A Case Study of the Shortcut Effects in Visual Commonsense Reasoning

Keren Ye, Adriana Kovashka
University of Pittsburgh, Pittsburgh PA 15260, USA
{yekeren,kovashka}@cs.pitt.edu

Abstract

Visual reasoning and question-answering have gathered attention in recent years. Many datasets and evaluation protocols have been proposed; some have been shown to contain bias that allows models to "cheat" without performing true, generalizable reasoning. A well-known bias is dependence on language priors (frequency of answers) resulting in the model not looking at the image. We discover a new type of bias in the Visual Commonsense Reasoning (VCR) dataset. In particular we show that most state-of-the-art models exploit co-occurring text between input (question) and output (answer options), and rely on only a few pieces of information in the candidate options, to make a decision. Unfortunately, relying on such superficial evidence causes models to be very fragile. To measure fragility, we propose two ways to modify the validation data, in which a few words in the answer choices are modified without significant changes in meaning. We find such insignificant changes cause models' performance to degrade significantly. To resolve the issue, we propose a curriculum-based masking approach, as a mechanism to perform more robust training. Our method improves the baseline by requiring it to pay attention to the answers as a whole, and is more effective than prior masking strategies.

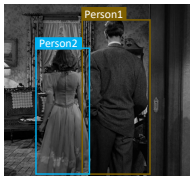
Introduction

Models for vision-and-language (VL) tasks, such as visual question answering (VQA) and visual commonsense reasoning (VCR), perceive the features of an image and provide natural language responses regarding the visual contents. The comprehensiveness of the VQA process seems to require complete human-like intelligence, and has inspired great interest. Unfortunately, in practice, models have many opportunities to bypass "reasoning" and instead find shallow patterns in the data in order to match answers to image-question pairs. By "reasoning" we mean a generalizable process that analyzes the structure of the world as demonstrated by training data, pays attention to links between participants in the scene as well as between entities and their semantic properties, and analyzes how these correspond to the entities or events indicated in the question. Such a process ideally persists when small changes are made to the potential answer options without changing their meaning, because the entities represented by these options remain the same.

To solve visual question answering tasks, studies used feature fusion (Antol et al. 2015; Zhou et al. 2015), attention (Lu et al. 2016; Fukui et al. 2016; Anderson et al. 2018), or question-related modular structures (Andreas et al. 2016b,a). Recently, transformers modeling cross-modal attention (Alberti et al. 2019; Chen et al. 2020; Lu et al. 2019a; Li et al. 2019; Su et al. 2020) have also been applied. These methods are trained with supervision i.e. the correct answers are provided by human annotators on the training set. The nature of supervised training means methods are rewarded for finding *any* connection between inputs (image-question) and outputs (answer options). In other words, methods can do well without performing complex reasoning, if they can find enough "shallow" matches between input and output. We refer to such shallow matches as "shortcuts".

Shortcut effects: Example and definition. Consider the example in Fig. 1 from the VCR dataset (Zellers et al. 2019). In the figure, [person1] (male) is on the right and [person2] (female) is on the left. The **correct option has the most overlap with the question**: the "[person1]" and "[person2]" tags, and the word "dress". Thus, to answer this question, the model need not perform reasoning or even look at the image. Examples in VCR vary: not all contain shortcuts of this nature, yet others contain even more severe shortcuts. For example, some incorrect answer choices mention entities entirely unrelated to question and image, which are thus easy to eliminate.

We define "shortcuts" as a way of achieving the correct answer by simply **matching repeated references** to the same entities in the question and answer options. We find that in 67.8% samples for the Q→A task in VCR, and 65.2% samples for the QA→R task, the correct choices have the most overlapped referring tags among the candidates. Further, state of the art methods' performance drops signifi-



*Question: What does **[person1]** think of **[person2]**'s **dress**?*
*Correct answer: **[person1]** thinks **[person2]** looks stunning in her **dress**.*
Incorrect #1: She does not approve.
Incorrect #2: [person2] is a girl and girls like to wear makeup.
Incorrect #3: [person1] is confused and annoyed by [person2] following her in the store.

Figure 1: Shortcut effects: An example.

cantly when these shortcuts are removed.

One reason for shortcuts is that humans often repeat the keywords or essential entities of the question to give a complete answer; this is hard to avoid during data collection. Further, the shortcuts may have broader forms across different modalities. E.g., in language "excited" is a common association to "feeling", people often perform action "eating" at visual environment "restaurant", etc. We emphasize that researchers that train models for VCR should pay more attention given these inevitable shortcuts. Yet, prior methods have sometimes exacerbated shortcuts. E.g., the "grounding" of objects in (Zellers et al. 2019) enables feature-level shortcuts since the same object feature may appear in both question and answer. We specifically examine shortcuts in the case of VCR, while the same phenomenon is likely to present in other datasets where question-answering is formulated as multiple-choice task and features full-sentence answers e.g. (Tapaswi et al. 2016; Zadeh et al. 2019).

While machine learning methods for other tasks also find easy ways to do well at the target task, we argue that "shortcuts" are a particular type of dataset bias whose reduction requires specific mechanisms. What exacerbates the problem is that such shortcutting is easier in the multiple-choice VQA setting compared to classification. In image classification, a shortcut has to be found across modalities, i.e. pixels to labels. In VQA, a shortcut between input and output can easily be found within the same modality, i.e. text in the question and text in answers. However, shortcuts are distinct from prior biases discovered in VQA datasets (Goyal et al. 2017), because they have more to do with shallow string matches than modes in the answer distribution. No prior dataset bias work has studied shortcut effects.

In this work, we first quantify the impact of shortcuts on state-of-the-art models. We propose two methods to augment VCR evaluation. One makes small word-level changes while maintaining the original meaning, while the other examines which word a VCR method most depends on. We show the performance of SOTA methods drops significantly on the modified evaluation data. Second, we propose a novel masking technique to make training more robust and make models rely on more extensive evidence compared to individual shortcuts. Because masking may under-utilize useful information, we perform masking on curriculum, with a large masking ratio initially and gradually reducing it. We show our robustly trained method collapses less when partial evidence is missing, and curriculum masking is more effective than prior masking techniques in both the original and modified settings. Our paper is an initial exploration of shortcut effects in VQA and a case study of VCR. We expect it to inspire future ideas of overcoming shortcut effects. Our code and data are available at https://github.com/yekeren/VCR-shortcut-effects-study.

Related Work

Dataset biases vs shortcuts. Many works studied VQA dataset biases to improve data acquisition. For example, (Goyal et al. 2017; Zhang et al. 2016) optimized the annotation pipeline to cope with questions being answerable without examining the visual contents. These biases arise because often, the language prior coming from the distribution in answer frequencies (e.g. yes/no) is very strong. The problem we study is orthogonal in two ways. First, it has to do with question-answer shortcuts rather than the presence of modes in the answer class distribution. For example, a language prior is exemplified by the following: "Q: How many dogs are in this image? A: 2 [most common answer]." In contrast, a shortcut is exemplified by the following: "Q: What does [person1] think of [person2]'s dress? Correct: [person1] thinks [person2]'s dress is... Incorrect: [person2] thinks that..." Second, shortcuts arise due to the multiple-choice setting in VCR (e.g. 4 answer options), where it is easy to eliminate some options because they have less overlap with the question compared to the correct option. In contrast, in the classic VQA setting, answering is approached as classification among e.g. 1000 options, and bias arises from the modes in the answer distributions for specific question types, not because answers match parts of the question.

Coping with dataset bias. Prior work has largely focused on biases in the classification probability given the question, but the shortcuts we study take the broader form of co-existing words or objects, in the question-answer pair. The VCR authors (Zellers et al. 2019) trained an adversarial matching model to provide suggestions for the distracting options, but we show shortcuts still exist. (Hudson and Manning 2019) developed a question engine to leverage scene graph structures to dispatch diverse reasoning *questions*, thus tightly control the answer distribution; this does not remove question-answer shortcuts. (Johnson et al. 2017a) proposed the procedurally generated *synthetic* CLEVR dataset and minimized the biases of the annotations through random sample generation; this is not possible for VCR. (Agrawal et al. 2018) propose train-test splits that have different answer distribution priors, but over-reliance on priors is not the only problem. For the shortcuts problem we examine, diversifying the options does not ensure shortcuts will not be exploited (and may make the problem worse if done naively). Constructing adversarial data to attack the trained models is a way to diagnose the effects of dataset biases, and we propose a technique in our work. In *text* question answering (QA), (Jia and Liang 2017; Wang and Bansal 2018; Jiang and Bansal 2019) applied adversarial evaluation on the SQuAD dataset (Rajpurkar et al. 2016). They turned questions into confusing facts that should have no impact on the answers and added them to the knowledge context to distract models. Our strategy for modifying the evaluation is simpler—we only replace pronouns with existing person tags or we mask, rather than generating new phrases. We are not aware of prior adversarial evaluation in the VCR setting.

Robust training. General-purpose techniques, e.g. dropout (Srivastava et al. 2014), regularization, or pre-training, potentially benefit VL tasks through learning robust feature representations. In NLP, distributed representations (Mikolov et al. 2013; Pennington, Socher, and Manning 2014) project words with similar context to neighboring points in feature space and are often used to initialize sequence models. ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019) learn context word embeddings through left-to-right/right-to-left or masked language modeling, and are

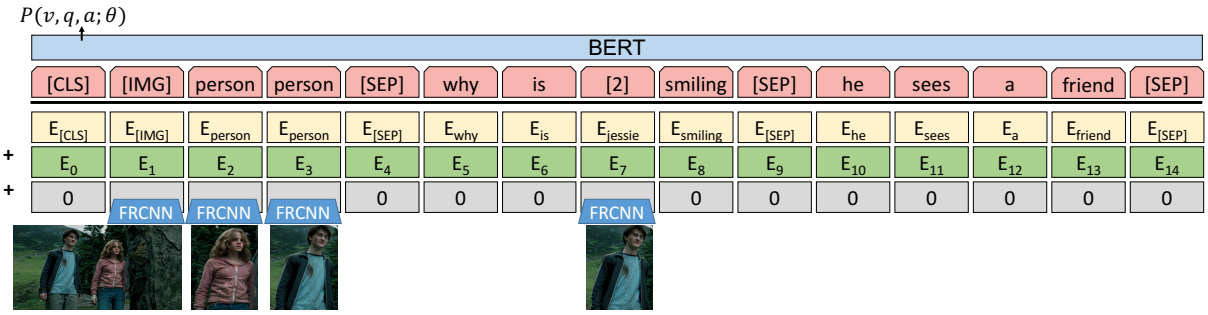


Figure 2: Model architecture for our shortcut effects study. We use BERT as the language model backbone and add the tag sequence features generated by Fast-RCNN to the token and positional embeddings. The contextualized feature of [CLS] is used to predict the answer-question matching score.

often used for pretraining to benefit downstream tasks. In vision-and-language, (Chen et al. 2020; Lu et al. 2019a; Li et al. 2019) extend BERT to the multi-modal setting and pretrain on large VL datasets e.g. Conceptual Captions (Sharma et al. 2018). The methods we study also use some forms of pre-training (e.g. embeddings), but still suffer from shortcut effects. To cope with specific dataset biases, (Ramakrishnan, Agrawal, and Lee 2018) push their full VQA model away from a question-only one, thus encourage the former to pay more attention to the visual features. (Lu et al. 2019b) train textual distractors using reinforcement learning to confuse the answering module thus partially resolve the priors in question type. However, these are limited to the classification setting, and tackle a different type of bias. We instead propose a technique to cope with the input-output shallow matches.

Alternative methods for VQA. While neuro-symbolic modular reasoning disentangled from perception (Johnson et al. 2017b; Yi et al. 2018; Hudson and Manning 2018; Mao et al. 2019) may increase robustness, thus far these methods target the CLEVR dataset, with limited extensions to VQA, and no extensions that we know of for the VCR task.

Approach

First, we develop techniques to quantify the detrimental effect of shortcuts, by removing some them at test time. Second, we propose a technique to make training more robust.

VCR Task and Basic Model

The visual Commonsense Reasoning (VCR) task involves two subtasks. The first one (Q→A) requires predicting if an answer choice $\boldsymbol{a}$ fits the context of both visual information $\boldsymbol{v}$ and question $\boldsymbol{q}$ (i.e., multi-choice VQA). The second subtask (QA→R) predicts the likelihood of a rationale $\boldsymbol{r}$, given $\boldsymbol{v}$, $\boldsymbol{q}$, and $\boldsymbol{a}^*$ ($\boldsymbol{a}^*$ is the correct answer). For each question, the dataset provides one correct choice (answer or rationale) as well as three distracting (incorrect) options. The evaluation protocol also involves a combined Q→AR metric without separate training. Fig. 3 shows examples of Q→A. Unlike other VQA datasets, VCR mixes person/object tag annotations with the questions and answers, denoting that the text refers to a particular image region. We find these tags create problematic shortcuts.

To achieve unified modeling $\mathcal{P}$ of both subtasks, we follow (Alberti et al. 2019; Zellers et al. 2019; Yu et al. 2019; Lin, Jain, and Schwing 2019) to reparameterize the formulation of QA→R (Eq. 1). We concatenate $\boldsymbol{q}$ and $\boldsymbol{a}^*$ to obtain question $\boldsymbol{q}'$ in QA→R, and treat rationale $\boldsymbol{r}$ as answer $\boldsymbol{a}'$. Thus both VCR models differ only in parameters $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$.

$$\begin{aligned} Q \rightarrow A &: \mathcal{P}(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{a}; \boldsymbol{\theta}) \\ QA \rightarrow R &: \mathcal{P}(\boldsymbol{v}, \boldsymbol{q}', \boldsymbol{a}'; \boldsymbol{\theta}'), \text{where } \boldsymbol{q}' = [\boldsymbol{q}; \boldsymbol{a}^*],\ \boldsymbol{a}' = \boldsymbol{r} \end{aligned} \tag{1}$$

For our modified, more challenging evaluation setting (Methods to Evaluate the Shortcut Effects), we use four recent, diverse methods. To show improvements through robust training, we focus on B2T2 (Alberti et al. 2019) to implement $\mathcal{P}$. We choose B2T2 because: (1) The architecture is simple. It is essentially a BERT (Devlin et al. 2019) model with multimodal inputs, with the next sentence prediction of BERT modified to be the matching prediction of the answer given question-image pair. (2) BERT-based architectures are popular for the VCR task (Alberti et al. 2019; Chen et al. 2020; Lu et al. 2019a; Su et al. 2020; Li et al. 2020; Gan et al. 2020; Yu et al. 2020) hence our choice of method is representative. (3) B2T2 achieves good results without expensive pre-training on external, non-VCR data, while models like UNITER (Chen et al. 2020) are more dependent on expensive out-of-domain pre-training.

Fig. 2 shows how we predict the joint probability of $\boldsymbol{v}$, $\boldsymbol{q}$, $\boldsymbol{a}$. Similar to B2T2, we create a token sequence by concatenating the image object labels (from the VCR dataset, e.g. "person") and textual words. We also create a tag features sequence using the associated Fast-RCNN features (Girshick 2015), adapted to the same dimensions as the word embeddings; for words not mentioning any visual objects, we pad with zeros. Then, the embeddings of the token sequence and the tag features are pointwise added and normalized before being fed to the BERT model to get the contextualized feature vectors. Next, we add a linear layer on the feature of the `[CLS]` token to estimate $\mathcal{P}(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{a}; \boldsymbol{\theta})$ (a scalar). We use *sigmoid cross-entropy* to optimize the model. Thus, $\mathcal{P}(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{a}; \boldsymbol{\theta})$ approximates a probability which is large if the answer is appropriate for this image and question.

Question	Original	Changed to
Why is [person2] in such a rush?	*He used the wrong ingredients to make the meal.*	*[person2] used the wrong ingredients to make the meal.*
How is [person2] feeling?	*[person1] is very excited.*	*[person2] is very excited.*

Table 1: Examples - Modifying distractor answer options.

All models that we train, including baselines, use BERT-Base (12 layers, 768 hidden units) and ResNet-101 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009), as the language and vision models' backbones, respectively. We keep all the layers in BERT-Base trainable while we freeze the ResNet-101 layers until the ROIAlign. We use 4 GTX1080 GPUs, batch size of 48 (12 images per GPU), learning rate of 1e-5, ADAM optimizer, and the Tensorflow framework. We train for 50k steps (roughly 11 epochs) on the 212,923 training examples and save the model performing best on the validation set (26,534 samples), for each method in Table 5. Each model took 10 hours to train.

Methods to Evaluate the Shortcut Effects

We propose two methods (rule-based and adversarial) to modify the answer candidate options in the evaluation set. Both methods keep meanings unchanged in most cases, but the second does change meaning in some cases and is primarily used to gauge what kind of words in the answer options a VCR method relies on. The methods highlight shortcuts and test the models' capability of utilizing comprehensive features instead of shortcuts.

Rule-based modification. Inspired by the observations in Introduction, we first use a set of simple rules to modify references to persons. While individual words in the answers are changed, the meaning of the answer choices remains unchanged or almost unchanged. We always modify both the distracting and correct options. Depending on whether the question contains one or multiple person tags, we refer to the rule as RULE-SINGULAR or RULE-PLURAL. This method only covers a proportion of the validation data but causes a significant drop for several recent methods.

For ground-truth options, we turn person tags into pronouns to make the answer less associated with the question-image pair at the surface (removing tag matches). To choose the proper gender pronouns, we first check the hints ("his", "her", etc.) in both the question and answer. For groups of tags ("`[person1,person2]`"), we replace with the pronoun "they". Since the distracting options are semantically unrelated to the image, we assume the pronouns and person tags do not matter in *most* cases. We turn pronouns ("he", "she", "they") and any other person tags, into the person tags asked in the question. Tab. 1 shows some examples, where the question is about `[person2]`, and both "he" and `[person1]` are changed to `[person2]`.

Discussion: Shortcuts vs distribution shifts. Changing the distribution of the evaluation set compared to the training set naturally causes a drop in performance. What this modified evaluation allows us to do is measure precisely how much different methods rely on person tag shortcuts. Further, it creates a more realistic, less inflated setting to demonstrate the reasoning capacity of different models, including ours which enables robust training. The shortcuts we highlight through our modified evaluation, are distinct from distribution shifts. In particular, our robust training algorithm that copes with shortcuts (next section) improves performance in both the modified evaluation and the original setting. In contrast, a method that exploits the distribution shifts created with our modification by training on such modified data, degrades performance in the original setting.

Adversarial modification. We next propose an adversarial modification. First, we train a B2T2 model $\mathcal{P}(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{a}; \boldsymbol{\theta})$ to solve the VCR problem using unmodified data. Given ground-truth label information $\mathcal{C}(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{a}) \in \{0, 1\}$ ($\boldsymbol{a}$ *is* or *is not* the answer to $\{\boldsymbol{v}, \boldsymbol{q}\}$), we define the potential shortcut evidence in Eq. 2, where $|\cdot|$ denotes the length of the sequence and $\Psi(\boldsymbol{x}, i)$ is a function to replace the i-th token in sequence $\boldsymbol{x}$ with a special token `[MASK]`. Eq. 2 looks for the evidence in the answer choices that makes the model most "fragile", i.e. the special position in answer $\boldsymbol{a}$ such that after replacing that token with a mask, the cross-entropy loss is *maximized* (because we want to confuse models).

$$\begin{aligned}\mathbf{argmax}_{i\in[1,|\boldsymbol{a}|)]}[&-\mathcal{C}(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{a}) \log \mathcal{P}(\boldsymbol{v}, \boldsymbol{q}, \Psi(\boldsymbol{a}, i); \boldsymbol{\theta}) \\ &-(1-\mathcal{C}(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{a})) \log(1-\mathcal{P}(\boldsymbol{v}, \boldsymbol{q}, \Psi(\boldsymbol{a}, i); \boldsymbol{\theta})]\end{aligned} \tag{2}$$

Intuitively, there should be more than one word in the correct answer ($\mathcal{C}(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{a}) = 1$) that allows a method to find that answer. However, compared to the rule-based revisions, we expect that performance will drop for the adversarial setting because the adversarial method potentially changes the meaning. Thus, in this setting, we are more interested in **what words cause performance to drop the most when masked**, rather than how much performance drops. We provide statistics regarding the masked words in Experiments. Adversarial modification mostly attacks word repetitions, pronouns, and word tenses. **This supports our intuition about shortcut effects: models use trivial, content-free hints** to make decisions instead of real reasoning. We expect the rule-based modification to more precisely show the effect of a specific type of shortcut (person tag), while adversarial revision will show the broader effects in a less controlled environment (as any word can be chosen in Eq.2).

Robust Training with Curriculum Masking

We propose a new way to make training more robust such that it can overcome shortcut effects, using masking on a curriculum. We describe two masking baselines, then our new masking technique. Note the strategies we used to create the modified evaluation sets are not appropriate to augment the training set because they potentially add new shortcuts, as we show in Experiments.

Masking baselines: Masked VCR and language modeling. We randomly replace tokens in answers with the `[MASK]` during training, with a probability of 5%, 10%, 15%, or 30%. We predict whether a masked answer follows the question, and refer to this technique as MASKING in Experiments. The `[MASK]` token is not applied in inference.

We also use masked language modeling (MLM), where the task is to predict the missing tokens in the masked sentence. We use a 0.001 coefficient to weigh the MLM softmax-cross-entropy loss; this is because too large weighting negatively affects the main loss (answer choice cross-entropy). We jointly train for the two objectives and refer to the approach as MASKING+MLM. Both of these masking strategies are inspired by BERT (Devlin et al. 2019).

Our method: Masked VCR on a curriculum. There is a tradeoff between masking to increase robustness and maintaining the required information. We found that the more masking is applied during training, the better the result in the modified settings, but the worse it is on the original standard validation. Thus, we propose a new curriculum masking approach which slowly decays the amount of masking that is applied during training. It uses a high masking probability at the beginning, then gradually reduces the masking ratio:

$$\text{Masking ratio} = \text{Initial ratio} * e^{-(\text{Decay rate} * \text{Train steps})}$$

We feed hard examples (higher masking ratio) at the start because this regularizes the model to pay more attention to the inputs as a whole, while in later stages the model leverages examples that have closer distribution to the unmasked validation data. We refer to this method as OURS-CL, and show its benefit in Experiments. While curriculum learning (Jiang et al. 2015; Zamir et al. 2017; Zhang, David, and Gong 2017; Jiang et al. 2018) has been tried to decide the *order* of tasks for pre-training (Ma et al. 2019; Wang et al. 2020; Clark et al. 2020), to our knowledge, **ours is the first method to mask using a curriculum.**

Discussion. None of our robust training approaches focus on pre-training on large external corpora, because its effect makes it unclear how a method makes its decisions, and this pre-training incurs a large computational cost. The contribution of pre-training on an external dataset gives mixed results: B2T2 (Alberti et al. 2019) show pretraining on Conceptual Captions improves accuracy by 0.4%, vs 1% (and 2% for second-stage in-domain pre-training) for UNITER (Chen et al. 2020). Our experiments with existing masking techniques resemble in-domain pretraining, but we show these are inferior to masking using a curriculum, in both the original and modified evaluation settings.

Experiments

We qualitatively demonstrate, then quantitatively measure, the effect of shortcuts through our modified evaluations, on four recent and competitive VCR methods. We then test how well our robust training strategy copes with the challenge.

Qualitative Results on the Modified Options

We show that R2C (Zellers et al. 2019) (checkpoint by authors) is confused once the expected shortcuts are no longer available. In Fig. 3, we show the option chosen by the method in bold, and the correct one is underlined. In Fig. 3 top, in the original setting, only options A0 and A1 contain the person tag `[2]`, hence the model only had to rule out "carriage". In the rule-modified setting, the model confused

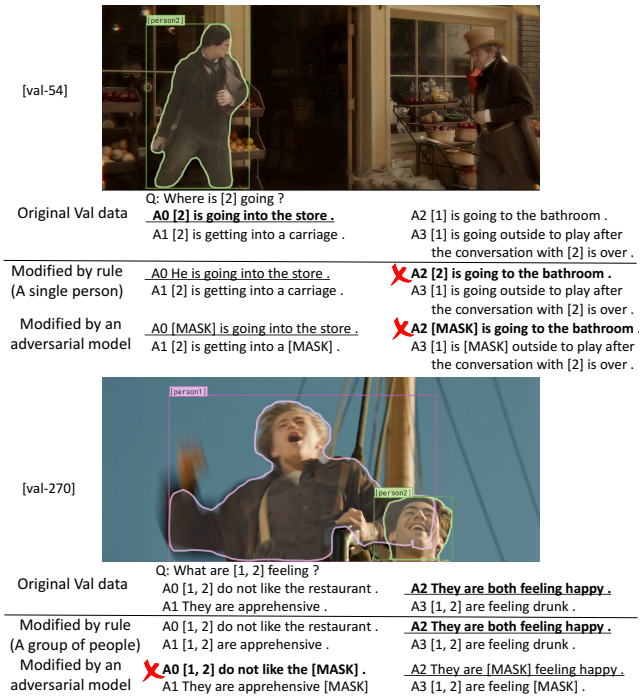


Figure 3: Qualitative study of shortcuts. We underline the ground-truth and bold the prediction of R2C. R2C was fooled by negligible changes in the answer options.

"store" with "bathroom" once the easy way of ruling out non-matching references (`[person1]` v.s. `[person2]`) is no longer applicable. The adversarial method has detected the same shortcut, replaced `[person2]` with `[MASK]` and tricked the model. In Fig. 3 bottom, the model relied too much on detecting the incompatibility between the image and concept "restaurant": when the word "restaurant" is masked in the adversarial setting, the model chooses the incorrect option A0 rather than detecting the "happy" people.

The rule-based method, targeting the over-relying of person tags, is focused and precise. In comparison, it is not that intuitive what the adversarial method attacks. We hence show statistics of the top-20 masked words. In Tab. 2, p(mask x) denotes the frequency the adversarial method chose the token x to mask; $\sum_x$ p(mask x)=1. Since token appearance frequency varies, we also report p(mask x|exist x). We observe that the adversarial method chose to hide the top-20 words in most cases ($\sum_{x \in \text{top-20}}$ p(mask x)=45.37%). However, it is hard to say these words are crucial for human reasoning. For example, "#PERSON", "he", "they", "she" are pronouns referring to persons; "is", "a", "are" are articles with hints regarding numbers; "will", "going" involve tense information. "Not" and "yes" are two exceptions, and hiding them will change the meaning. However, the proposed adversarial method relies on no human intervention and such simple cases can be ruled out by extra rules. Besides, they only constitute 2% of the revised evaluation data while the person tags are the leading choice of masking.

Many words in Table 2 are "content-free", in the sense

Token x	p(mask x)	p(mask x| exist x)	Token x	p(mask x)	p(mask x| exist x)
#PERSON	25.71%	27.84%	will	0.77%	11.33%
.	3.82%	3.79%	to	0.65%	2.04%
he	2.53%	12.09%	going	0.59%	14.13%
is	1.56%	2.78%	are	0.59%	3.72%
they	1.54%	11.70%	feeling	0.56%	22.25%
not	1.29%	24.36%	him	0.47%	12.09%
she	1.20%	12.86%	it	0.41%	7.27%
yes	0.86%	22.47%	her	0.40%	8.99%
the	0.82%	2.97%	something	0.40%	11.62%
a	0.80%	3.06%	someone	0.39%	15.43%

Table 2: Statistics of top-20 words removed by the adversarial revision. Note how often content-free words (e.g. pronouns) are key for answering, hence removed.

that other nouns and verbs should intuitively be more important. We conclude that **meaning does not change greatly when a single, however important, word is removed**, yet method performance drops by 14-34%. We thus emphasize that researchers should pay special attention to the issue at both the data acquisition and model learning phases. Besides VCR, shortcuts may also arise in other multiple-choice VQA tasks, e.g. MovieQA (Tapaswi et al. 2016) and Social-IQ (Zadeh et al. 2019), when fragments of the question and answer can be trivially matched.

Shortcut Effects on Rule-based Modified Setting

We next quantitatively demonstrate how our modified evaluation setting affects the following four VCR methods.

- B2T2 (Alberti et al. 2019) proposes early integration of visual features in BERT to benefit from stacked attention
- HGL (Yu et al. 2019) uses vision-to-answer and question-to-answer graphs using BERT/CNN embeddings
- TAB-VCR (Lin, Jain, and Schwing 2019) incorporates objects and attributes into the R2C tag matching
- R2C (Zellers et al. 2019) builds RNN layers on the pre-extracted BERT embeddings and uses attention mechanisms to highlight important visual/language elements

For HGL, TAB-VCR and R2C, we download the best-trained checkpoints provided by the authors and run inference using our modified validation. We refer to the reference implementation to implement B2T2, since no checkpoint was provided. Note B2T2, HGL and TAB-VCR are competitive in the VCR leaderboard, achieving ranks 17, 20, and 24. The better ranks are occupied by other BERT-based models (Chen et al. 2020; Lu et al. 2019a; Su et al. 2020; Li et al. 2020; Yu et al. 2020; Gan et al. 2020) focusing on **pre-training using large external VL datasets** and even object, attribute and relationship predictors (Yu et al. 2020). These settings incur significant additional data collection cost.

We observe that **merely replacing the pronouns and person tags confuses the state-of-the-art models.** Tab 3 shows the results. For RULE-SINGULAR, the average drop in accuracy, between the standard and modified validation sets, is 5% for Q→A and 6% for QA→R. Although the performance of QA→R is better than that of Q→A in the original setting, the performance drop was higher on QA→R. Thus,

Questions regarding	Count	METHOD	Q→A STD VAL	Q→A MOD VAL	QA→R STD VAL	QA→R MOD VAL
A single person e.g., *Where is [2] going ?* (RULE-SINGULAR)	16,154	R2C	64.5	58.5	67.8	62.0
		HGL	69.8	66.1	70.8	64.5
		TAB-VCR	70.5	65.4	72.4	66.3
		B2T2	69.9	63.3	69.1	64.9
A group of people e.g., *What are [1,2] feeling ?* (RULE-PLURAL)	3,657	R2C	62.2	59.7	66.9	65.4
		HGL	69.2	67.5	70.7	69.8
		TAB-VCR	69.8	66.8	71.3	70.9
		B2T2	67.6	65.3	69.3	67.9

Table 3: Shortcuts in VCR: rule-based modified evaluation.

	Method	STD VAL	Rm. a shortcut ADVTOP-1	Utilizing the potential shortcuts KEEPTOP-1	KEEPTOP-3	KEEPTOP-5
Q→A	R2C	63.8	49.8	51.8	65.9	67.5
	HGL	69.4	54.5	51.8	68.4	71.5
	TAB-VCR	69.9	54.9	49.6	65.1	69.7
	B2T2	68.5	37.0	51.0	75.0	80.4
QA→R	R2C	67.2	47.0	31.3	44.5	55.3
	HGL	70.6	51.6	33.7	48.8	60.2
	TAB-VCR	72.2	53.9	32.6	44.5	55.7
	B2T2	68.5	34.7	28.1	37.6	54.5

Table 4: Shortcuts in VCR: adversarially-modified data.

we question if models have learned to reason instead of utilizing the shortcuts. The average drops for RULE-PLURAL are 2% and 1%, respectively, likely because annotators were less willing (lazy) to point out each individual if there are too many of them. Thus the referring preference of the correct and distracting choices are similar in the RULE-PLURAL (both options prefer "they" to the person tags).

Shortcut effects on Adversarially-Modified Setting

We constructed the following validation sets to check the shortcut effects. ADVTOP-1 removes the most probable evidence (see Fig. 3), while in contrast KEEPTOP-K only uses the top-K potential pieces of evidence. Tab. 4 shows the results. Compared to STD VAL, ADVTOP-1 is more challenging since one important piece of evidence is masked out, thus performance drops by 14-32% accuracy on Q→A and 18-34% on QA→R. Given that the average length of the answer choices in both tasks are 7.65 and 16.19 tokens respectively, it is not understandable that masking out one token shall have such a big impact unless the models are fragile and base their decisions on single tokens. Finally, the strong performance in the KEEPTOP-K setting further shows models made decisions based on little facts instead of comprehensive thinking. For example, based on carefully chosen[1] three tokens, R2C is able to improve accuracy from 63.8% (full answers) to 65.9% (3-word answers). Note that we used a *single B2T2 model* (different initialization) to generate the same adversarial evaluation data *for all models*. This is why the performance drop is larger on B2T2 in Tab. 4.

[1]The adversarial model used the label information to look for the token positions (see Eq. 2).

Method	Std Val	Rule-Singular	Rule-Plural	AdvTop-1
Baseline (B2T2)	68.5	63.3	65.3	37.0
Aug Rule	**67.0**	**78.8**	**69.9**	31.6
Aug AdvTop-1	64.4	57.3	57.0	**81.4**
Masking 0.05	**69.3**	**63.9**	**66.0**	48.8
Masking 0.10	68.7	62.8	64.7	50.1
Masking 0.15	68.2	62.0	63.3	**50.6**
Masking 0.30	64.1	56.6	56.8	47.5
Masking 0.05 + MLM	68.5	62.9	64.8	47.3
Masking 0.10 + MLM	69.1	63.8	65.0	**50.6**
Ours-CL Init0.30 Decay1e-4	69.6	64.5	64.7	51.7
Ours-CL Init0.30 Decay5e-5	**69.9**	**65.9**	**66.8**	54.5
Ours-CL Init0.50 Decay1e-4	69.4	65.0	65.0	53.0
Ours-CL Init0.50 Decay5e-5	69.8	65.4	66.3	**54.9**
Baseline (B2T2)	68.5	64.9	67.9	34.7
Ours-CL Init0.30 Decay5e-5	70.6	66.6	70.4	47.9

Table 5: Our method enables the most robust training. All results show Q→A except for the bottom two which show QA→R. The best method per group on Q→A is bolded, and the best method per task is underlined.

Contribution of Our Robust Training

We next verify the extent to which robust training enables us to recover some of the lost performance. We train B2T2 based on the authors' reference implementation, but skip the expensive pre-training stage (contributing only 0.4% in (Alberti et al. 2019)). We refer to this method as Baseline, and compare it to the strategies described in Approach: Robust Training. Tab. 5 shows the results. First, we found using the rule-based and adversarial strategies (Aug Rule, Aug AdvTop-1) to augment the training data achieved better performance in the corresponding evaluation settings (as expected), but did not perform well in the original nor the other modified setting. On the Q→A task (first 13 rows), when the probability of replacing a random token is small (e.g., Masking 0.05), it leads to robust results in both the original and rule-modified settings (69.3% vs. 68.5%, 63.9% vs 63.3%, etc.) However, performance degrades (64.1% v.s. 68.5%, 56.6% vs 63.3%) once too few pieces of evidence are used in training (Masking 0.30). Masking 0.10 + MLM slightly outperforms the baseline in some settings, but is worse than Masking 0.05. In contrast, **our best curriculum learning method,** Ours-CL Init0.30 Decay5e-5, **outperforms all masking/MLM methods and the B2T2 baseline.** We observe the benefit of dynamic, curriculum masking, compared to static masking from prior work, in both the original and modified settings.

Attention Weights Show Broader Use of Evidence

Next, we show that robust training leads to models' broader attention to various evidence. We use BertViz (Vig 2019) and examine attention strength. In Fig. 4, we observe that to determine the effect of "turned around", Ours-CL (right) pays attention to more tokens in the question, and determines "walk away" to be important as the result of "turned around". In contrast, the baseline without robust training (middle) based the prediction of "turned" on "would" *because this content-free word* is in the question (a shallow match), thus did not learn to *reason*.

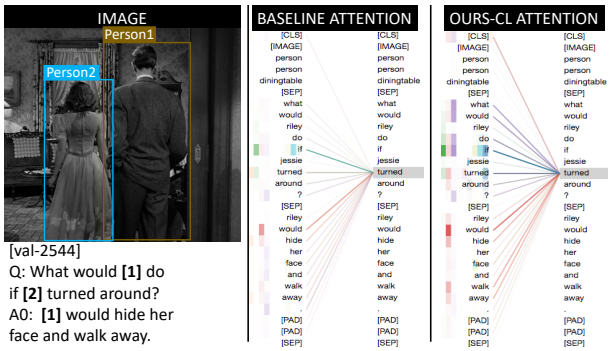


Figure 4: Learned attention of the baseline and Ours-CL. Attention strength is denoted by darker/lighter shaded boxes under word "Layer". We show weights in BERT-Base Layer-9, which is potentially the last layer of interpretable high-level reasoning. In Layer-10, word features are aggregated in [SEP], while in the last Layer-11, [SEP] is gathered in [CLS]. Please zoom figure to 300%.

Entropy of	Layer11	Layer10	Layer9	Layer8	Layer7	Layer6
Ours-CL/Baseline	**104.62%**	**105.20%**	98.97%	98.07%	102.13%	101.83%
Entropy of	Layer5	Layer4	Layer3	Layer2	Layer1	Layer0
Ours-CL/Baseline	102.53%	99.61%	100.67%	101.52%	97.74%	98.93%

Table 6: Our model pays attention to broader evidence: The numbers shown are the ratios of attention entropies for Ours-CL and those corresponding to Baseline.

Quantitatively, we compute the attention distribution on the validation set and the average *entropy* per BERT layer (from different attention heads and image examples). We show in Tab. 6 the ratio of entropy for Ours-CL vs Baseline. In the last layers (11 and 10), which are used to compute the answers, the entropy of Ours-CL is larger, which means our model pays attention to broader evidence.

Conclusion

We evaluated the effect of shortcuts, i.e. shallow matching between questions and answers in the VCR dataset. We demonstrated that subtle changes to the answer options, which should not change the meaning or correct choice, do successfully trick methods, causing large drops in performance for four recent models. We further proposed a novel technique for robust training, which applies masking on a curriculum, starting with a large amount of masking and gradually reducing it. We showed that our method was more successful in undoing the harmful effect of shortcuts, compared to techniques that have been previously used for achieving robustness through masking.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No.1718262. We thank the reviewers for their feedback.

References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alberti, C.; Ling, J.; Collins, M.; and Reitter, D. 2019. Fusion of Detected Objects in Text for Visual Question Answering. 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)*, 2131–2140.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016a. Learning to Compose Neural Networks for Question Answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016b. Neural Module Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. VQA: Visual Question Answering. In *The IEEE International Conference on Computer Vision (ICCV)*.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Learning universal image-text representations. In *European Conference on Computer Vision (ECCV)*.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations (ICLR)*.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 457–468.

Gan, Z.; Chen, Y.-C.; Li, L.; Zhu, C.; Cheng, Y.; and Liu, J. 2020. Large-Scale Adversarial Training for Vision-and-Language Representation Learning. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Girshick, R. 2015. Fast R-CNN. In *The IEEE International Conference on Computer Vision (ICCV)*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hudson, D. A.; and Manning, C. D. 2018. Compositional Attention Networks for Machine Reasoning. In *International Conference on Learning Representations (ICLR)*.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jia, R.; and Liang, P. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021–2031.

Jiang, L.; Meng, D.; Zhao, Q.; Shan, S.; and Hauptmann, A. G. 2015. Self-Paced Curriculum Learning. In *Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Jiang, Y.; and Bansal, M. 2019. Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017a. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Hoffman, J.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017b. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2989–2998.

Li, G.; Duan, N.; Fang, Y.; Jiang, D.; and Zhou, M. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* .

Lin, J.; Jain, U.; and Schwing, A. 2019. TAB-VCR: Tags and Attributes based VCR Baselines. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019a. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lu, J.; Ye, X.; Ren, Y.; and Yang, Y. 2019b. Good, Better, Best: Textual Distractors Generation for Multi-Choice VQA via Policy Gradient. *arXiv preprint arXiv:1910.09134* .

Ma, X.; Xu, P.; Wang, Z.; Nallapati, R.; and Xiang, B. 2019. Domain Adaptation with BERT-based Domain Classification and Data Selection. In *Workshop on Deep Learning Approaches for Low-Resource NLP*.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations (ICLR)*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, 3111–3119.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ramakrishnan, S.; Agrawal, A.; and Lee, S. 2018. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1541–1551.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2556–2565.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15(56): 1929–1958.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. MovieQA: Understanding Stories in Movies Through Question-Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vig, J. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, 37–42.

Wang, C.; Wu, Y.; Liu, S.; Zhou, M.; and Yang, Z. 2020. Curriculum Pre-training for End-to-End Speech Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wang, Y.; and Bansal, M. 2018. Robust Machine Comprehension Models via Adversarial Training. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 575–581.

Yi, K.; Wu, J.; Gan, C.; Torralba, A.; Kohli, P.; and Tenenbaum, J. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2020. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph. *arXiv preprint arXiv:2006.16934* .

Yu, W.; Zhou, J.; Yu, W.; Liang, X.; and Xiao, N. 2019. Heterogeneous Graph Learning for Visual Commonsense Reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zadeh, A.; Chan, M.; Liang, P. P.; Tong, E.; and Morency, L.-P. 2019. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zamir, A. R.; Wu, T.-L.; Sun, L.; Shen, W. B.; Shi, B. E.; Malik, J.; and Savarese, S. 2017. Feedback networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1808–1817. IEEE.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2016. Yin and Yang: Balancing and Answering Binary Visual Questions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Y.; David, P.; and Gong, B. 2017. Curriculum domain adaptation for semantic segmentation of urban scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, 2020–2030.

Zhou, B.; Tian, Y.; Sukhbaatar, S.; Szlam, A.; and Fergus, R. 2015. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167* .