.0 compared to the model trained on the

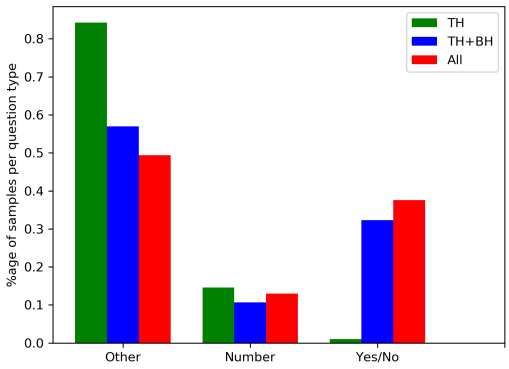

Figure 2: Percentage of samples per question type in VQA2.0-train for each of the three splits used in the HF training regime. *Other* contains all *wh-* questions, *Number* count questions, *Yes/No* polar questions.

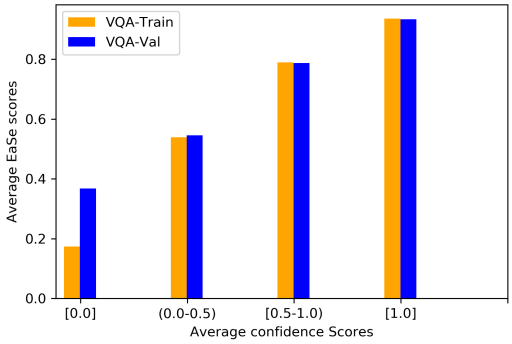

Figure 3: Average EASE scores against binned confidence scores in VQA2.0. Closed/open brackets indicate that values are included/not-included in the bin.

whole data (Table 2). This is even more pronounced in VizWiz, where using TH+BH during training (68% of total data) leads to a comparable performance as the one obtained with the whole training data. Similarly, LXM achieves 98% VA using only 52% of training data for VQA2.0, and 97% VA with 68% training data in VizWiz.

(2) Compared to the TH(R) condition, models trained/fine-tuned with TH achieve higher results in the TH split of both VQA2.0 and VizWiz, which confirms that TH samples are particularly beneficial for dealing with challenging cases. At the same time, when evaluated on the entire data (*all*), they perform similarly to TH(R) in VizWiz and slightly worse than TH(R) in VQA2.0. This is to be expected: randomly sampling from VizWiz—where 68% cases are either BH or TH—will likely produce a more similar distribution to that of TH as compared to sampling from VQA2.0, where E cases are 48% of the total. Since proportions are the same in the validation set, training/fine-tuning with easier cases in VQA2.0 will have a positive impact on E, which will drive performance on *all*.

Overall, these results indicate that the hard samples selected by EASE are more informative than easier ones and help models obtain comparable performance with significantly less training data.

6 Analysis

6.1 EASE vs. Question Types

We explore whether the hard splits selected by EASE contain question types that are known to be particularly challenging for VQA models, e.g., count and *wh-* questions. As can be seen in Fig. 2, a higher proportion of *wh-* (*Other*) and count (*Number*) questions is observed in the hardest split compared to the other splits of VQA2.0.[4] In contrast, polar questions (*Yes/No*) are poorly represented in TH, which indicates they are overall less challenging for humans and less informative for the models.

6.2 EASE vs. Confidence Scores

We test whether EASE correlates with human intuition of when is *difficult* to answer a question. To this end, we use the confidence scores provided by annotators along with their answers in VQA2.0,[5] which self-evaluate whether annotators are confident in providing their answer. We map confidence scores *yes*, *maybe*, *no* to 1, 0.5, and 0, respectively, and compute the average confidence score for each sample. We then compute *Spearman's* correlation between confidence scores and EASE scores, and find a substantial positive correlation both in train ($\rho = 0.49$) and val ($\rho = 0.48$) sets. This trend is also clear in Fig. 3, where higher confidence scores correspond to increasingly higher EASE values.

7 Conclusion

We present EASE, a simple diagnostic tool which quantifies the difficulty of a VQA sample based on its pattern of answers. We show that EASE selects the most informative samples of a dataset, which is helpful to train/fine-tune VQA models more efficiently with less, but highly-informative data. In future work, we plan to combine model prediction for difficulty estimation in EASE.

[4] A similar, though less pronounced pattern, is observed in VizWiz; see Fig. 6 in Appendix.

[5] We perform the same analysis for VizWiz (Appendix).

Acknowledgements

Shailza Jolly was supported by the TU Kaiserslautern CS Ph.D. scholarship program, the BMBF project XAINES (Grant 01IW20005), and the NVIDIA AI Lab (NVAIL) program. Sandro was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455 awarded to Raquel Fernández).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 431–441.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Danna Gurari and Kristen Grauman. 2017. 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*, pages 3511–3522.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Shailza Jolly, Sandro Pezzelle, Tassilo Klein, Andreas Dengel, and Moin Nabi. 2018. The wisdom of MaSSeS: Majority, subjectivity, and semantic similarity in the evaluation of VQA. *arXiv preprint arXiv:1809.04344*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13–23. Curran Associates, Inc.

Tomáš Mikolov, Édouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Kento Terao, Toru Tamaki, Bisser Raytchev, Kazufumi Kaneda, and Shun'ichi Satoh. 2020. Which visual questions are difficult to answer? Analysis with entropy of answer distributions. *arXiv preprint arXiv:2004.05595*.

Chun-Ju Yang, Kristen Grauman, and Danna Gurari. 2018. Visual question answer diversity. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6.

A Appendix

B Dataset analysis

As described in Section 3.2 of the main paper, we did a preliminary analysis of the answers to the validation split questions. Each $\langle image, question \rangle$, is coupled with 10 answers provided by as many annotators. We use these annotations to see the human agreement for a given $\langle image, question \rangle$ pair. Fig. 4 and Fig. 5 shows the statistics for VQA2.0 and VizWiz. It clearly shows that in VQA2.0, 33% of the questions are assigned the same answer string by all annotators (i.e., in $1/3$ questions, there is a perfect agreement between them); as for VizWiz, this percentage drops to only 3%. If we consider the questions with no more than 3 unique answers, this is the case for 71% cases in VQA2.0 and just 30% in VizWiz. We use this disagreement as a proxy for the difficulty of these datasets.

C EASE vs. Question Type

As described in Section 6.1 of the main paper, EASE selects samples with difficult question types for VQA2.0. Figure 2 (main paper) reports the proportion of question types that present in each split defined by EASE: as conjectured, we see a higher proportion of the *other* question type (i.e., *wh-*) and number questions in the hardest split of both datasets compared to the others. *Yes/No* questions are poorly represented in the hardest split, which suggests they are less challenging for humans and the models.

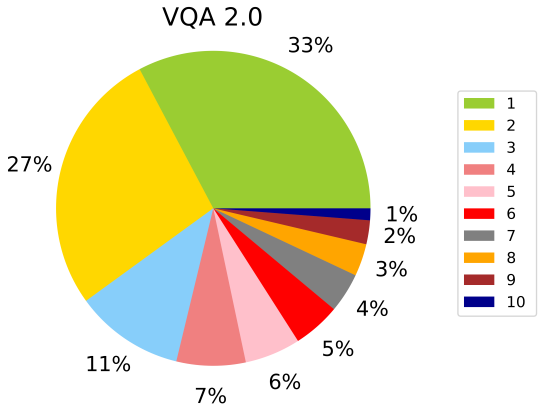

Figure 4: Distribution of samples in the validation splits of VQA2.0, against number of unique answers. E.g., in 33% samples in VQA2.0, all annotators gave the same answer.

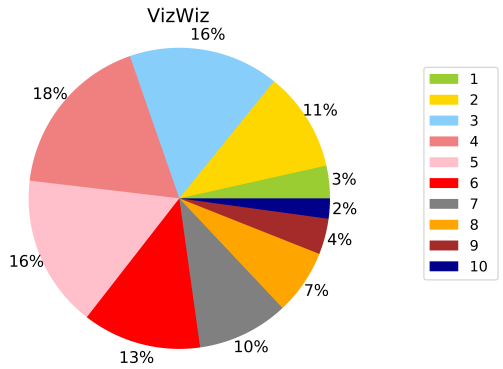

Figure 5: Distribution of samples in the validation splits of VizWiz against number of unique answers. E.g., in 3% samples in VizWiz, all annotators gave the same answer.

Figure 6 shows similar pattern in VizWiz where the percentage of Other question types is higher in TOP-HARD split selected by EASE. It is interesting to see that the number of Unanswerable questions are very low in TOP-HARD. This shows another property of EASE in which it didn't consider the usual notion of associating Unanswerable questions with hard ones, while look at human agreement/disagreement to decide difficulty.

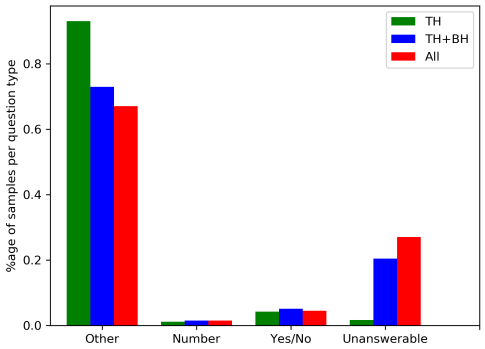

Figure 6: Number of samples per question-type in VizWiz-train for each of the three splits used in HF training regime. Here, Other belongs to reasoning questions (why, which, where), Number to counting questions, and Yes/No to polar questions.

Dataset/Split		BUTD	LXM	LXM-S
VQA	TH	34.73	42.2	34.89
	BH	71.31	78.66	71.22
	E	74.98	84.38	74.21
VW	TH	44.40	46.79	41.48
	BH	59.25	64.02	53.31
	E	52.25	64.86	40.54

Table 4: Accuracy by BUTD, LXM, and LXM-S on three validation splits of VQA2.0 (VQA) and VW. The splits are obtained via Entropy.

D Other methods to split evaluation data

As discussed in Section 3.3, we obtained TH, BH, E splits using Entropy and Random Selection. We use Eq. 4 to compute Entropy over the original answer distribution, and then subtract the score from 1.

$$E(p_f) = \frac{-1}{\eta} \sum_{k=1}^{M} p_k * log(p_k) \qquad (4)$$

We use the same criterion, as EASE, to divide our samples into TH, BH, and E. Table 4 shows that, contrary to EASE splits, in VizWiz, both BUTD and LXM-S achieve higher results in BH compared

Dataset/Split		BUTD	LXM	LXM-S
VQA	TH	63.42	71.56	63.10
	BH	63.45	71.43	63.16
	E	63.35	71.46	63.17
VW	TH	50.31	53.59	45.79
	BH	49.88	53.31	45.97
	E	50.26	53.77	46.18

Table 5: Accuracy by BUTD, LXM, and LXM-S on three random splits of validation data of VQA2.0 (VQA) and VW. The random splits are of same size as that of TH, BH, and E as mentioned in Section 3.3

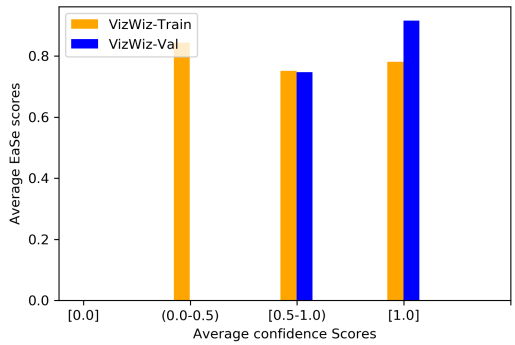

Figure 7: Average EaSe scores per confidence scores provided by annotators for both splits of VizWiz. Open/close brackets indicate that values are not/ included.

to E, which indicates that Entropy is not as effective as our tool in measuring the difficulty of a sample.

In the Random Selection, we tested the model's performance on splits having similar sizes of EASE's TH, BH, and E split. Table 5 shows that the three splits have the almost same accuracy. In particular, the random splits don't show any pattern, unlike EASE in which TH always gets the lowest accuracy and E gets the highest (Table 2 of main paper). These results clearly shows that EASE selects the smallest subset of training data which is both informative and hard.

E EASE vs. Confidence scores

As discussed in Section 6.2 of main paper, we test correlation of EaSe scores with already available human confidences. We map confidence scores *yes*, *maybe*, *no* to 1, 0.5, and 0, respectively, and compute the average confidence score for each sample. We then compute relationship between confidence scores and EASE scores. Fig. 7 shows the analysis for VizWiz data, where higher confidence scores correspond to increasingly higher EASE values. This shows that EaSe correlates with human intuition of having difficultly to answer a question.