Learning to Caption Images by Asking Natural Language Questions

Kevin Shen
shenkev@cs.toronto.edu

Amlan Kar
amlan@cs.toronto.edu

Sanja Fidler
fidler@cs.toronto.edu

Abstract

In order to bring robots into our lives, we will need to go beyond supervised learning on closed datasets to having the ability to continuously expand knowledge. Inspired by a student learning in a classroom, we present an agent that can continuously learn by posing natural language questions to humans. Our agent is composed of three interacting modules, one that performs captioning, another that generates questions and a decision maker that learns when to ask questions by implicitly reasoning about the uncertainty of the agent and expertise of the human. As compared to current active learning methods which query images for full captions, our agent is able to ask pointed questions to improve generated captions. The agent trains on the improved captions, expanding its knowledge base. We show that our approach achieves better performance using less human supervision than baselines.

1 Introduction

Children learn from teachers in an active way: asking questions about concepts that they are unfamiliar or uncertain about. In doing so, they make learning more efficient – the child who learns exactly the information they are missing – and the teacher who answers the question instead of needing to explain many aspects of a concept in full detail. As A.I. becomes more and more integrated in our everyday lives, be it in the form of personal assistants or household robots [29, 17, 24], we need it to be able to actively seek out missing information from humans – by asking questions in the form of natural language which non-experts can understand and answer.

Most existing work on complex scene understanding tasks such as VQA [6, 26, 31, 7] and captioning [12, 22, 4] has mostly focused on a closed world setting, i.e. consuming the knowledge provided by a labeled dataset. On the other hand, the goal of active learning is to be able to continuously update the model by seeking for the relevant data to be additionally labeled by a human [23]. Most active learning approaches, however, ask the human to provide a full labeling of an example, and the main challenge is in identifying the examples to be labeled, to ensure annotation efficiency. In our work, we go beyond this, by endowing the model with the ability to ask for a particular aspect of a label, and do so in natural language in order to unambiguously identify the missing information.

We focus on the task of image captioning as a proxy task for scene understanding. In order to describe the image, a model needs to generate words describing the objects, their attributes, and possibly relationships and interactions between objects. This is inherently a multi-task problem. Our goal is to allow a captioning agent to actively ask questions about the aspects of the image it is uncertain about, in a continual learning setting in which examples arrive sequentially. Thus, instead of having humans provide captions for each new training image, our agent aims to ask a minimal set of questions for the human to answer, and learn to caption from these answers. We showcase our method on the challenging MSCOCO dataset [12]. To the best of our knowledge, this is the first time that natural language question asking has been explored in a continual learning setting with real-world images.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

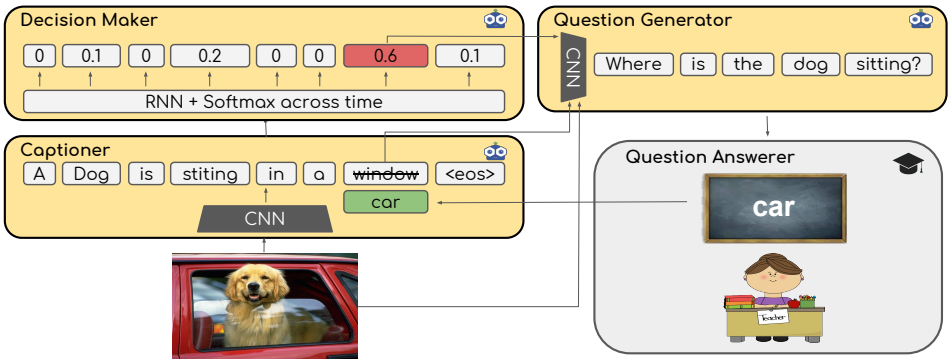

Figure 1: Improving captions by asking natural language question.

2 Our approach

Our model learns to describe images in a lifelong learning setting by asking questions to the teacher. This is illustrated in Algorithm 2. Data arrives in chunks. The first chunk D_w has complete ground truth (GT) i.e. human written captions. We refer to it as the warmup chunk. The agent learns from the remaining K unlabelled chunks $D_u = [D_{u1}, D_{u2}, \dots D_{uK}]$ with partial supervision from the teacher. The agent attempts to caption each image in the unlabelled chunks, and decides whether to replace words with answers obtained by asking questions.

Let $\mathbf{w} = (w_1, w_2, \dots w_L)$ denote a caption, I an image, $\mathbf{q}$ a question, and a an answer from the teacher. **Captioner** - $C(\mathbf{w}|I)$ is implemented as a CNN-RNN model [32]. We pretrain the captioner using MLE with teacher forcing and scheduled sampling on D_w. **Decision maker** - $DM(t|\mathbf{c})$ predicts for which word in a caption (time step t) a question should be asked (if at all).This module conditions on the context $\mathbf{c} = (c_1, c_2, \dots c_L)$ computed from the captioner. More details about c are in Appendix. Asking on the `<eos>` token is interpreted as "not asking". **Question generator** - $Q(\mathbf{q}|I, c_t)$ is also implemented as a CNN-RNN model and conditions on the context at time t. Q is trained on a custom dataset derived from VQA2.0 and MSCOCO (details found in the Appendix) and is not fine-tuned.

Fig. 1 and Algorithm 1 show the agent interacting with the teacher, which consists of two parts: a **QA bot** - $V(a|I, \mathbf{q})$ implemented following [26] and a caption scorer composed of BLEU [20], ROUGE [11], METEOR [2], and CIDEr [28]. We call this the `Mix` score, or reward, and denote it by r. Given an image, the captioner produces the caption $\mathbf{w}^0$. Let w_t be the word for which the decision module decides to ask a question. The question generator produces a question and the agent receives an answer a. The agent then replaces word w_t in $\mathbf{w}^0$ with a and predicts new caption $\mathbf{w}^1 = (w_1 \dots w_{t-1}, a, w'_{t+1}, \dots w'_L)$, by rolling out the rest of the caption from time step t using the previous hidden state h_{t-1} of the captioner and a. Finally the teacher scores both the original and improved captions. The process can be repeated by asking a second question and replacing another word at time $t' > t$. In general, the agent can ask up to N questions for a single caption. In practice, we use $N = 1$ in our experiments.

Algorithm 1 Improve captions by asking

procedure IBYASK(I)
  $\mathbf{w}^0, \mathbf{c}^0 \leftarrow C(\cdot|I)$
  for $n = 1$ to N **do**
    $t^n \leftarrow DM(\cdot|\mathbf{c}^{n-1})$
    $\mathbf{q} \leftarrow Q(\cdot|I, c_{t^n}^{n-1})$
    $a \leftarrow V(\cdot|I, \mathbf{q})$
    $\mathbf{w}^n \leftarrow [\mathbf{w}_{0:t^n-1}^{n-1}, a, C(\cdot|I, h_{t^n-1}, a)]$
    $r^n \leftarrow \texttt{Mix}(\mathbf{w}^n)$
  $n^* = \arg\max_n r^n$
  return $r^{n^*}, \mathbf{w}^{n^*}$

Algorithm 2 Lifelong learning

procedure LIFELONG(D_w, D_u)
  pretrain: C, Q, V
  initialize: MK
  $D \leftarrow D_w$
  $D_u = [D_{u1}, D_{u2}, \dots D_{uK}]$
  for D_{uk} in D_u **do**
    $D_c \leftarrow [\,]$
    for epoch $= 1$ to P **do**
      for I in D_k **do**
        $\mathbf{w}, r \leftarrow \texttt{IBYASK}(I)$
        $\mathbf{w}^*, r^* \leftarrow \texttt{IBYASK}(I,$ greedy=True$)$
        $\theta_d \leftarrow \theta_d + (r - r^*)\nabla_{\theta_d} \log p_{\theta_d}(t|\mathbf{c})$
        $D_c \mathrel{+}= (\mathbf{w}, r, \mathbf{w}^*, r^*)$
    $D \leftarrow \texttt{filter}(D_c, H)$
    train: C on D using $L(\theta_c)$

Method	$H\%$	GT %	Sup %	Mix [2]	C [28]	M [2]	R [11]	B4 [20]	B2 [20]
Equal GT	-	45.2 %	45.2 %	98.9	91.5	24.7	52.3	28.0	53.4
All GT	-	100 %	100 %	101.1	95.3	25.2	52.8	28.5	54.4
Ours	100%	21.8 %	49.9 %	99.4	91.1	24.8	52.6	28.1	54.7
Ours	85%	33.5 %	61.6 %	101.8	94.6	25.3	53.3	29.4	55.8
Ours	70%	45.2 %	73.5 %	**102.4**	**96.0**	**25.3**	**53.3**	**29.4**	**56.1**

Table 1: Evaluation on the Karpathy test split. Our model was trained using a 10% warmup chunk and 3 unlabelled chunks. Methods see all images at least once for fairness. **Note:** 100% GT% in table corresponds to 46% of MSCOCO train captions (See line 73).

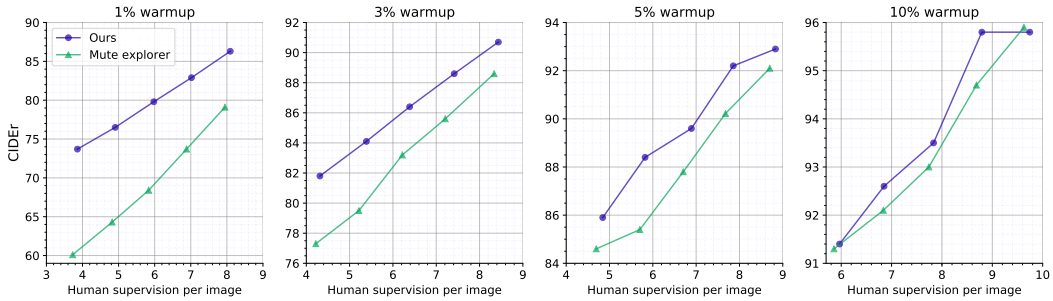

Figure 2: Caption quality on the val split. Both models are decoded greedily. Refer to the Appendix for how human supervision is calculated. For each plot, supervision is varied by changing the percentage of captions $H\%$ collected by the agent.

Lifelong learning. We now describe how the agent learns from the data obtained via a teacher. In particular, the agent makes P passes over each chunk. As the agent sees images, it stores the teacher-improved captions in a buffer D_c and trains the decision maker on the teacher's rewards. Because auto-eval metrics are non-differentiable, we train DM using REINFORCE [25]. We baseline the reward with the greedy decision reward $(r^*)^0$ (that is, what the improved-caption would have been had DM sampled greedily), following the self-critical policy gradient [22]. See lines 11-12 of Algorithm 2. In the general case where N questions are asked, the gradient for the parameters of the decision maker θ_d is:

$$\sum_{n=1}^{N}[r^n - (r^*)^n]\nabla_{\theta_d}\log p_{\theta_d}(t^n|\mathbf{c}^{n-1}), \quad t^i > t^j \text{ for } i > j \tag{1}$$

The agent collects the top m captions for each image seen during learning to distill the teacher's knowledge back into C. We choose $m = 2$ to trade off the amount of data collected and redundancy in the captions. We allow the agent to "give up" if the improved caption is still bad, and the teacher writes a new caption. In practice, the agent keeps the top $H\%$ of images based on the average caption reward from the buffer. For the other 100-$H\%$ images, the agent is given two GT captions.

Define D as the union of warmup and collected data. We assume the agent has full access to any data it has trained on in the past. After each chunk, the captioner trains on D according to a joint loss over collected and GT captions,

$$L(\theta_c) = -\sum_{\mathbf{w}} r_{\mathbf{w}}\log p_{\theta_c}(\mathbf{w}|I) - \lambda\sum_{\mathbf{w}^*}\log p_{\theta_c}(\mathbf{w}^*|I) \tag{2}$$

where $\mathbf{w}$ are collected captions, $\mathbf{w}^*$ GT captions, $r_{\mathbf{w}}$ is `Mix` reward, and λ a tuned hyperparameter.

3 Experiments

We evaluate our approach on the MSCOCO dataset [12]. In Table 3 we evaluate our lifelong learning setting against training only on GT data. All results are reported using greedy decoding. Our model was trained with a 10% warmup chunk, 3 unlabelled chunks and varying collect percentage. We report two baselines: All GT - the same number of captions as our model and Equal GT - fewer total captions but the same number of GT captions as our model. All models have the same architectures and hyperparameters and are trained on all 117,843 training images to ensure fairness.

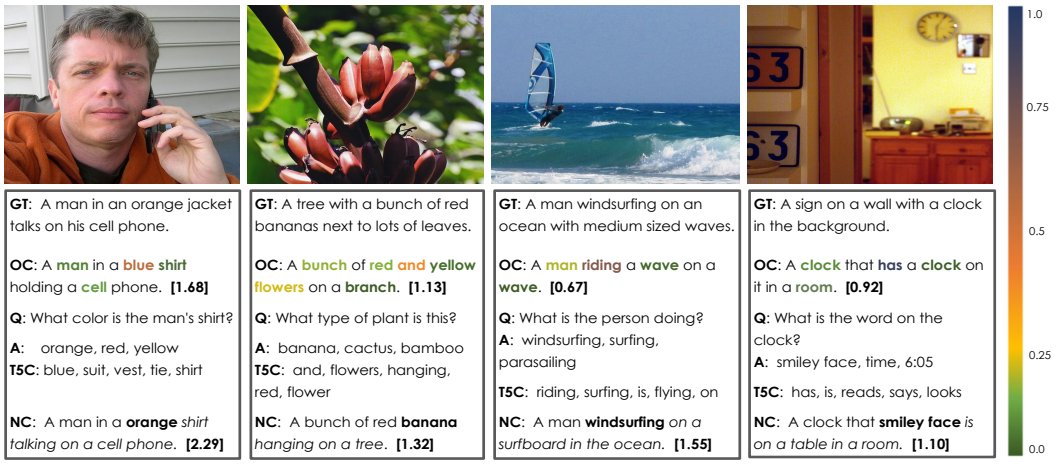

Figure 3: (best viewed in color) Three positive and one negative example from lifelong learning. T5C is the top-5 words predicted by the captioner at the time step when the question is asked. Colors in OC indicate the probability the decision maker asks about that word (scale shown on the right).

We introduce a metric to approximate human supervision effort shown in column "Sup", details can be found in the Appendix. As compared to training only on GT captions, our lifelong model achieves 1.3 mixed or 0.7 CIDEr higher while using only 45.2% of GT and 73.5% human supervision. Given the same number of GT captions, our model performs 3.5 mixed and 4.5 CIDEr better than baseline. These findings are consistent with [13] who showed that training with corrected captions performs better than purely GT captions. In summary, lifelong learning by asking questions can achieve greater performance than training only on GT with not only fewer GT captions but less human supervision.

Fig. 2 shows the regimes where learning by asking questions is most effective. The baseline is a "mute explorer" - a model that is trained in exactly the same lifelong setting as our model but samples from its own distribution to explore new captions rather than ask questions. We vary the amount of human supervision by adjusting the percentage of captions $H\%$ collected by the agents in the lifelong setting. Question exploration (QE) outperforms mute exploration (ME) in almost all settings but the difference is greater when using smaller warmup data. For 1% warmup, QE improves 13.6 CIDEr over ME (73.7 vs 60.1) but for 10% warmup the largest difference is 1.1 CIDEr (95.8 vs 94.7). This is likely because with a smaller warmup set, there are still many concepts left unexplored, hence asking the teacher pays off. Another trend is that QE becomes better than ME as the amount of human supervision decreases and the agent relies more on its collected captions rather than GT. For 3% warmup, the CIDEr difference at 8 supervision is 2.1 (90.7 vs 88.6) and at 4 supervision is 4.5 (81.8 vs 77.3). Finally, performance increases with the warmup set size for a constant amount of human supervision. For 8 supervision QE achieves achieves 93.5 CIDEr in the 10% warmup setting and 86.3 CIDEr in the 1% warmup setting. The reason for this is likely because at 1% warmup, the captioner is making too many mistakes to fix with only 1 question asked. In summary, asking questions is a more efficient way to learn from a human teacher than mute exploration. This is especially true at the low warmup regime and when the amount of human supervision is limited.

Some selected examples are shown in Fig 3. We can see that question asking is able to fix incorrect concepts in the original caption and retrieve novel nouns and verbs from the teacher. The fourth example is a failure case where the question generator produces a semantically wrong question and the QA model returns a nonsensical answer. Interestingly, the reward is higher for the new caption despite the semantic and grammatical mistakes. This highlights the weaknesses of auto-eval metrics.

4 Conclusion

In this paper, we addressed the problem of active learning for the task of image captioning. In particular, we allow the agent to ask for a particular concept related to the image that it is uncertain about, and do not require the full caption from the teacher. Done this way, the learning and teaching efficiency is shown to be improved on the challenging MS-COCO dataset.

Our work is a step towards a more natural learning setting in which data arrives continuously, and robots learn from humans through natural language questions. There are many challenges ahead in making the continual learning model more efficient, and incorporating real humans in the loop.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[2] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[4] B. Dai, S. Fidler, R. Urtasun, and D. Lin. Towards diverse and natural image descriptions via a conditional gan. *arXiv preprint arXiv:1703.06029*, 2017.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[6] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] A. K. Gupta. Survey of visual question answering: Datasets and techniques. *arXiv preprint arXiv:1705.03865*, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] H. Ling and S. Fidler. Teaching machines to describe images via natural language feedback. *arXiv preprint arXiv:1706.00130*, 2017.

[14] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun. iVQA: Inverse Visual Question Answering. *ArXiv e-prints*, Oct. 2017.

[15] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. *arXiv preprint arXiv:1612.00370*, 2016.

[16] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[17] M. J. Matarić. Socially assistive robotics: Human augmentation versus automation. *Science Robotics*, 2(4):eaam5410, 2017.

[18] I. Misra, R. Girshick, R. Fergus, M. Hebert, A. Gupta, and L. van der Maaten. Learning by asking questions. *arXiv preprint arXiv:1712.01238*, 2017.

[19] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating Natural Questions About an Image. *ArXiv e-prints*, Mar. 2016.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[21] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[22] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*, 2016.

[23] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[24] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*, volume 2, page 6, 2015.

[25] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[26] D. Teney, P. Anderson, X. He, and A. v. d. Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*, 2017.

[27] K. Uehara, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada. Visual question generation for class acquisition of unknown objects. *arXiv preprint arXiv:1808.01821*, 2018.

[28] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[29] O. Vinyals and Q. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.

[30] L. Wang, A. G. Schwing, and S. Lazebnik. Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space. *ArXiv e-prints*, Nov. 2017.

[31] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

[32] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[33] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh. Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition. *ArXiv e-prints*, Oct. 2018.

Appendix

4.1 Question generator

Method	a@1	a@3	a@5	a@10
Baseline	37.8	50.2	55.3	62.7
+CE	45.9	60.2	65.5	72
+PE	49.2	63.9	69.3	75.4
+PE +CE	52	67.2	73	79.4

Table 2: Comparing question generation models using different context inputs. (+PE) with position encoding, (+CE) with RNN encoding of the caption.

We pretrain the question generator on tuples of image, context and questions. We exploit the fact VQA2.0 and MSCOCO share images and match answers from QA pairs of VQA2.0 to words in the captions of MSCOCO to generate training samples. Doing this gives 135,670 training samples. Context for the question generator consists of: (1) the POS which determines the "question type", (2) attention weights predicted by the captioner which guide the question generator to look, (3) an encoding of the caption which provides global context and prevents asking for redundant concepts, (4) position encoding of the time step. The question generator is trained similarly to the captioner using MLE with teacher forcing and scheduled sampling. In table 2 we show the accuracy of the question generator trained using various variables in the context. Accuracy is measured by passing a *greedily* decoded question through the VQA module and comparing the teacher's answers with the ground truth answer. The baseline is a model trained only with POS and an attention maps as context. Both position and caption encoding give a boost to the accuracy. Combining both achieves the highest performance, 14.2% over baseline for the top answer. We use the full model as our question generator in the experiments.

4.2 Decision maker

Method	Mix	C	B4
No questions	86.4	74.1	22.1
Random	88.3	76.2	22.2
Entropy	88.9	76.5	22.4
Unc. metrics	89.6	77.5	22.5
Unc. metrics learned	90.8	79.3	23.2
Full learned	91.9	80.6	23.7

Table 3: Comparing different decision maker models.

The context for the decision maker consists of (1) distribution over top-k words (2) the POS (3) an encoding of the caption and (4) entropy and closeness metrics computed from the top-k words including the cosine and L2 distance between word embeddings. The closeness metrics are motivated by the fact that the captioner predicts synonyms which increase the entropy but do not suggest that the model is uncertain. Table 3 ablates different decision makers. The baseline is simply a pretrained captioner model evaluated without asking questions. For the other settings, a pretrained captioner is paired with a decision maker and question generator. The agent is evaluated by asking a single question and rolling out the caption. Entropy is picking the time step with the highest top-k word entropy. Unc. metrics includes entropy and closeness metrics. Unc. metrics learned contains a MLP to predict the logit. Full learned includes POS and an encoding of the caption. Both learned models are trained using RL on a single chunk of continual learning data. As seen from table 3, the full model gives a 6.5 CIDEr improvement over no questions. Picking the time step with maximum entropy is not very effective, only giving 0.3 CIDEr over random. Adding the closeness metrics yields 1.0 CIDEr improvement over maximum entropy. In all cases, learning improves performance, with the best learned model achieving 3.1 CIDEr more than the best non-learning model. We use the full context model as our decision maker in experiments.

Task	Avg. time (s)	Std. (s)	Time ratio
CAP	34.55	23.11	1.0
SCO	6.62	2.20	5.13
ANS	7.74	3.87	4.51

Table 4: Human time for tasks: captioning an image (CAP), scoring a caption of an image (SCO), and answering a question about an image (ANS). Time ratio is calculated relative to CAP. $N = 22$ humans were surveyed for a total of $n_c = 220$ captions, $n_q = 550$ questions, $n_s = 550$ scores.

4.3 Evaluation

The goal of active learning is to maximize performance on the test set while minimizing human supervision. We adopt this philosophy and argue that for humans-as-teachers systems, it is crucial to measure the human cost of providing feedback and make sure that it's cheaper than getting full data labels. For our experiment, a human teacher has three possible tasks given an image: produce a full caption, answer a question, score a caption. We take time taken to complete a task as a proxy to human effort. Table 4 shows the average amount of time taken by humans to do each task. Specifically, it takes on average 5.13 and 4.51 times longer to caption than score a caption or answer a question. We normalize the cost of each task to caption scoring. During lifelong learning, we charge the agent a single unit of human effort for each caption scored, 5.13 for full caption labels and 1.14 for questions answered. For full disclosure, we include the assumptions in how we evaluated our model. We filtered out questions that were repeats. We assumed no cost for answers from the teacher because we are using synthetic teacher noisier than a human. Finally, we assume the agent has some way to choose the best caption from three alternatives, for example by training a discriminator. We leave relaxing these assumptions to future works.

4.4 Dataset

We used Karpathy splits containing 117,843 training, 5K validation and 5K test images for training the captioner [9]. We randomly split the training set into warmup and continual learning portions. For each image in the warmup set, there are 5 ground truth captions. Each image in the continual set has 2 captions, collected by the agent. We used the Stanford NLP parser to get the ground truth POS labels [16]. The synthetic teacher was trained on the VQA2.0 dataset [1]. We followed a simplified implementation of [26] using a multi-answer binary cross entropy loss function. Our model achieved 64.2% on the VQA2.0 val without ensembling. `yes/no` questions were removed in the actual implementation. Image features are precomputed from ResNet-101 trained on ImageNet [5] [8]. The vocabulary sizes for the captioner, question generator and VQA were 11253, 9755 and 3003 respectively.

4.5 Implementation details

All modules are trained with batch size 20. Image features are precomputed from ResNet-101 trained on ImageNet [5] [8]. In particular from the `conv4_23` layer with adaptive pooling resulting in 14x14 spatial dimensions and 2048 channels. Gated reccurrent units (GRUs) were used for RNNs [3]. The captioner and question generator were trained using ADAM with a learning rate of $2e-4$ and with learning rate scheduling and scheduled sampling [10]. The VQA module was trained using a learning rate of $1e-3$ and word embeddings initialized with GloVe6B trained on Wikipedia2014 and Gigaword5 [21].

4.6 Related work

Image captioning, VQA, VQG. Our approach touches on each of these multimodal tasks. A popular model for image captioning is CNN-RNN with attention first proposed by [32]. Recent works have extended captioning models by fine-tuning using policy gradient [22] and [15], using generative models to increase diversity [30], and introducing a discriminator to make captions agree more with human judgement [4]. However, all works to date have approached image captioning from a closed dataset setting. Visual question answering in the real world domain remains a generally unsolved problem [7]. Recent works have exploited attention based models, top-down and bottom-up attention,

data augmentation and more to achieve state of the art on the VQA2.0 dataset [26]. Visual question generation is a recently proposed task closely related to VQA. Some recent works include [19] [14].

Active learning. In active learning, the agent selects unlabelled data to label from an oracle using various expected improvement metrics such as uncertainty sampling, query by committee, expected model change in order to maximize performance on the test set [23]. Our work differs from active learning because we don't directly optimize over which next sample gives the most improvement to the model, instead we ask natural language questions to explore novel concepts. The teacher only provides an answer to a question, rather than a caption which is a much cheaper form of supervision. However, learning by asking comes with its own challenges: specifically asking intelligent questions and reconciling the teacher's answer. Future works can explore combining active learning with learning by asking by also optimizing over the best next image to ask about.

Human in the loop captioning Human in the loop captioning has been studied [13]. Specifically, in [13] the captioner poses candidate captions to human teacher who identify incorrect phrases and corrects them. The agent then finetunes itself on the corrected captions. While both [13] and our work utilizes human feedback, our works differ in that we actively ask natural questions to learn new concepts and make corrections while [13] presents the captions to the teacher for correction. Actively asking questions requires us to introduce the decision maker module which must reason about the agent's uncertainties and the teacher's expertise. Furthermore, we present a lifelong learning training regime where the agent starts with a weak captioner that improves over time.

Learning by asking questions [18] explores question asking for the VQA task, making their work closer to traditional active learning where image and question are the unlabelled data and answers are received from an oracle to form novel training pairs. Our work differs by introducing question-asking as a surrogate task to the main task, in our case captioning. Furthermore, [18] works in a much simpler, synthetic environment with questions represented as CLEVR programs rather than natural languages. [18] does not require the agent to learn when or if to ask a question. [33] explores question asking for visual recognition. A model is trained to pose template questions to a teacher to learn about objects, attributes and relationships in a scene. [33] is also limited to training in synthetic environments with a limited set of objects and relationships and posing template questions. [27] explores using question generation as a way to explore new object classes in the context of image classification. However, they do not retrain their classifier to continually improve in a lifelong setting.