Are Red Roses Red?
Evaluating Consistency of Question-Answering Models

Marco Tulio Ribeiro
Microsoft Research
marcotcr@microsoft.com

Carlos Guestrin
University of Washington
guestrin@cs.uw.edu

Sameer Singh
University of California, Irvine
sameer@uci.edu

Abstract

Although current evaluation of question-answering systems treats predictions in isolation, we need to consider the relationship between predictions to measure true understanding. A model should be penalized for answering “no” to “Is the rose red?” if it answers “red” to “What color is the rose?”. We propose a method to automatically extract such implications for instances from two QA datasets, VQA and SQuAD, which we then use to evaluate the *consistency* of models. Human evaluation shows these generated implications are well formed and valid. Consistency evaluation provides crucial insights into gaps in existing models, and retraining with implication-augmented data improves consistency on both synthetic and human-generated implications.

How many birds?	**A:** 1
Is there 1 bird?	**A:** no
Are there 2 birds?	**A:** yes
Are there any birds?	**A:** no

(a) Input image from the **VQA dataset**.

(b) Model (Zhang et al., 2018) provides inconsistent answers.

Kublai originally named his eldest son, Zhenjin, as the Crown Prince, but he died before Kublai in 1285.

(c) Excerpt from an input paragraph, **SQuAD dataset**.

Q: When did Zhenjin die?	**A:** 1285
Q: Who died in 1285?	**A:** Kublai

(d) Model (Peters et al., 2018) provides inconsistent answers.

Figure 1: **Inconsistent QA Predictions:** Models that are accurate for questions from these datasets (first row in (b) and (d)) are not able to correctly answer follow-up questions whose answers are implied by the original question/answer. We generate such questions automatically, and evaluate existing models on their consistency.

1 Introduction

Question-answering (QA) systems have become popular benchmarks for AI systems, as they require the ability to comprehend and employ complex reasoning about the question and the associated context. In order to really excel in machine comprehension (Rajpurkar et al., 2016), for example, models need to understand the entities, coreferences, and relations in the paragraph, and align them to the information need encoded in the question. Similarly, Visual Question Answering (Antol et al., 2015) requires not only perception abilities (fine-grained recognition, object detection), but also “higher level reasoning” about how the question is related to the visual information, commonsense reasoning, knowledge based reasoning, and the understanding of location/color/size attributes.

However, recent work has shown that popular benchmarks have crucial limitations in their ability to test reasoning and comprehension. For example, Weissenborn et al. (2017) show that models can do well in the SQuAD dataset by using heuristic lexical and type overlap between the context and the question. Biases have also been observed in the popular VQA dataset, e.g. answering questions starting with “Do you see a ...” with “yes” results in 87% accuracy, and “tennis” is the correct answer for 41% of questions starting with “What sport is ...” (Goyal et al., 2017).

While there are laudable efforts to try to diminish such biases (Rajpurkar et al., 2018; Goyal et al., 2017), they do not address a fundamental evaluation question: it is not only individual predictions that matter, but also whether multiple answers reflect a *consistent* and *coherent* model. For example, in Figure 1, models answer original questions correctly but answer follow-up questions in an inconsistent manner, which indicates they do not really understand the context or the questions (e.g. simultaneously predicting 0, 1, and 2 birds in Figure 1b).

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6174–6184
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

In this paper, we propose evaluation for QA systems that measures the extent to which model predictions are consistent. We first automatically generate new question-answer pairs that are implied by existing instances from the dataset (such as the ones in Figure 1). We use this generated dataset to evaluate models by penalizing them when their predictions are not consistent with these implications. Human evaluation verifies that the generated implications are valid and well formed when compared to original instances, and thus can be used to evaluate and gain insights into models for VQA and SQuAD. Finally, we propose a simple data augmentation procedure that results in models nearly as accurate as the original models on the original data, while being more consistent when measured by our implications and by human generated implications (and thus expected to generalize better in the real world).

2 Related Work

Since QA models often exploit *shortcuts* to be accurate without really understanding questions and contexts, alternative evaluations have been proposed, consisting of solutions that mitigate known biases or propose separate diagnostic datasets. Examples of the former include adding multiple images for which the answer to the same question is different (Goyal et al., 2017; Zhang et al., 2016), or questions for which an answer is not present (Rajpurkar et al., 2018). While useful, these do not take the relationship between predictions into account, and thus do not capture problems like the ones in Figure 1. Exceptions exist when trying to gauge robustness: Ribeiro et al. (2018) consider the robustness of QA models to automatically generated input rephrasings, while Shah et al. (2019) evaluate VQA models on crowdsourced rephrasings for robustness. While important for evaluation, these efforts are orthogonal to our focus on consistency.

Various automatically generated diagnostic datasets have been proposed (Weston et al., 2015; Johnson et al., 2017). While these recognize the need to evaluate multiple capabilities, evaluation is still restricted to individual units and thus cannot capture inconsistencies between predictions, like predicting that an object is at the same time to the left and to the right of another object. Furthermore, questions/contexts can be sufficiently artificial for models to *reverse-engineer* how the dataset was created. An exception contemporaneous with our work is GQA (Hudson and Manning, 2019), where real images are used, and metrics such as consistency (similar to our own) are used for a fraction of inputs. Since questions are still synthetic, and "not as natural as other VQA datasets" (Hudson and Manning, 2019), it remains to be seen whether models will overfit to the generation procedure or to the implications encoded (e.g. many are simple spatial rules such as "X to the left of Y implies Y to the right of X"). Their approach is complementary to ours – they provide implications for ~54% of their synthetic dataset, while we generate different implications for ~67% of *human generated* questions in VQA, and ~73% of SQuAD questions.

(a) Example input image.

Q: What room is this?
A: bathroom

(b) Example (q, a) pair.

Type	Cov	Example
Logeq	56.8%	Is this a bathroom? Yes
Nec	50.2%	Is there a bathroom in the picture? Yes
Mutex	34.6%	Is this a kitchen? No

(c) Implication types, with coverage and examples.

Figure 2: **VQA Implications and examples.** Implications can be generated for 67.3% of the original data.

3 Generating Implications

Let an instance from a QA datset be represented by (c, q, a) denoting respectively the context (image or paragraph), question, and answer (c may be omitted for clarity). We define logical implications as $(c, q, a) \rightarrow (c, q', a')$, i.e. an answer a to q implies that a' is the answer for question q' for the same context. We now present a rule-based system that takes (q, a) and generates $(q, a) \rightarrow (q', a')$.

Visual QA (q, a) pairs in VQA often have both positive and negative implications that we encode into three types of *yes/no* implications, illustrated in Figure 2: *logical equivalence* (**Logeq**), *necessary condition* (**Nec**) and *mutual exclusion* (**Mutex**) (more examples in appendices). To generate such instances, we use a dependency parser (Dozat et al., 2017) to recognize root/subject/object and build the implication appropriately, and to detect auxiliary/copula that may need to be moved. Logical equivalence implications are generated by transforming the original (q, a) into a proposition, and then asking the "yes-no" equivalent by moving auxiliary/copula, adding "do" auxiliaries, etc (e.g. "Who painted the wall? man" → "Did the man paint the wall? yes"). Necessary conditions are created via heuristics such as taking numerical answers to "How many X" questions and asking if there are any X present (e.g. "How many birds? 1" → "Are there any birds? yes"), or asking if answer nouns are in the picture (e.g. bathroom in Figure 2c). We used WordNet (Miller, 1995) to find antonyms and other plausible answers (hyponyms of the original answer's hypernym) when generating mutual exclusion implications, as illustrated in changing "bathroom" to "kitchen" in Figure 2c. We also used a 4-gram language model (Heafield et al., 2013) to smooth implication questions (e.g. adding "the", "a", etc before inserting the original answers into implication questions).

SQuAD Since the answers need to be spans in the paragraph, we cannot generate the same kinds of implications (e.g. yes/no questions are not suitable). Instead, we use the QA2D system of Demszky et al. (2018) to transform a (q, a) into declarative form d, and then use the dependency parse of d to extract questions about the subject (**Subj**), direct object (**Dobj**), adjectival modifiers (**Amod**), or prepositional phrases (**Prep**) (Table 1). To decide which WH-word to introduce, we use a NER tagger (Honnibal and Montani, 2017) coupled with heuristics, e.g. if the answer is "in DATE" or "in LOC", the WH-words are "when" and "where", respectively.

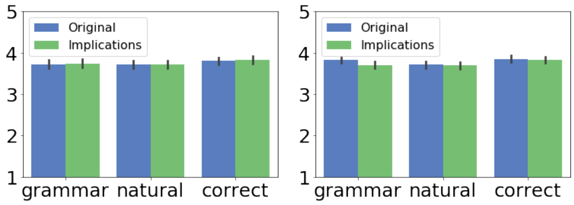
Evaluating consistency We want the generated implications to meet the following criteria: (1) the questions are well formed, (2) the answers are correct, and (3) the implication is valid, i.e. if we generate an implication $(q, a) \rightarrow (q', a')$, an answer a to q really implies that a' is the answer to q'. If these are met (Section 4), we can evaluate the consistency of a large fraction of predictions in these datasets (67.3% of VQA and 73.2% of SQuAD) by taking (q, a) instances predicted correctly by the model, generating implications $(q, a) \rightarrow (q', a')$, and measuring the frequency at which the model predicts the generated questions correctly.

4 Experiments

In this section, we assess the quality of the generated (q', a') pairs, measure consistency of models for VQA and SQuAD, and evaluate whether data augmentation with implications can improve the consistency of existing models.

Type	Cov	Example
Subj	29.3%	When did Zhenjin die? 1285 →Who died in 1285? Zhenjin
Dobj	10.0%	When did Denmark join the EU? 1972 →What did Denmark join in 1972? the EU
Amod	29.7%	When did the Chinese famine begin? 1331 → Which famine began in 1331? Chinese
Prep	46.1%	Who received a bid in 1915? Edison →When did Edison receive a bid? 1915

Table 1: **SQuAD Implication types and examples.** Implications cover 73.2% of the original data.



(a) VQA

(b) SQuAD

Figure 3: **Quality of implications** (q', a') and original (q, a) as judged by workers: grammaticality and naturalness of questions, and correctness of answers.

4.1 Quality of Implications

We randomly select 100 generated implications and original instances for each dataset, and ask 5 different crowd workers on Amazon Mechanical Turk to rate each question for grammaticality and naturalness on a scale of 1 to 5 (following Demszky et al. (2018)). We also ask workers to evaluate the correctness of the answer given the question and context (image or paragraph). The results presented in Figures 3a and 3b show that the average scores on all criteria are nearly indistinguishable between original instances and the generated implications, which indicates that implication questions are well formed and answers are correct.

4.2 Validity of Implications

In order to check if (q, a) really implies (q', a') (i.e. check if the implication is valid), we show workers the (q, a) *without* the context and ask them to answer the implication question q' assuming the original answer a is correct. If $(q, a) \rightarrow (q', a')$, workers should be able to answer q' correctly even in the absence of the image or paragraph. As an example, the answer to the implication question in Figure 4a should be "yes" for any image, if the original

Original	**Q:** How many zebras are there? **A:** 4
Implication	**Q:** Are there any zebras?
Control	**Q:** Is this scene taken in the wild?

(a) Example from the **VQA** dataset.

Original	**Q:**Which IPCC author criticized the TAR? **A:** Richard Lindzen
Implication	**Q:** What did Richard Lindzen criticize?
Control	**Q:** Who responded to Lindzen's criticisms?

(b) Example from the **SQuAD** dataset.

Figure 4: **Testing the validity of implications:** given an original (q, a) pair, humans should be able to deduce the answer for the implication question without context, but not necessarily for the control question.

	VQA		**SQuAD**	
	Impl	**Control**	**Impl**	**Control**
#Answered	99%	13%	95%	4%
#Correct|Answered	97%	77%	97%	50%

Table 2: **Validating Implications:** Crowd evaluation of the validity of implications, where the first row indicates how often workers provide an answer, while the second row indicates the *precision* of their answers.

(q, a) holds. For control purposes, we also include question-answer pairs asked of the same context from the dataset, expecting that workers would not be able to answer these without the original context most of the time (Figure 4a provides an example where a reasonable guess can be made, which is not true in Figure 4b). We take the same 100 implications from the previous experiment and add 100 control questions, each evaluated by 5 workers. Workers are instructed to abstain from answering if the original (q, a) does not give them enough information to answer q' or the control question. For each question, we evaluate the worker majority answer w.r.t. the implication or control answer. The results in Table 2 are quite positive: workers almost always provide the correct answer a' to our implication question q' when given only the original (q, a) pair and no additional context, which indicates the implication is valid. On the other hand, workers under-predict and are inaccurate for the control questions, which is expected since there is no necessary logical connection between (q, a) and the control question.

4.3 Evaluating Consistency of QA Models

Having concluded that our generated implications are high quality and typically valid, we proceed to use them to evaluate the logical consistency of models. For VQA, we evaluate the **SAAA** baseline (Kazemi and Elqursh, 2017), a recent model with a counting module (**Count**; Zhang et al., 2018), and bilinear attention networks (**BAN**; Kim et al., 2018). For SQuAD, we evaluate **bidaf** (Seo et al., 2017), bidaf with ELMO embeddings (**bidaf+e**; Peters et al., 2018), **rnet** (Wang et al., 2017), and Mnemonic Reader (**mnem**; Hu et al., 2018). All models are trained with available open source code with default parameters.

The results for VQA are presented in Table 3. Note that more accurate models are not necessarily more consistent, and that all models are particularly inconsistent in the Mutex category. One specific category of Mutex that affects all models was asking the equivalent $n+1$ questions when the answer is a number n, e.g. "How many birds? 1" implies "Are there 2 birds? no". SAAA, Count, and BAN had, respectively, 35.3%, 22.4% and 32.2% consistency in this category even though Count has a module specific for counting (implications are binary yes/no questions, and thus random guessing would give 50% consistency). This is probably because the original dataset contains numbers in 12.3% of answers, but only in 0.3% of questions, thus models learn how to *answer* numbers, but not how to reason about numbers that appear in the question. Evaluating consistency in this case is useful for finding gaps in models' understanding, and similar insights can be reached by considering other violated implications.

For SQuAD (Table 4), we consider a prediction as consistent if it had *any* overlap with the implied answer. Again, models with different accuracies do not vary as much in consistency. All models are less consistent on direct object implications. Interestingly, ~12% of questions in the training data have the WH-word in the direct object subtree (e.g. "Who did Hayk defeat?"), while 53% are in the subject subtree (e.g. "Who is Moses?"), which may warrant further investigation.

All models had average consistency lower or equal to 75%, which indicates they do not possess real comprehension of the concepts behind many of their correct predictions. Besides surfacing this, consistency evaluation provides clues as to potential sources of such problems, such as the lack of questions with numbers in VQA.

Model	Acc	LogEq	Mutex	Nec	Avg
SAAA	61.5	76.6	42.3	90.2	72.7
Count	65.2	81.2	42.8	92.0	75.0
BAN	64.5	73.1	50.4	87.3	72.5

Table 3: Consistency of VQA Models.

Model	F1	Subj	Dobj	Amod	Prep	Avg
bidaf	77.9	70.6	65.9	75.1	72.4	72.1
bidaf+e	81.3	71.2	69.3	75.8	72.8	72.9
rnet	79.5	68.5	67.0	74.7	70.7	70.9
mnem	81.5	70.3	68.0	75.8	71.9	72.2

Table 4: Consistency of SQuAD Models.

4.4 Data Augmentation with Implications

We propose a simple data augmentation technique: for each (q, a) in the training set, add a generated implication (q', a') if one exists. We evaluate the consistency of models trained with augmentation on held-out implications, to check whether they generalize to unseen generated implications. Further, to verify if augmentation improves consistency "in the wild", we collect *new* implications from Mechanical Turk by showing workers (q, a) pairs without context (image or paragraph), and asking them to produce new (q', a') that are implied by (q, a) for any context. For VQA, we restrict a' to be yes / no, while for SQuAD we filter out all a' that are not present in the original paragraph, resulting in a total of $3,277$ unique implication annotations for VQA and $1,027$ for SQuAD. While workers sometimes create implications similar to ours, they also include new patterns; implications that contain negations (all models are very inconsistent on these), word forms for numbers (e.g. "one"), comparatives ("more", "less"), and implications that require common sense, such as ("What type of buses are these? double decker"→"Do the buses have 2 levels? yes"). The results are presented in Table 5. Accuracy on the validation set remains comparable after augmentation, while consistency on both generated and worker-provided implications improves across models and tasks. We also evaluate **SAAA** on the GQA dataset (Hudson and Manning, 2019) (Count and BAN use features that are not allowed in GQA): while accuracy is comparable (41.4% before augmentation, 40.4% after), consistency goes up significantly (59.3% before, 64.7% after). These results indicate that data augmentation is useful for increasing consistency with a small trade off in accuracy. We leave more sophisticated methods of enforcing consistency (e.g. in models themselves) for future work.

	Model	Validation Accuracy		Consistency (rule-based)		Consistency (crowdsourced)	
VQA	SAAA	61.5	60.8	72.7	94.4	73.0	75.6
	Count	65.2	64.8	75.0	94.1	73.8	77.3
	BAN	64.5	64.6	72.4	95.0	72.3	77.9
SQuAD	bidaf	77.9	76.4	72.1	79.1	68.2	70.9
	bidaf+e	81.3	80.7	72.9	81.2	70.7	70.6
	rnet	79.5	79.5	70.9	79.8	66.5	68.1
	mnem	81.5	81.3	72.2	81.5	68.7	73.9

Table 5: **Data Augmentation:** Accuracy (F1 for SQuAD) and consistency results before and after data augmentation. Consistency (rule-based) is computed on our generated implications, while (crowdsourced) is computed on crowdsourced implications.

5 Discussion

We argued that evaluation of QA systems should take into account the relationship between predictions rather than each prediction in isolation, and proposed a rule-based implication generator which we validated in crowdsourcing experiments. The results of this approach are promising: consistency evaluation reveals gaps in models, and augmenting training data produces models that are more consistent even in human generated implications. However, data augmentation has its limitations: it may add new biases to data, and it cannot cover all the different implications or ways of writing questions. Ideally, we want models to be able to reason that "What color is the rose? Red" implies "Is the rose red? Yes" without needing to add every possible implication or rephrasing of every (q, a) to the training data. We hope that our work persuades others to consider the importance of consistency, and initiates a body of work in QA models that achieve real understanding by design. To support such endeavours, generated implications for VQA and SQuAD, along with the code to generate them and for evaluating consistency of models, is available at https://github.com/marcotcr/qa_consistency.

Acknowledgments

We would like to thank Sara Ribeiro, Julian Michael, Tongshuang Wu, Tobias Schnabel, and Eric Horvitz for helpful discussions and feedback. This work was funded in part by the NSF award #IIS-1756023.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Timothy Dozat, Peng Qi, and Christopher D Manning. 2017. Stanford's graph-based neural dependency parser at the conll 2017 shared task. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

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 *CVPR*, volume 1, page 3.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified kneserney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 690–696.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4099–4106. AAAI Press.

Drew A Hudson and Christopher D Manning. 2019. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE.

Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1571–1581.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations (ICLR)*.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. *arXiv preprint arXiv:1902.05660*.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 189–198.

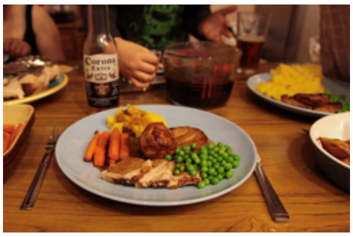
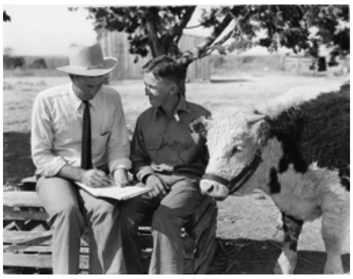
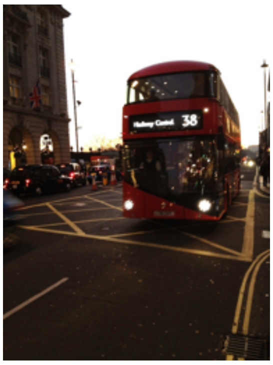
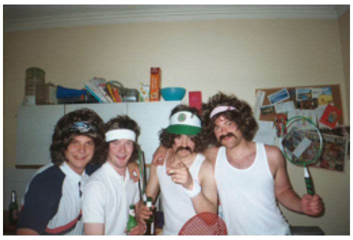
Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural qa as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280.

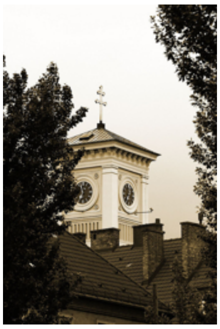
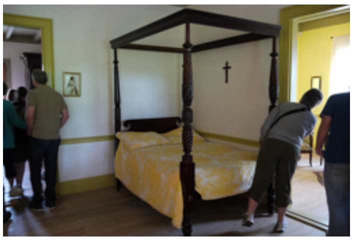
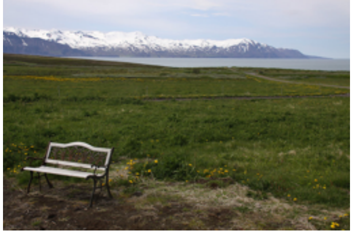
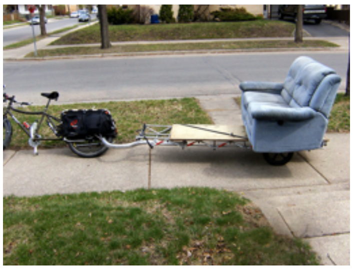
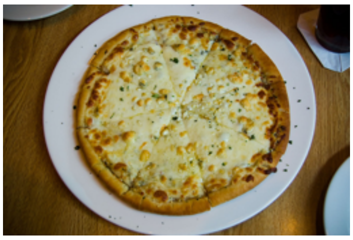
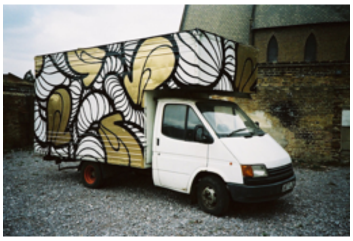
Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yan Zhang, Jonathon Hare, and Adam Prgel-Bennett. 2018. Learning to count objects in natural images for visual question answering. In *International Conference on Learning Representations*.

A Implications selected at random: VQA dataset

Context	Question/Answers		
	**Original:**	Where is the fork?	**A:** left of plate
	Logeq:	Is the fork on the left of the plate?	**A:** yes
	**Original:**	What are the men sitting on?	**A:** bench
	Logeq:	Are the men sitting on the bench?	**A:** yes
	Nec:	Is there a bench in the picture?	**A:** yes
	**Original:**	What is the number on the bus	**A:** 38
	Logeq:	Is the number on the bus 38?	**A:** yes
	**Original:**	What kind of headwater is the man on the left wearing?	**A:** headband
	Logeq:	Is the man on the left wearing a headband?	**A:** yes
	Nec:	Is there a headband in the picture?	**A:** yes

Context	Question/Answers		
	**Original:**	What is on the top of the clock tower	**A:** cross
	Logeq:	Is the cross on the top of the clock tower?	**A:** yes
	Mutex:	Is the area on the top of the clock tower?	**A:** no
	Nec:	Is there a cross in the picture?	**A:** yes
	**Original:**	Is this a Christian home?	**A:** yes
	Mutex:	Is this an unchristian home?	**A:** no
	Nec:	Is this a home?	**A:** yes
	**Original:**	What separates the meadow from the mountains in the background?	**A:** water
	Logeq:	Does water separate the meadow from the mountains in the background?	**A:** yes
	**Original:**	What color is the couch?	**A:** blue
	Logeq:	Is the couch blue?	**A:** yes
	Mutex:	Is the couch orange?	**A:** no
	Nec:	Is there anything blue in the picture?	**A:** yes
	**Original:**	How many toppings are on this pizza?	**A:** 2
	Logeq:	Are 2 toppings on this pizza?	**A:** yes
	Mutex:	Are 3 toppings on this pizza?	**A:** no
	Nec:	Are any toppings on this pizza?	**A:** yes
	**Original:**	What material is the building in the back, made of?	**A:** brick
	Logeq:	Is the building in the back, made of brick?	**A:** yes
	Mutex:	Is the building in the back, made of stone?	**A:** no
	Nec:	Is there a brick in the picture?	**A:** yes

B Implications selected at random: SQuAD dataset

Context:	The first commercially viable process for producing liquid oxygen was independently developed in 1895 by German engineer Carl von Linde and British engineer William Hampson.	
Original:	When was liquid oxygen developed for commercial use?	**A:** 1895
Subj:	What was developed for commercial use in 1895?	**A:** liquid oxygen
Amod:	Liquid oxygen was developed for which use in 1895?	**A:** commercial

Context:	In the 1960s, a series of discoveries, the most important of which was seafloor spreading, showed that the Earth's lithosphere, which includes the crust and rigid uppermost portion of the upper mantle	
Original:	Which parts of the Earth are included in the lithosphere?	**A:** the crust and rigid uppermost portion of the
Amod:	Which portion of the upper mantle are included in the lithosphere?	**A:** crust and rigid uppermost
Amod:	The crust and rigid uppermost portion of which mantle are included in the lithosphere?	**A:** upper
Prep:	The crust and rigid uppermost portion of what are included in the lithosphere?	**A:** upper mantle
Prep:	Where are the crust and rigid uppermost portion of the upper mantle included?	**A:** lithosphere

Context:	Around 1800 Richard Trevithick and, separately, Oliver Evans in 1801 introduced engines using high-pressure steam; Trevithick obtained his high-pressure engine patent in 1802.	
Original:	In what year did Richard Trevithick patent his device?	**A:** 1802
Subj:	Who patented his device in 1802?	**A:** Richard Trevithick

Context:	The average Mongol garrison family of the Yuan dynasty seems to have lived a life of decaying rural leisure, with income from the harvests of their Chinese tenants eaten up by costs of equipping and dispatching men for their tours of duty.	
Original:	How were the Mongol garrison families earning money?	**A:** harvests of their Chinese tenants
Amod:	The Mongol garrison families were earning money by the harvests of their which tenants?	**A:** Chinese
Prep:	The Mongol garrison families were earning money by the harvests of what?	**A:** their Chinese tenants

Context:	Of particular concern with Internet pharmacies is the ease with which people, youth in particular, can obtain controlled substances (e.g., Vicodin, generically known as hydrocodone) via the Internet..	
Original:	What is an example of a controlled substance?	**A:** Vicodin
Amod:	An example of which kind of substance is Vicodin?	**A:** controlled
Prep:	An example of what is Vicodin?	**A:** controlled substance

Context:	...the exterior mosaic panels in the parapet were designed by Reuben Townroe who also designed the plaster work in the library	
Original:	Who designed the plaster work in the Art Library?	**A:** Reuben Townroe
Dobj:	What did Reuben Townroe design in the Art Library?	**A:** plaster work
Prep:	Where did Reuben Townroe design the plaster work?	**A:** Art Library

Context:	Combustion hazards also apply to compounds of oxygen with a high oxidative potential, such as peroxides, chlorates, nitrates, perchlorates, and dichromates because they can donate oxygen to a fire.	
Original:	What other sources of high oxidative potential can add to a fire?	**A:** compounds of oxygen
Prep:	Compounds of what can add to a fire?	**A:** oxygen
Prep:	What can compounds of oxygen add to?	**A:** fire

Context:	In 1881, Tesla moved to Budapest to work under Ferenc Pusks at a telegraph company, the Budapest Telephone Exchange.	
Original:	Which company did Tesla work for in 1881?	**A:** the Budapest Telephone Exchange
Subj:	Who worked for the Budapest Telephone Exchange in 1881?	**A:** Tesla
Prep:	When did Tesla work for the Budapest Telephone Exchange?	**A:** 1881

Context:	...membrane is used to run proton pumps and carry out oxidative phosphorylation across to generate ATP energy.	
Original:	What does oxidative phosphorylation do?	**A:** generate ATP energy
Subj:	What generates ATP energy?	**A:** oxidative phosphorylation
Dobj:	What does oxidative phosphorylation generate?	**A:** ATP energy

Context:	formerly model C schools tend to set much higher school fees than other public schools.	
Original:	How do the fees at former Model C schools compare to those at other schools?	**A:** much higher
Amod:	The fees at former Model C schools compare to those at which schools by much higher ?	**A:** other