

Counterfactual Vision and Language Learning

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, Anton van den Hengel

{ehsan. abbasnejad, damien. teney, amin. parvaneh, javen. shi, anton. vandenhengel}@adelaide.edu.au

Australian Institute for Machine Learning & The University of Adelaide, Australia

Abstract

The ongoing success of visual question answering methods has been somewhat surprising given that, at its most general, the problem requires understanding the entire variety of both visual and language stimuli. It is particularly remarkable that this success has been achieved on the basis of comparatively small datasets, given the scale of the problem. One explanation is that this has been accomplished partly by exploiting bias in the datasets rather than developing deeper multi-modal reasoning. This fundamentally limits the generalization of the method, and thus its practical applicability. We propose a method that addresses this problem by introducing counterfactuals in the training. In doing so we leverage structural causal models for counterfactual evaluation to formulate alternatives, for instance, questions that could be asked of the same image set. We show that simulating plausible alternative training data through this process results in better generalization.

1. Introduction

Recent advances in computer vision and natural language understanding have paved the way for a variety of tasks that combine visual and textual modalities [28, 15, 7, 5, 33]. Visual Question Answering (VQA) is one such task in which the goal is to answer a question framed in natural language that relates to an image. VQA thus requires a high-level understanding of the visual scene and the question, and an ability to relate (or ground) the two. Much of the interest around VQA, and the associated vision-and-language problems, stems from the fact that success might represent a step toward artificial intelligence. A variety of real-world applications have arisen also, including aiding the visually impaired, searching through large quantities of visual data via natural language interfaces, and flexible tasking of robots.

Current end-to-end VQA models achieve high accuracies on most of the available benchmarks and surpass human performance in a selection of cases (compositional reasoning [25], for example). It has been shown, however, that

these methods exploit statistical regularities and biases in the data to achieve this performance [25, 33, 23, 6]. In addition, although these approaches are expected to merge information from multiple modalities, in practice they often exploit unimodal biases and ignore the other modalities entirely. In addition, particular signals in the input trigger specific answers; for instance, when the image contains a banana, the answer is most likely to be yellow, irrespective of the remainder of the image, or the question. This dependence on spurious correlations in the training data leaves VQA methods vulnerable to a failure to generalize. In addition, this phenomenon highlights the lack of high-level understanding of the input and its connection to other modalities.

To remedy the weaknesses identified above and improve generalization, we propose to utilize *counterfactuals* [30, 12] in the learning process. In traditional causal inference counterfactuals are *unobserved* scenarios, and are often used to estimate the effect of an intervention that is not directly represented in the data. In machine learning they can equally represent a potential training data element for which we do not have a label, or a data-label pair for which we do not have a reward. This is particularly relevant in those supervised learning settings where more than one true label might apply to each training data element, yet only one true answer is typically observed. This is the case in many vision-and-language problems, as the fact that the training set documents a particular answer to a VQA question does not mean that every alternate answer is wrong. This is referred to as *bandit feedback* [24], and such problems are labelled *nonstochastic multiarmed bandit problems* [9]. In the context of VQA, counterfactual analysis leads us to ask “*what would have happened if we observed a different image or asked a different question, given the past observations*”.

We consider the causal model underlying the training data, and introducing an extra (exogenous) variable that governs the question and image generation (from which the observed answers are produced). Then, we learn a distribution for that variable, providing a model of how the observational data was generated. Subsequently, we ask

“what would be the minimum alteration to the question or image that could change the answer”. To that end, we choose the exogenous variable such that the question or image generated using that variable yields an incorrect answer, thus effectively injecting an *intervention* into our causal model. Since the intervention can degrade the model’s performance, we “reason” about these counterfactual instances by formulating an alternative to conventional empirical risk minimization, allowing the model to learn both from observational and counterfactual instances. This implicitly forces the VQA model to use both input modalities instead of relying on statistical regularities specific to either of them. Further, training a model to both learn to answer and “reason” about the intervention in questions and images, encourages generalization. In Fig. 1, our approach is summarized.

By effectively “asking the algorithm” what would have happened, we aim to highlight the most interesting cases of disagreement between the counterfactuals and the training observations, while also demonstrating implicitly why the learned model is preferred.

We describe extensive experiments on VQA-CP [6], VQA 2.0 [7] and Embodied QA [14] (where agent requires navigation to answer questions) and demonstrate the ability of our approach to improve generalization. Our contributions in this paper are:

- We provide a counterfactual framework under which the interventions in the inputs, either the question or image, are anticipated. We show that a simple model of learning the distribution of an exogenous intervention variable of the observational data, and subsequently counterfactual samples generated from that variable improves generalization. We encourage the model to reason about “what the answer could be about a counterfactual image or question”.
- We provide a theoretical analysis for the proposed approach to shed light on its underlying working mechanism. In addition, we show a lower bound on the likelihood of the counterfactuals based on the observations.
- Our extensive experiments show that our simple yet powerful approach is capable of improving the generalization ability of diverse multimodal and unimodal vision and language tasks. In VQA-CP we observe more than 2% improvement over the baseline when using the full set and 7% when using a fraction of the dataset. In Embodied QA, our approach improves the state-of-the-art by more than 2%.

2. Related Work

Counterfactuals [12, 30] have gained recent interest in various areas in machine learning, in particular in applying insights from causal inference to augment the training as in

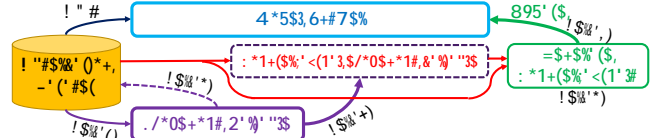


Figure 1: The training process with counterfactuals. We infer the posterior on the exogenous variables. Subsequently generate counterfactual samples using that variable and evaluate its output.

bandit settings [24, 2], reinforcement learning [10], recommendation [39] and explanation [19]. *Adversarial* learning [17] is a prime instance of use of counterfactuals in learning and was shown to improve performance (e.g. [42]). However, most of the state-of-the-art in this area focus on the analysis of the outcome of an intervention of sorts, i.e. change in the input or model. Our approach however, focuses on both proper generation of the counterfactuals from intervention and ensuring the outcome is adjusted in an alternative risk minimization.

Data Augmentation lies at the heart of successful machine learning where substantial domain knowledge is leveraged to design suitable data transformations (e.g. rescaling, rotation, etc) leading to improved generalization. While learning these invariances, using for instance generative models, can potentially alleviate the problem, their use is nontrivial.

Recently, MixUp [41] was proposed as a simple means for data augmentation and regularization which does not require significant domain knowledge. Similar to label smoothing, the supervision of every example is not overly dominated by the ground-truth label. Moreover, the augmented data is transformed from training instances to establish a linear relationship between data augmentation and the supervision signal. However, it requires sampling a mixing parameter that is not trivial to choose. Our approach on the other hand, learns to interpolate depending on the difficulty of producing its output for the model and the landscape in the feature space, hence harnessing the advantages of MixUp for sample generation.

Biases in VQA datasets and models are major pitfalls in current models where superficial correlations between inputs from one modality and the answers are exploited by models [29, 18, 33]. Unfortunately, biased models that exploit statistical shortcuts from one modality usually reach impressive accuracy on most of the current benchmarks. VQA-CP [6] is a recent diagnostic datasets containing different answer distributions for each question-type leading to different distribution of train and test splits. Consequently, models biased towards one of the modality often fail at this benchmark. Human provided additional balancing data, for instance in the case of VQA v2 [18] has not resolved the issue. More elaborate models to avoid biases

such as Grounded VQA [6] introduces additional submodules that are not trivial to be used with novel architectures. Similarly, [33] proposed a model-agnostic learning strategy to overcome language priors in VQA models by directly penalizing the input question-only bias. In [13], the authors cluster training questions using to their prefix to prevent the model from relying on them as features.

Our method is model-agnostic, easy to implement and does not need an elaborate parameter tuning or prior knowledge. In addition, our approach naturally leverages inherent dependencies to improve generalization and discourage simple exploitation of the biases by the model. Our counterfactual training approach discourages learning the biases by relying on the capacity to generate samples that can change the predictions.

2.1. Visual Question Answering

Visual Question Answering (VQA) is the task of answering previously unseen questions framed in natural language about a previously unseen image. For training, we are interested in learning a model from a training set made up of image v , question q and answer a triplets $D = \{q_i, v_i, a_i\}_{i=1}^n$. During test time, given an image and question, the trained model predicts the correct answer. The classical approach for VQA is to use an embedding of the questions $e^q = f_q(q)$, an embedding of the image $e^v = f_v(v)$ and a fusion function of the two $z = h(e^q, e^v)$ into what is known as the joint space. We denote by θ all of the parameters of the deep models used to learn these representations and generate answers. Using better embeddings yields better joint space representations and consequently more accurate answers. For brevity below we omit the parameters in the models, i.e. we use $p(a|q, v)$ as a shorthand for $p(a|q, v, \theta)$.

2.2. Counterfactuals

In the following we provide a background on counterfactuals that will form the basis for the rest of this paper. Interested readers are referred to [30] for further details

Definition 1 (Structural Causal Model (SCM)). A structural causal model M consists of a set of independent (exogenous) random variables $u = \{u_1, \dots, u_n\}$ with distribution $P(u)$, a set of functions $F = \{f_1, \dots, f_n\}$, and a set of variables $X = \{X_1, \dots, X_n\}$ such that $X_i = f_i(PA_i, u_i)$, i , where $PA_i = X \setminus X_i$ is the subset of X which are parents of X_i . As a result, the prior distribution $P(u)$ and functions determine the distribution P^M .

An SCM defines the data generating process and the distribution of the observations. Using this model, we can investigate the consequences of intervention.

Definition 2 (Interventional Distribution). For an SCM M , an intervention $I = \text{do } X_i := \tilde{f}_i(PA_i, u_i)$ corresponds

to replacing the structural mechanism $f_i(PA_i, u_i)$ with $\tilde{f}_i(PA_i, u_i)$. We can simply write $\text{do}(X_i = x)$ to denote the intervention. The resulting SCM is denoted M^I , and the resulting interventional distribution is denoted P^{M^I} .

We can also define the *counterfactual distribution* which tells us what might have happened had we acted differently.

Definition 3 (Counterfactual Distribution). Given an SCM M and an observed assignment $X = x$ over any set of observed variables, the counterfactual distribution $P^{M|X=x;I}$ corresponds to the distribution entailed by the SCM M^I using the posterior distribution $P(u|X = x)$.

For an SCM M , the counterfactual distribution can be estimated by first inferring the posterior over exogenous variables and then passing that distribution through the modified structural model M^I to obtain a counterfactual distribution over other variables¹.

3. Counterfactual Vision and Language (CVL)

Our intuition is that the functions that extract the features in a VQA system, either from the image or the question, are prone to focusing on spurious correlations in the data, which diverts them from modeling the deeper relations that generalize better. Hence, we encourage the learning algorithm to consider counterfactuals—a set of imaginary alternative samples. Training a model to both learn to answer, and “reason” about the intervention in the questions and images allows better generalization. To that end, we construct the SCM as shown in Fig. 2 where the functions for learning the embeddings are conditioned on the exogenous variables.

As is the convention for intervention in counterfactual reasoning, we are interested in replacing the embedding functions by their corresponding counterfactuals, that is, f_v is replaced by $\tilde{f}_v(v, u^v)$ or f_q by $\tilde{f}_q(q, u^q)$ where u^v and u^q are exogenous variables for image (vision module) and question (language module), respectively. Note that $\tilde{f}_v(\cdot, \cdot)$ and $\tilde{f}_q(\cdot, \cdot)$ are the functions of the exogenous variables for a given image and question pair. Effectively our approach reasons about the interventions in the embedding extractions. We use $u = [u^v, u^q]$ to denote both of the exogenous variables. We denote by \tilde{q} and \tilde{v} the variables obtained after the intervention and \tilde{e}^q and \tilde{e}^v as their corresponding embeddings. This intuitively allows our model to answer image-based questions it has never observed. We are generally interested in the following objectives: (1) the joint space of the question-image embedding must lead to a low-error rate on the factual data; (2) the conditional distribution of the factual and counterfactual data considering the exogenous distribution must be similar; (3) the distribution of the exogenous variables must be obtained from the observations; and (4) the embedding has to yield small error on

¹Called abduction, action, and prediction in [30]

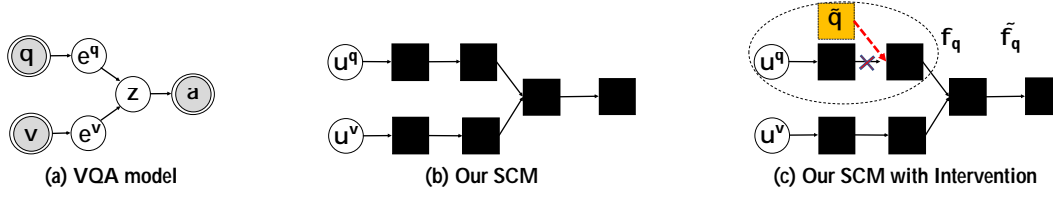


Figure 2: The difference between a typical VQA graphical model (in Fig. 2a), our corresponding causal model (in Fig. 2b) and an example of intervention in the question representation of this model (in Fig. 2c). In our model two exogenous variables u^q and u^v are incorporated to learn and reason about the intervention caused by these variables.

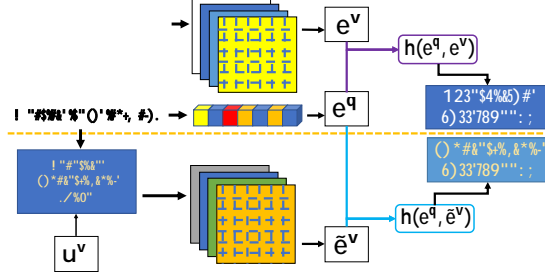


Figure 3: Counterfactual examples that can be generated for a given image. The fusion function h is used with the observational data as well as the counterfactual data to predict the answer. For the counterfactual loss, we need to consider the relationship between the predicted counterfactual answer and its observational counterpart.

the unobserved counterfactual distribution (obtained from the intervention in the structural model).

The first objective is the same as any other vision and language task. The second is a necessary constraint to ensure using a model from the observations we can predict answers for counterfactuals. The third objective ensures the possible intervention distribution from the exogenous variable is learned as part of the model. Lastly, our approach should be able to reason about the answer to the counterfactual instances (see Fig. 3 for an example). As such, we devise the following steps through which our model is trained and the distribution of the exogenous variable is found:

1. Infer the predictive model for the observed data using one step of the conventional risk minimization.
2. Perform intervention I on M . This yields M^I , which entails the counterfactual distribution $p^{do(I)}(\bar{q}, \bar{v})$.
3. Reason about the effect of that intervention on the answer and the loss that incurs.

Intuitively, first we learn what distribution of the exogenous variable is obtained from the observations, then model how the answer is affected by the intervention on this variable.

4. Counterfactual Distribution

The counterfactual distribution is the posterior of the exogenous variables obtained from the observations. Hence, using the training data we are interested in²

$$p(u|D) = p(u) \prod_{i=1}^n p(a_i|q_i, v_i) p(v_i|u^v) p(q_i|u^q). \quad (1)$$

We use independent priors, i.e. $p(u) = p(u^q)p(u^v)$ with Beta distributions for u^v and u^q (i.e. $u^v \sim \text{Beta}(\alpha, \beta)$). Although we could estimate $p(v_i|u^v)$ and $p(q_i|u^q)$ using various methods (including autoencoders [27, 1] and GANs [16, 4, 3]), we use a simple approach to model the question or image’s conditional likelihood. To obtain the posterior, considering the generating process of q_i and v_i for a given sample of the variable u^q, u^v and an arbitrary constant $0 < 1$, we have

$$p(q|u^q) = \begin{cases} q & u^q = 1 \\ u^q q & (1 - u^q)q, \text{ otherwise} \end{cases}, \text{ and} \quad (2)$$

$$p(v|u^v) = \begin{cases} v & u^v = 1 \\ u^v v & (1 - u^v)v, \text{ otherwise} \end{cases}$$

where q and v are uniformly sampled at random from the dataset and \cdot denotes an interpolation. It is easy to see that for 0 we have more interpolated samples and for 1 , we obtain samples that are independent of the prior. An advantage of this approach of sampling the observations is that we effectively reduce the conditional independence assumption of the training data allowing for the relation between observations to be established.

Since we use all conjugate priors, the posterior is also a Beta distribution with parameters α, β , where $\alpha = \alpha_0 + I[a_i = \arg \max p(a_i|q_i, v_i)]$ and $\beta = \beta_0 + I[a_i \neq \arg \max p(a_i|q_i, v_i)]$. Intuitively, samples from the regions of the prior that produce the correct answers are “successful” and encourage the posterior to concentrate. Notice that the samples from the posterior are drawn from the regions where the likelihood of the correct answer is higher (since the expectation of the posterior is $\alpha / (\alpha + \beta)$).

²We note that without loss of generality and for brevity we drop the dependence on the embedding features $p(a_i|q_i, v_i) = p(a_i|e^v, e^q) \times (e^q - f_q(q_i)) \times (e^v - f_v(v_i))$ where δ is the Dirac delta.

4.1. Generating Counterfactuals

Once the posterior on the exogenous variables $p(u|D)$ is obtained, we perform the intervention. That is, we generate the counterfactuals and replace the v (or q) with its alternative \tilde{v} (or \tilde{q}) and anticipate the answer. This corresponds to replacing the function $f_v(\cdot, \cdot)$ (or $f_q(\cdot, \cdot)$) with an alternative $\tilde{f}_v(\cdot, \cdot)$ (or $\tilde{f}_q(\cdot, \cdot)$) which leads to a different answer prediction.

In obtaining the counterfactual samples we are interested in the minimum interventions that will change the answer for a given question-image pair (q, v) to (\tilde{q}, \tilde{v}) when using the generating process in Eq. 2. This corresponds to a sample from the posterior of the exogenous variable with high likelihood (minimum intervention) that will alter the answer for (q, v) to an incorrect one. As such, we formalize the problem as:

$$\begin{aligned} \max_u \quad & \log(p^{\text{do}(I)}|_{q,v}(\tilde{q}, \tilde{v}|u)) \\ \text{s.t.} \quad & \tilde{a} = \operatorname{argmax}_a p^{\text{do}(I)}|_{q,v}(a|\tilde{q}, \tilde{v}), \quad \tilde{a} \neq a \\ & 0 \leq u < 1 \end{aligned}$$

Considering the generative process in Eq. 2, the minimum intervention (the minimum edit of the factual [32, 19]) is achieved when u is largest. Since the constraint is not computationally feasible, we relax the objective and choose the variable that has the minimum likelihood of having the same answer as the observations. Thus, we choose u from the relaxed alternative (we project u to be bounded in $[0, 1]$)

$$\max_u \quad u^2 - \log p^{\text{do}(I)}|_{q,v}(a|\tilde{q}, \tilde{v}) \quad (3)$$

where γ is a hyper-parameter. We note that simply sampling from the posterior $p(u|D)$ and generating v (or q) to infer the answer, is not the counterfactual (alternating between sampling the variable u and learning parameter resembles conventional Gibbs sampling). Hence, this step is critical to obtain instances that are *not* merely from the learned distribution, yet very likely. Consequently, enabling our approach to generalize better beyond observations.

4.2. Counterfactual Loss

We alternate between intervening in the inputs, and minimizing the risk on the corresponding counterfactual along with the observations. As is common practice in empirical risk minimization (ERM), the objective in using observational training instances is minimizing $E_{q,v} E_{p(a|q,v)}[(f(q, v), a)]$ where $(f(q, v))$ is the loss of the function predicting the answer. Note that in practice f and $p(a|q, v, \cdot)$ may be the same function or share architecture (e.g. $p(a|q, v, \cdot) = \text{softmax}(f(q, v))$). In the case of using counterfactuals, we can rewrite the risk by changing the distribution [12]:

$$\begin{aligned} R(\cdot) &= E_{q,v} E_{p(a|q,v)}[(f(q, v), a)] \\ &= E_{q,v} E_{p^{\text{do}(I)}|_{q,v}(a|\tilde{q}, \tilde{v})}[(f(q, v), a)] \frac{p(a|q, v, \cdot)}{p^{\text{do}(I)}|_{q,v}(a|\tilde{q}, \tilde{v}, \cdot)} \end{aligned}$$

Note that $p^{\text{do}(I)}|_{q,v}(a|\tilde{q}, \tilde{v}, \cdot)$ has part of SCM altered. Intuitively, the counterfactuals that have smaller scores are more penalized and conversely the over-confident ones are discouraged. This subsequently adjusts the decision boundary to be discriminative for both observations and counterfactuals. Furthermore, since this risk can have a very high variance we can clip this value similar to [12],

$$R^M(\cdot) = E_{q,v} E_{p^u(a|q,v)}[(f(q, v), a)] \times \min_i M_i \frac{p(a|q, v, \cdot)}{p^{\text{do}(I)}|_{q,v}(a|\tilde{q}, \tilde{v}, \cdot)}$$

This is because we may have very low probability in predicting an output of an intervened observation. Thus, the empirical counterfactual risk is,

$$\hat{R}^M(\cdot) = \frac{1}{n} \sum_{i=1}^n (f(q_i, v_i), a_i) \times i(\cdot) \quad (4)$$

$$\text{where } i(\cdot) = \min_i M_i \frac{p(a_i|q_i, v_i, \cdot)}{p^{\text{do}(I)}|_{q,v}(a|\tilde{q}, \tilde{v}, \cdot)}.$$

Here, $i(\cdot)$ is the clipped ratio of evaluation of the factual sample i and its corresponding counterfactual. We intentionally use a shorthand to underscore the fact that the parameters are optimized with respect to γ in p . The objective of the counterfactual risk minimization for vision and language tasks is therefore

$$\hat{R}^M = \operatorname{argmin} \hat{R}^M(\cdot)$$

In practice, we alternate between the conventional ERM (i.e. when $\gamma = 1$) and the counterfactual risk.

4.3. Further Analysis

When we generate samples in Eq. 2, q is likely to have a different answer to q (with probability $(1 - n_a/n)$ for n_a denoting the number of instances with answer a). As such, interpolating between the questions and images will lead to samples for which the answer is uncertain. In the case of the generated counterfactuals, however, such interpolations are in fact close to the decision boundary. Hence, when weighted by the confidence of the classifier in Eq. 4, the connection between samples in the fusion space (i.e. the common semantic space) is adjusted to account for the sensitivity of the representations to changes in the input.

Furthermore, one main question is how do we know that the interventions won't lead to divergence, or learning useless models. We can derive the bound on the risk using the following theorem:

Theorem 4. Denote $u^i(\cdot) = (f(q_i, v_i), a_i) \cdot i(\cdot)$, $\bar{u} = \frac{1}{n} \sum_{i=1}^n u^i(\cdot)$, $\hat{V}(u) = \frac{1}{n} \sum_{i=1}^n u^i(\cdot) - \bar{u}^2 / (n-1)$ and $Q = \log(10 \cdot \gamma / \epsilon)$ for $0 < \epsilon < 1$ and the γ -cover for the function class that predicts the answer. With probability at least $1 - \epsilon$ for $n \geq 16$ we have

$$R(\cdot) \leq \hat{R}^M(\cdot) + \frac{1}{18\hat{V}(u)Q} \gamma / n + 15MQ \gamma / (n-1)$$

Proof. See the supplementary material. \square

This result implies that when we have the counterfactual risk minimized, we achieve the minimum variance.

We note that we can compute the density of the counterfactuals based on the observations, i.e.

$$p^{\text{do}(I)}(\tilde{q}, \tilde{v}) = E_{(q,v)} p_{(q,v)} p^{\text{do}(I)|q,v}(\tilde{q}, \tilde{v}) \quad (5)$$

This result shows that the density of intervened variables (\tilde{q}, \tilde{v}) is the marginal of the observations. Hence, the factual, counterfactual and exogenous variables are connected with the following lemma:

Lemma 5. *We have the following lower bound on the log-density of the counterfactuals:*

$$\log(p^{\text{do}(I)}(a, \tilde{q}, \tilde{v})) = E_{(q,v)} p_{(q,v)} \log(p^{\text{do}(I)|q,v}(a|\tilde{q}, \tilde{v})) + E_u p_{(u)} \log(p^{\text{do}(I)}(\tilde{q}, \tilde{v}|u)) .$$

Proof. See the supplementary material. \square

In fact we can show that even if u is not drawn from the true generating prior, we can use an arbitrary distribution q and obtain an alternative lower bound to that of Lemma 5:

$$\log(p^{\text{do}(I)}(a, \tilde{q}, \tilde{v})) = E_{(q,v)} p_{(q,v)} \log(p^{\text{do}(I)|q,v}(a|\tilde{q}, \tilde{v})) + E_q [\log(p^{\text{do}(I)}(\tilde{q}, \tilde{v}|u))] + H(q) - H_q(p). \quad (6)$$

Effectively using Lemma 5, we know even if the distribution of the exogenous variable for generating the counterfactuals deviates from the true posterior obtained from observations, we can lower-bound the marginal of the counterfactuals which depends on the likelihood of predicting the correct answer, the difference of entropy of the true prior versus the one used and the likelihood of the counterfactual examples.

5. Experiments

To evaluate the performance of our approach, we construct experiments on various datasets. We note that our approach is agnostic to the base model used and as such is widely applicable to a wide range of applications. To optimize the objective in Eq. (3), we use a simple gradient ascent where we set the learning rate to a constant. We use prior for the exogenous variable as Beta(0.1, 0.1) for the experiments unless otherwise stated. We alternate between the observational training and the counterfactuals.

5.1. Unimodal Problems

The motivation of our approach is multimodal problems, but it is equally effective for problems involving only a single modality. In this case the description of the process

	LSTM	T	LSTM+P	T+P	LSTM+C	T+C
Random	84.4	82.0	84.53	85.21	85.61	85.56
GloVe	84.9	86.4	85.77	87.1	87.24	88.4

Table 1: Accuracy (%) obtained by the testing methods using LSTM (with randomly initialized, trainable embeddings). Best results highlighted in Bold. T abbreviates TreeLSTM [40]; +P and +C indicate posterior and Counterfactuals respectively.

stands, with the exception that either u^v or u^q is inferred and used for counterfactual generation.

Stanford Sentiment Treebank (SST) [38] is a natural language dataset of movie reviews (neutrals are removed in our experiments). This dataset contains 11855 instances with vocabulary size of 17836 and 5 classes. We follow the implementation of [40] where a tree structured LSTM is used with this dataset. We use two alternative baselines for embedding words to be used when sampling in Eq. (2): random embedding and trainable GloVe [31] initialized word embeddings. We report mean scores over 5 runs and use 10 epochs for training. Here we examine how the change in the embedding representation effects the performance of the model. Since we don't have the image input, we only infer u^q with prior Beta(0.1, 0.1) and the counterfactual learning rate is set to 0.01. As shown in Table 1 using either the posterior (+P models) or the optimized exogenous variable (+C) from Eq. (3) improves algorithm accuracy. As expected, when pretrained models are tuned, the overall performance is better.

We further evaluate the generalization performance of our approach when only the visual data is available on the **CIFAR-10 and CIFAR-100** image classification datasets. In particular, we compare the baseline architectures for: VGG-19 [35], ResNet-18 [21], ResNet-101 [20], and DenseNet [22]. All models are trained for 100 epochs on the training set with 128 examples per minibatch and learning rate 0.1, using SGD and evaluated on the test set. The learning rate is then reduced to 0.001 for an additional 150 epochs. We use the interpolations in the input images for Eq. (2). In the experiments we have not observed any noticeable

difference (see the supplementary material for additional results). We set the prior of u^v to Beta(0.1, 0.1) and run the counterfactual optimizer for 10 iterations.

We summarize our results in Table 2. In both CIFAR-10 and CIFAR-100 classification problems, the models trained

(a) Values of u^v in Training

(b) Variance of Loss

Figure 4: Training metrics in CIFAR experiments.

Dataset	Model	Baseline	Ours+P	Ours+C
CIFAR-10	VGG-19	95.04	95.92	96.73
	ResNet-18	93.02	94.2	94.91
	ResNet-101	93.75	94.1	95.34
	DenseNet-121	95.04	95.92	96.73
CIFAR-100	VGG-19	72.23	73.45	74.8
	ResNet-18	75.61	76.5	77.75
	ResNet-101	77.78	78.9	80.0
	DenseNet-121	77.01	79.67	79.67

Table 2: Test errors for the CIFAR experiments.

Model	Overall	Yes/No	Number	Other
Question-Only [6]	15.95	35.09	11.63	7.11
RAMEN [34]	39.21	-	-	-
BAN [26]	39.31	-	-	-
MuRel [11]	39.54	42.85	13.17	45.04
UpDn [8]	39.74	42.27	11.93	46.05
UpDn+Q-Adv+DoE [33]	41.17	65.49	15.48	35.48
UpDn+C Images	41.01	44.61	12.38	46.11
UpDn+C Questions	40.62	42.33	14.17	48.32
UpDn+C (Q+I)	42.12	45.72	12.45	48.34

Table 3: State-of-the-art results on VQA-CP test. **UpDn+C** indicates our approach based on UpDn baseline. **(Q+I)** denotes both question and images are intervened.

using our approach consistently improve on the baselines by a margin. As seen in Fig. 4, the variance is also reduced during training which, as discussed in Theorem 4, is an indication of the convergence of counterfactual training. As observed, the values of u^v decreases over time to find the samples that are harder to predict. Our experiments thus indicate that our approach provides improvements to even unimodal problems.

5.2. Multimodal Problems

Visual Question Answering is used to evaluate our model with two datasets: VQA-CP [6] and VQA v2 [18]. VQA-CP is specifically designed to measure the generalization ability of VQA models. Since our model learns how the data is generated, we expect it to be particularly robust towards bias. We follow the same training and evaluation protocol as [8] (see the supplementary material for implementation details). For each model, we report the standard VQA accuracy metric [7]. In this experiment, we interpolate the word/visual embeddings rather than actual inputs to generate counterfactuals.

In Table 3, we compare our approach consisting of our baseline architecture trained with additional counterfactual training on VQA-CP against the state-of-the-art. To be fair, we only report approaches that use the visual features from [8]. Our approach improves the baseline more than 2 percentage point beyond UpDn+Q-Adv+DoE which regularizes the model for better performance. In addition, our approach gains most from the “other” category that encom-

pass the most valuable improvement indicating better reasoning about the answers. We should note that since our approach is architecture agnostic, we expect more against better baselines.

Ablation Study on Modality Intervention: In Table 3, we perform an ablation study of learning to intervene in multimodal problems by only either inferring u^q (i.e. intervention in the question) or u^v (intervention in the images). Even though intervening in both u^q, u^v improves performance, counterfactual questions lead to better “number” results indicating strong bias in the baseline for questions with number answers.

Smaller Training Sets: As shown in Fig. 6, when the number of training instances is smaller our approach achieves significantly better performance compared to the baseline. This is due to our approach being able to exploit the alternative instances with counterfactuals.

Impact on VQA v2: We use the standard VQA v2 dataset [18] by following the implementation in [37, 36]. Since by exploiting statistical regularities in this dataset it is easier to achieve better performance, large gains are not expected. As shown in this section, counterfactual samples improve the accuracy in VQA-CP, while marginally improving in VQA v2 compared to its baseline. It is interesting to note that in adversarial training in UpDn+Q-Adv+DoE, the performance drops in VQA v2 indicating the same phenomenon.

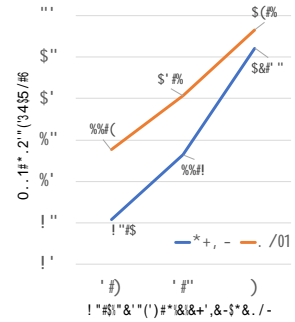


Figure 6: The performance of our approach vs. the baseline using fraction of the training data.

In Fig. 5 we show samples of the counterfactuals for the given question-image pairs from the test set. These samples are generated by following Eq. (2) (i.e. randomly sampling another question-image pair and interpolating the embeddings using

the samples from the posterior) and subsequently finding the closest instances (either question or image with smallest Euclidean distance in the embedding space) in the test

Model	Overall
Question-Only [6]	25.98
BAN [26]	69.08
MuRel [11]	65.14
UpDn [8]	63.48
UpDn+Q-Adv+DoE [33]	62.75
Pythia [37]	68.49
Pythia+C	68.77

Table 4: Performance of our approach on VQA v2 validation. **Pythia+C** is our counterfactual implementation of [37].

! "\$%&'()*+,-#	. ' " (%#/0,1%,2)!"#\$%&'(\$. ' " (%#/0,1%,2)*+,-#&
!"#\$%&'()*+,-#	. /#!"#\$%&'()*+,-#1"00' - 2/#!"#\$%&'()*+,-#3"001"3) '4 - 5/#("0#\$%&'()*+,-#4! '13"001"00' #!"#\$%&'()*+,-# 7/0,4 - 8/#!"#\$%&'()*+,-#39#11%00' -	
: %&'()*+,-#3"001"3) '4 - %0#\$%&'()*+,-#3"001"3) '4 - %0, ? 0\$-	. /# : %&'()*+,-#3"001"3) '4 - >0+1"1 - 2/# : %&'()*+,-#3"001"3) '4 - >0+1"1 - 5/# : %&'()*+,-#3"001"3) '4 - 8/# : %&'()*+,-#3"001"3) '4 -	
: %0"0414# %0#"%+43># 3"001"3) '4 - ;3?0#?"3? -	. /# : %&'()*+,-#437431# "001"3) '4 - 2/# : %&'()*+,-#90437431# "001"3) '4 - 5/# : %&'()*+,-#437431# "001"3) '4 - 8/# : %&'()*+,-#00437431# "001"3) '4 -	

Figure 5: Given the image-question pair in the first column, the closest instances of the questions (in second column) and images (in the third column) are found from the VQA v2 test dataset corresponding to the generated counterfactuals (using the exogenous variables).

set. As observed, some of the questions are reasonable alternatives to the ones asked and conversely, the given question can be asked of the counterfactual images showing that our approach successfully generates alternatives.

Embodied Question Answering (EQA) [14] is proposed as a novel variant of VQA where an agent is spawned at a random location in a 3D environment and asked a question for which the answer requires exploration in the environment. We closely follow the instructions of [14] for the experimental setup. Similar to VQA, the agent is tasked with utilizing both vision (i.e. the input ego-centric RGB image from the robot’s camera) and language (i.e. the given instructions) to answer questions. However, a distinct feature of this task is, unlike VQA, the final answer is produced after the agent takes a finite number of intermediate actions (i.e. navigation by choosing the action right, left, straight, stop at each step for which we use a 2-layer GRU to predict). During training, each batch contains a random environment, a question in that environment and its corresponding answer along with the path to reach the corresponding location in the target room.

In our approach, we intervene in both the image and question embeddings using a randomly sampled environment and question to generate counterfactual instances in Eq. (2). We set the prior for the exogenous variables u^q and u^v to Beta(0.75, 0.75). We trained the model based on shortest path trajectories to target objects inside 640 houses (total 6,912 questions) for 30 epochs and then evaluated it on 57 unseen environments during the inference. In particular we consider three cases which correspond to being 10, 30 and 50 steps away from the target room, with the distance corresponding to 0.94, 5.47 and 10.99 respectively. In this experiment we measure the number of correct intermediate steps that the agent correctly takes to increase its proximity to the room with the answer. The results are shown in Table 5. As is shown, our approach of allowing the agent to contemplate counterfactual questions and images enables the robot to travel closer to the target room and improves

Model	d _T Lower is better			d Higher is better		
	T ₋₁₀	T ₋₃₀	T ₋₅₀	T ₋₁₀	T ₋₃₀	T ₋₅₀
PACMAN	1.39	4.98	9.33	-0.45	0.49	1.66
[14]						
GRU [14]	0.74	3.99	8.74	0.20	1.48	2.26
GRU+C	0.67	3.90	8.47	0.26	1.57	2.52

Table 5: Evaluation metrics for EQA navigation. Spawning the agent 10, 30, or 50 steps away from the target location, d_0 shows the distance between these initial locations and the target location, while d_T reveals the distance of the final locations and the target ones by starting from these initial location and using the model for the maximum of 100 steps. Finally, $d = d_T - d_0$ measures the overall progress of the agent towards the target. GRU+C is ours.

generalization. This further illustrates our approach’s success in improving generalization in various tasks and input-output alternatives. Note that in this task the output is a sequence of actions to be predicted (before the answer).

6. Conclusion

The tendency to focus on spurious correlations in the training data is one of the key factors limiting the practical application of modern machine learning methods. We have shown that this failure to generalize can, in part, be tackled by generating a set of counterfactual examples to augment the training data. This is motivated by the success that the counterfactual approach has had in causal reasoning. We have demonstrated the effectiveness and generality of the proposed approach on a wide variety of problems including multimodal vision-and-language tasks. An additional advantage of the method is that the sample generation strategy relieves the conditional independence assumption of the training data, which is too strong for most real datasets.

Acknowledgement: This work is partly supported by DP160100703.

References

- [1] Ehsan Abbasnejad, Anthony R. Dick, and Anton van den Hengel. Infinite variational autoencoder for semi-supervised learning. In *CVPR*, pages 781–790. IEEE Computer Society, 2017.
- [2] Ehsan Abbasnejad, Justin Domke, and Scott Sanner. Loss-calibrated monte carlo action selection. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [3] Ehsan Abbasnejad, Javen Shi, and Anton van den Hengel. Deep lipschitz networks and dudley gans. 2018.
- [4] Ehsan Abbasnejad, Qinfeng Shi, Anton van den Hengel, and Lingqiao Liu. A generative adversarial density estimator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [5] Ehsan Abbasnejad, Qi Wu, Qinfeng Shi, and Anton van den Hengel. What’s to know? uncertainty as a guide to asking goal-oriented questions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [6] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [7] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. VQA: visual question answering. *Int. J. Comput. Vis.*, 123(1):4–31, 2017.
- [8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
- [9] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.
- [10] Lars Buesing, Theophane Weber, Yori Zwols, Nicolas Heess, Sebastien Racaniere, Arthur Guez, and Jean-Baptiste Lespiau. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *International Conference on Learning Representations*, 2019.
- [11] Remi Cadene, Hedi Ben-Younes, Nicolas Thome, and Matthieu Cord. Murel: Multimodal Relational Reasoning for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2019.
- [12] Denis Charles, Max Chickering, and Patrice Simard. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14:3207–3260, November 2013.
- [13] Anton van den Hengel Damien Teney, Ehsan Abbasnejad. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*, 2020.
- [14] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [15] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.
- [17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR 2015 : International Conference on Learning Representations 2015*, 2015.
- [18] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vision*, 127(4):398–414, April 2019.
- [19] Yash Goyal, Ziyang Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384, Long Beach, California, USA, 09–15 Jun 2019. PMLR.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 2016.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016.

- [22] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Computer Vision and Pattern Recognition*, 2017.
- [23] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [24] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.
- [25] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.
- [26] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems 31*, pages 1571–1581, 2018.
- [27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.
- [28] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc., 2015.
- [29] Varun Manjunatha, Nirat Saini, and Larry S. Davis. Explicit bias discovery in visual question answering models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9562–9571, 2019.
- [30] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.
- [31] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [32] Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- [33] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems, NIPS’18*, pages 1548–1558, USA, 2018. Curran Associates Inc.
- [34] Robik Shrestha, Kushal Kafle, and Christopher Kanan. Answer them all! toward universal visual question answering models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [36] Amanpreet Singh, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia-a platform for vision & language research. In *SysML Workshop, NeurIPS*, volume 2018, 2018.
- [37] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.
- [38] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.
- [39] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 814–823, Lille, France, 07–09 Jul 2015. PMLR.
- [40] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics.

- [41] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018.
- [42] Ruixiang Zhang, Tong Che, Zoubin Ghahramani, Yoshua Bengio, and Yangqiu Song. Metagan: An adversarial approach to few-shot learning. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS’18, pages 2371–2380, USA, 2018. Curran Associates Inc.