

Unshuffling Data for Improved Generalization in Visual Question Answering

Damien Teney Ehsan Abbasnejad Anton van den Hengel
Australian Institute for Machine Learning
The University of Adelaide
Adelaide, Australia

{damien.teney, ehsan.abbasnejad, anton.vandenhengel}@adelaide.edu.au

Abstract

Generalization beyond the training distribution is a core challenge in machine learning. The common practice of mixing and shuffling examples when training neural networks may not be optimal in this regard. We show that partitioning the data into well-chosen, non-i.i.d. subsets treated as multiple training environments can guide the learning of models with better out-of-distribution generalization. We describe a training procedure to capture the patterns that are stable across environments while discarding spurious ones. The method makes a step beyond correlation-based learning: the choice of the partitioning allows injecting information about the task that cannot be otherwise recovered from the joint distribution of the training data.

We demonstrate multiple use cases with the task of visual question answering, which is notorious for dataset biases. We obtain significant improvements on VQA-CP, using environments built from prior knowledge, existing meta data, or unsupervised clustering. We also get improvements on GQA using annotations of “equivalent questions”, and on multi-dataset training (VQA v2 / Visual Genome) by treating them as distinct environments.

1. Introduction

The best of machine learning models can sometimes be right for the wrong reasons [2, 21, 26, 69]. The ubiquitous paradigm of empirical risk minimization (ERM) produces models that capture all statistical patterns present in the training data¹. However, not all of these patterns are reliable and reflective of the task of interest. Some of them result from confounding factors, sampling biases, and other annotation artifacts specific to a given dataset. We call these patterns *spurious*² and a model that relies on them will gen-

¹ We use *patterns* and *correlations* interchangeably to refer to statistical relationships between observed random variables, typically the input(s) and output(s) of a supervised learning task.

² The literature also uses *dataset biases* to refer to spurious correlations between inputs and outputs that are dataset-specific [69].

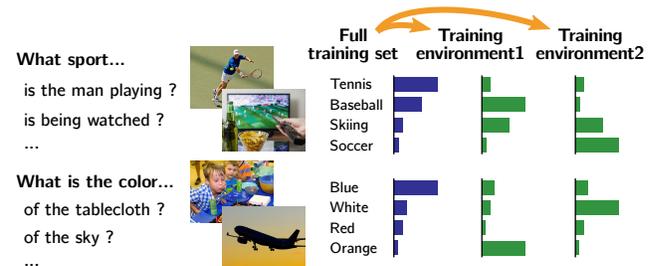


Figure 1. Datasets for visual question answering contain biases and spurious correlations: the first few words of a question are associated with a peaky distribution over answers (blue histograms). Models that guess their answers using these correlations generalize poorly. We improve by **partitioning the data** into multiple training environments across which the spurious correlations vary (green histograms) while reliable correlations are stable. Our training procedure produces a model that relies on these stable correlations such that it generalizes much better at test time.

eralize poorly to test data obtained in different conditions (*i.e.* out-of-distribution or OOD data).

The **limits of ERM** on OOD data are oftentimes overlooked and eclipsed by the common practice of evaluating on test data i.i.d. to the training data – a central assumption of classical learning theory [70]. The awareness of these limits has grown with that of their practical implications, from poor transfer across datasets [69] to biases and fairness issues [1] and vulnerability to adversarial inputs [27]. As a result, benchmarks with OOD test sets are becoming increasingly common in vision and NLP [2, 4, 9, 29, 37, 44]. This paper presents a training paradigm to improve OOD generalization.

The study of **generalization in computer vision** has a long history [20, 41, 48]. The rise in popularity of high-level tasks like visual question answering (VQA) [7], visual dialogue [19], or vision-and-language navigation [6] has made the topic even more important. The complexity of these tasks and the combinatorial explosion of the size of their input domain make it impossible to process

training data densely spanning this space. Models trained with ERM are then more likely to latch on spurious correlations (because they are often easier to fit [58]) rather than on the true reasoning process that underlies the task [36]. VQA was shown empirically to be a prime example of this issue. Many methods have been proposed to address it [2, 13, 17, 27, 29, 47, 55].

We propose a **general method to improve OOD generalization**. We discourage the model from using spurious correlations that only appear in subsets of the training data, and rather ensure that it uses reliable ones that are more likely to generalize at test time. More precisely, we first partition the data into multiple training environments [8] such that spurious correlations vary across environments while reliable ones remain stable. We later describe multiple strategies to build such environments, using unsupervised clustering, prior knowledge, and auxiliary annotations in existing datasets. Second, we train multiple copies of a neural network, one per environment. Some of their weights are shared across environments, while others are subject to a variance regularizer in parameter space. This leads the model to extract features that are stable across environments (*i.e.* features that do not represent environment-specific properties) since they are optimized to be predictive under a classifier common to all environments (as encouraged by the variance regularizer). Additional intuitions and reasons why this approach is superior to ERM are discussed in Section 3.3.

We provide **empirical evidence of improvements** in three distinct use cases on the task of VQA. First, we demonstrate improved resilience to language biases with VQA-CP [2]. Second, on GQA [34], we show how to use annotations of equivalent questions (some training questions being rephrasings of others). We obtain substantial gains over simple data augmentation with these equivalent questions in a small-data regime. Third, we show a small benefit in training a model on multiple datasets by treating VQA v2 [26] and Visual Genome QA [42] as two environments rather than one aggregated dataset. Of these three use cases, the first is the most well-studied but **our method has a much wider scope than the VQA-CP dataset**.

The contributions of this paper are summarized as follows.

1. We propose to partition existing datasets into training environments to improve generalization. We describe the requirements for the partitioning and a procedure to train neural networks to rely on stable correlations across environments while ignoring spurious ones.
2. We apply the method to three use cases with the task of VQA: (1) resilience to language biases, (2) leveraging known relations of equivalence between specific training questions, and (3) multi-dataset training.
3. We provide empirical evidence of clear improvements and a extensive sensitivity analysis to hyperparameters

and implementation choices.

2. Related work

This paper touches on fundamental aspects of machine learning and so is related to many existing works.

Dataset biases in vision-and-language tasks. Several popular datasets used in vision-and-language [25] and natural language processing [76] have been shown to exhibit strong biases. A model trained naively on these datasets can exhibit surprisingly good performance by relying on dataset-specific biases without capturing the true mechanisms of the task. On the evaluation side, there is a trend towards out-of-distribution test data to better identify this behaviour (*e.g.* [2, 4, 9, 29, 37, 44, 76]).

Invariances and generalization. Resilience to dataset biases cannot be solved by simply collecting more data from the same distribution, since it would still contain the same unreliable patterns. The data collection process can be improved [25, 76, 77] but this option only addresses precisely identified biases and confounders. Improving generalization requires to bring in information (often implicitly) about the mechanisms of the task of interest that go beyond what is represented by the joint distribution of the training data (see discussion Section 3.3). Common methods include architecture design and data augmentation [43, 63] to specify input transformations the model should be invariant to. This helps ignoring spurious correlations and improves generalization but it requires explicit knowledge of desirable invariances. In comparison, our method discovers invariances implicitly. We train a model to rely on input features that are similarly predictive (*i.e.* invariant) across environments. The required expert knowledge is displaced to the specification of the partitioning into environments.

Aggregating datasets. Using training data collected in different conditions is often beneficial for generalization (*e.g.* in [66] for VQA) because biases in each dataset are likely different and they can “cancel each other out”. We argue however that treating aggregated datasets as a collection of samples from a unique distribution loses valuable information. Our method treats them as distinct training environments and seeks to identify patterns that are similarly predictive across them while discarding those that are dataset-specific. Khosla *et al.* [41] also showed that accounting for bias when combining datasets was beneficial for generalization.

Domain adaptation and domain generalization. Training a model under multiple environments is reminiscent of domain adaptation [22]. But our objective is not to adapt to one particular target domain but rather to generalize across a range of unknown conditions. Our setup is more similar to the recently-introduced terminology of domain generalization [72, 28, 72]. These methods specifically target image recognition benchmarks like PACS and VLCS.

Ensembles. Our method trains multiple copies of a model in parallel, which is superficially similar to ensembling [79] and bootstrap aggregation a.k.a. bagging [12]. Traditional ensembles however combine models in output space. We combine models in parameter space³. We show empirically that the improvements from our approach are distinct from (and complementary to) those of traditional ensembling. Bagging uses uniform sampling, whereas the point of our method is to exploit prior knowledge to build the training environments.

Robustness in VQA. State-of-the-art models for VQA have been shown to be strongly reliant language biases. Benchmarks have been designed to better study the issue [2, 34, 38]. VQA-CP [2] allows out-of-distribution evaluation, where the joint distribution of questions and answers is different in the training and test sets. Methods have been proposed with strong improvements on VQA-CP ([13, 17, 27, 29, 47, 55] among others). However, many were shown to cheat the evaluation by exploiting knowledge of the construction of VQA-CP that should have been kept private [67]. In [65], the authors use counterfactual examples to learn which input features to focus on. This is comparable to our method in that we learn by contrasting training environments, whereas [65] proceeds at the instance level. In [64], the authors exploited auxiliary annotations of the GQA dataset to improve robustness, which we also use in some of our experiments.

Fair and bias-resilient machine learning. Addressing dataset biases is also motivated by issues of fairness [1, 31, 73, 78]. These works aim to build predictive models that are invariant to specific attributes of the input, such as gender or ethnicity. These attributes need to be specified and annotated, which is very limiting. For example in VQA, there is a known desired invariance to some linguistic patterns in the question, but their exact form is not known and cannot be annotated as a discrete attribute.

Invariant risk minimization. This paper is strongly inspired the principle of invariant risk minimization (IRM) [8]. Arjovsky *et al.* showed how training under multiple environments can improve OOD generalization. Our contributions over [8] are threefold: an alternative, easy-to-train implementation, an application to a real large-scale dataset, and demonstrations of how to obtain training environments. Other implementations were proposed [3, 14, 16, 35] but mostly demonstrated on toy data. Theoretical and empirical comparisons of variants of IRM are missing pieces to address in the future. Creager *et al.* [18] attempted unsuccessfully to discover environments automatically for IRM. Their failure, even on toy data, supports the view that the true source of improvements is in the information used to obtain or create the environments.

³ Averaging predictions or final weights is equivalent with a linear classifier. It is not in our case of a non-linear output.

Learning from groupings of data has been studied from statistical [11, 30] and causal [49, 57] points of view. Heinze-Deml and Meinshausen [30] used a variance regularizer on predictions across versions of an example, such as multiple photos of a same individual. Our variance regularizer, in comparison, acts on the parameters of the model. And we do not require correspondences between specific training examples. [20] is another classic work using a variance regularizer for multi-task learning. And Vasilescu *et al.* [71] were among the first to compute invariant representations for the causal factors of image formation, which they did for face recognition and human motion signatures.

3. Proposed approach

3.1. Partitioning data into training environments

The main intuition behind our method is that the training data contains both reliable and spurious correlations between inputs and labels, and that it is sometimes possible to partition the data into “training environments” in which the strength of the spurious ones is affected more than the reliable ones. We then train a model to rely on the correlations that are stable across environments. The corollary is that it ignores the environment-specific spurious ones.

As an example in VQA, let’s imagine the question *What animal is in the picture ?* A reliable correlation is the presence of canines in images that have dog as answer. An example of a spurious correlation is that most questions starting with *What sport...* have *tennis* as answer. A model that relies on this correlation irrespective of image contents or of the rest of the question will generalize poorly. This spurious correlation results from annotation and selection biases. Conceivably, data collected from two groups of annotators will exhibit different biases, *e.g.* one group mostly picking images with tennis as the answer, the other football. Our approach identifies such groupings of the data as training environments then trains a model that uses the correlations that are stable (*i.e.* similarly predictive) across environments.

Concretely, we partition the training set $\mathcal{T}=\{(\mathbf{x}_i, \mathbf{y}_i)\}_i$ of inputs \mathbf{x}_i and labels \mathbf{y}_i (one-hot vectors in a classification task) into E disjoint training environments \mathcal{T}_e such that $\bigcup_{e=1}^E \mathcal{T}_e = \mathcal{T}$. The environments are built to isolate the effect of spurious correlations, such that only the strength of reliable correlations remains stable across all environments. We provide additional justification for the principle in Section 3.3 and we describe strategies to build environments from existing datasets in Section 4. We show that they can be built by unsupervised clustering, by injecting prior knowledge, and by leveraging auxiliary annotations from existing datasets. Next, we describe how to train a model across environments to rely on stable correlations while ignoring spurious ones.

3.2. Training over multiple environments

Our goal is to learn a predictive model Φ that maps an input \mathbf{x} to an output $\hat{\mathbf{y}} = \Phi(\mathbf{x})$ such as a vector of class probabilities in a classification task. We represent the model as the combination of a feature extractor and a subsequent linear classifier. The feature extractor $f_{\theta}(\mathbf{x})$ (typically with a deep neural network) uses parameters θ to extract a vector $\mathbf{h} = f_{\theta}(\mathbf{x})$. The subsequent linear classifier is a matrix of weights \mathbf{W} and the whole model is described as $\Phi(\mathbf{x}) = \mathbf{W}f_{\theta}(\mathbf{x})$. The standard training procedure is to optimize θ and \mathbf{W} for maximum likelihood on the training set \mathcal{T} under a loss \mathcal{L} , *i.e.* solving the following optimization problem:

$$\arg \min_{\theta, \mathbf{W}} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{T}} \mathcal{L}(\mathbf{W}f_{\theta}(\mathbf{x}), \mathbf{y}). \quad (1)$$

In our method, assuming a prior definition of training environments \mathcal{T}_e (see Section 3.1), we want to train the model to be highly predictive on the training environments as well as on an OOD test set, in which only the input/output correlations *common to all training environments* can be assumed to hold. We train a different model $\Phi_e(\mathbf{x}) = \mathbf{W}_e f_{\theta}(\mathbf{x})$ for each environment. The feature extractor $f_{\theta}(\cdot)$ is shared, such that it identifies features common to all environments. But a different classifier \mathbf{W}_e is optimized for each environment. We want the features extracted by $f_{\theta}(\cdot)$ to be stable, *i.e.* similarly predictive across environments. For this, we choose to encourage the parameters of the classifiers \mathbf{W}_e to converge to a common value. This is naturally implemented by minimizing their variance over $e = 1..E$.

At test time, we use $\Phi^*(\mathbf{x}) = \bar{\mathbf{W}} f_{\theta}(\mathbf{x})$, where $\bar{\mathbf{W}}$ is the arithmetic mean of \mathbf{W}_e over e . Since the variance regularizer brings all \mathbf{W}_e toward a common value during training, the arithmetic mean is a natural choice. The complete optimization task is defined as:

$$\arg \min_{\theta, \mathcal{W}} \sum_e \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_e} \mathcal{L}(\mathbf{W}_e f_{\theta}(\mathbf{x}), \mathbf{y}) + \lambda \text{Var}_e(\mathbf{W}_e) \quad (2)$$

where λ is a scalar hyperparameter, $\mathcal{W} = \{\mathbf{W}_e\}_{e=1}^E$, and $\text{Var}_e(\mathbf{W}_e)$ is the variance of classifier weights. The standard definition of the variance gives

$$\text{Var}_e(\mathbf{W}_e) = (1/E) \sum_e \|\mathbf{W}_e - \bar{\mathbf{W}}\|^2 \quad (3)$$

$$\text{with } \bar{\mathbf{W}} = (1/E) \sum_e \mathbf{W}_e. \quad (4)$$

We refer to this definition as the ‘‘absolute variance’’ in our experiments. Finding a unique best value for λ in Eq. 2 proved difficult because the magnitude of the weights can vary widely during the early stages of the optimization. As a remedy, we use a relative measure of variance that rescales each term by the weights’ magnitude:

$$\text{Var}(\mathbf{W}_e) = (1/E) \sum_e (\|\mathbf{W}_e - \bar{\mathbf{W}}\|_2 / \|\mathbf{W}_e\|_1)^2 \quad (5)$$

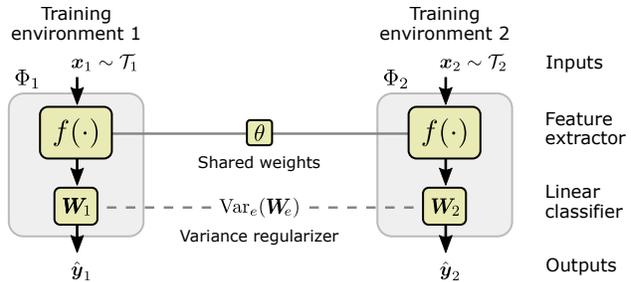


Figure 2. During training, we optimize a different copy of the model under each environment (each sees a different subset of the data). Two environments are pictured but our experiments use up to 18. The objective is to have the model rely on statistical patterns that are stable across environments. The weights of the feature extractor (θ) are shared across environments, those of the classifier (\mathbf{W}_e) are not. A regularizer encourages the latter to converge to a unique solution simultaneously optimal across environments. At test time, we use the arithmetic mean of these weights $\bar{\mathbf{W}}$.

It slightly improves our results (see Table 1) and makes λ easier to tune. We also found small improvements in optimizing Eq. 2 alternatively: one mini-batch serves to update θ , another one to update \mathbf{W} , repeating until convergence. It slightly improves the final accuracy but it is not crucial to the method and was only used in select experiments reported in Table 1.

3.3. Why it works

This section discusses further connections with existing methods and formal interpretations of generalization.

Causal view of generalization. Improving generalization requires distinguishing dataset-specific artifacts from the actual mechanisms of the task of interest. A causal view of generalization formalizes this as identifying causal properties⁴ of the data-generating process behind the observed data [8, 53]. A critical outcome of this formalization is that such causal properties are provably impossible to recover solely from the joint distribution over inputs and outputs that a standard training set represents. In other words, spurious and reliable correlations cannot be distinguished from one another.

Common avenues for bringing the missing information include architecture design (setting a prior on the causal structure), data augmentation (defining invariances in the input domain) and other task-specific inductive biases. Our method falls in another category that seeks to exploit implicit training signals found unexploited in existing datasets [8, 65, 30, 39]. More specifically, we use multiple

⁴ We do not claim to identify a causal model, we only aim at producing a predictive model. The causal formalization only serves to highlight that properties relevant to OOD generalization cannot be recovered from the joint distribution represented by a standard dataset, but can be informed by training across well-chosen environments.

training environments. These emerge naturally in data collected from multiple annotators, sites, or annotation interfaces. Data points are usually collated into a single dataset, but this loses information. Our premise is that such environments elicit different spurious correlations without affecting the reliable patterns inherent to the task of interest.

The fact that observational data from multiple environments can be informative about causal properties of the data generating process may seem at odds with basic principles of causal reasoning [50]. This is resolved when considering each environment as an intervention on the data generating process. Assuming these interventions only act on variables spuriously correlated with the output (hence the importance of a well-chosen partitioning), the causal mechanisms between non-intervened variables and the output will remain unchanged by the principle of independent causal mechanisms [54]. The statistical patterns inherent to the task thus remain stable across environments.

Source of improvements. The improvements obtained through our method therefore hinge on the information used to obtain training environments (*e.g.* multiple existing datasets) or to create them by “unshuffling” an existing dataset. The latter requires the practitioner to bring in meta knowledge about the task or dataset, such as axes along which to cluster the data. Conversely, environments made as random partitions are not expected to bring any benefit, as verified in our experiments (see Table 1).

Invariant risk minimization. Our training procedure is inspired by the principle of invariant risk minimization (IRM) [8]. IRM proposes to identify a representation of data such that the optimal classifier, on top of this representation, is identical across environments. Formally, using our notations, this amounts to optimizing the feature extractor $f_\theta(\cdot)$ and linear classifier \mathbf{W} for the following objective:

$$\min_{\theta, \mathbf{W}} \sum_e \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_e} \mathcal{L}(\mathbf{W}^* f_\theta(\mathbf{x}), \mathbf{y}) \quad (6)$$

$$\text{s.t. } \mathbf{W}^* \in \arg \min_{\mathbf{W}} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_e} \mathcal{L}(\mathbf{W} f_\theta(\mathbf{x}), \mathbf{y}), \forall e. \quad (7)$$

The constraint on \mathbf{W}^* is the crux of the principle. A classifier that is optimal in a given environment can only use the features that are reliable predictors in that environment. Requiring the classifier \mathbf{W}^* to be simultaneously optimal across all environments (*i.e.* at the intersection of all environment-specific optima) means that it can only use stable features. In other words, consider a spurious correlation, specific to an environment e , between the output labels and a feature $\tilde{\mathbf{x}}$. A model (feature extractor and classifier) trained in isolation on e would use this feature $\tilde{\mathbf{x}}$. However, this spurious correlation does not hold in another environment e' . Even though the shared feature extractor could extract some semblance of the feature $\tilde{\mathbf{x}}$ in e' , this feature

will not be predictive in the same way as in e . Therefore, the optimal classifier in e' will not use $\tilde{\mathbf{x}}$ in the same way. Since we are looking for a unique classifier that is simultaneously optimal in e and e' , the shared feature extractor must ignore this unreliable feature, and only extract those that are *similarly predictive* across environments.

Proposed method and IRM. The objective of Eq. 7 involves an impractical nested optimization. The approximation proposed in [8] replaces the constraint with a regularizing term in the objective that uses the gradient of the environment-specific risk with respect to the classifier. The resulting objective is highly non-convex and has been reported to be difficult to train in practice. Our method (Eq. 2) uses the variance of \mathbf{W}_e as a regularizer. The gradient of the risk in [8] is motivated as a measure of “how optimal” a classifier is. Our version operates directly in the parameter space of the classifier. Following Section 3.2, stationary points that satisfy the IRM principle are also solutions to our optimization problem, although the converse is not true, so further work is warranted to determine conditions in which guarantees discussed in [8] possibly hold with our formulation. Our results are only empirical but showed it to be very stable during training and highly effective in our use cases.

Finally, recent breakthrough theoretical [40] and empirical results [60] proved identifiability in non-linear ICA when the generating process is a conditional distribution of which the conditioning variable is observed. This setting is analogous to training across environments with the environment ID interpreted as the conditioning variable. Further work is needed to establish formal connections with these results.

4. Experiments

We present three applications on the task of VQA, which is notorious for dataset biases. Our strongest results are with the VQA-CP dataset [2] which is designed to test out-of-distribution generalization. The other two applications use GQA [34], and VQA v2 [25] combined with Visual Genome QA [42]. The quantitative improvements in these other two applications are smaller but they demonstrate the wider applicability of the method. Most other methods in Table 2 are specific to VQA-CP.

Implementation. We implemented the method on top of the “bottom-up and top-down attention” model [66] (details in supp. mat.) since it serves as the baseline of most competing techniques on VQA-CP [13, 17, 27, 29, 47, 55]. Our method should readily apply to recent models [15, 23, 24, 45, 46, 61, 75] including those with stronger baseline performance on GQA [32, 33]. Evaluations on VQA-CP follow the guidelines of Teney *et al.* [67], including performing ablations on “Other” questions only, and reporting both in-domain and OOD performance without retraining. Our results are also averaged across multiple runs (thus not to

outcome of one possibly-lucky random seed) unlike most results reported in the literature.

4.1. Robustness to language biases (VQA-CP)

Experimental setup. The VQA-CP dataset [2] was constructed by reorganizing VQA v2 [25] such that the correlation between the question type and correct answer differs in the training and test splits. For example, the most common answer to questions starting with *What sport...* is *tennis* in the training set, but *skiing* in the test set. A model that guesses an answer primarily from the question will perform poorly. In our experiments, we report the accuracy on the official test set, but also on a validation set that we built by holding out 8,000 random instances from the training set. This serves as to measure “in-distribution” performance, while the test set serves to measure generalization to out-of-distribution data. As discussed in [67], evaluation on the ‘yes/no’ and ‘number’ categories of VQA-CP have unintuitive issues (for example, randomly guessing yes/no on the former category achieves 72.9% while a method like [2] only gets 65.5%; thus, a random, untrained model is usually better than a trained one). For these reasons, our ablation study uses only the ‘other’ type of questions.

Environments from ground truth question types. We first present experiments for which we built training environments with the ground truth type of questions (provided with the dataset). Each training question has one label among 65. This label serves as a natural clustering of the data. We assign the 65 clusters randomly to E environments, splitting clusters as needed to obtain the same number of training questions per environment. We trained our method with a different number of environments (see Fig. 3b). The point $E=1$ corresponds a standard training of the model with the whole dataset. The plot shows a clear improvement with multiple environments, with a peak performance with $E=15$. Why does the accuracy decrease with more environments? We believe that the diversity and amount of data in each environment then gets too low. We experimented with other strategies (not reported in plots and tables) to assign clusters to environments other than randomly, by maximizing or minimizing the variation in the answer distribution in each environment (compared to the whole dataset). We found that the random assignment performed best. It keeps the distribution of answers relatively similar across environments, unless E is too large, which further explains the slight decrease in accuracy then.

Environments by clustering questions. We now present experiments where the environments are built through unsupervised clustering of the questions. We do not use the ground truth question types here. We rely on our prior knowledge that a model should not be overly reliant on the general form of a question. We represent the questions as

binary bag-of-words vectors (details in supp. mat.) and cluster them with K -means. As above, we then assign the clusters randomly to E environments ($E < K$). We plot in Fig. 3 (right) the accuracy of the model against the number of clusters K . There is a distinct broad optimum. The best accuracy is close but still inferior to the strategy that uses the ground truth question types (compare the peaks in Fig. 3 middle and right). We measured the similarity of the unsupervised clustering with the ground truth type in terms of Rand index, and noted that it was positively correlated with the accuracy. This shows that using ground truth types is the better strategy and the clustering approximates it.

Ablative analysis. We provide an ablation study in Table 1. The performance substantially increases on the test set with the proposed method compared to all baselines. The variance regularizer is crucial to the success of the method. We plot in Fig. 3 (left) the accuracy as a function of the regularizer weight (λ in Eq. 2). There is a clear optimum, with higher values being generally better (the plot uses a log scale). In Table 1, we also observe that the relative variance performs slightly better than the absolute variance. We also note that the alternating optimization scheme performs slightly better. It works best after a few epochs of non-alternating “warm-up”, during which all parameters are updated together. The use of the alternating optimization is not crucial to the overall success of the method, and it is not used in any other experiment.

Comparison to existing methods. We trained our method on the whole VQA-CP dataset, including ‘yes/no’ and ‘number’ questions to compare it against existing methods (see Table 2). Our method surpasses all others on ‘other’, most of them by a large margin. The method of Clark *et al.* [17] gets better results on the ‘yes/no’ and ‘number’ questions, but its results on the standard splits of VQA v2 are also down to baseline levels (*i.e.* similar to a random guess out of the subset of answers used in each category). In comparison, our performance on the standard splits remains higher. Note that some competing methods admittedly use the test set as a validation set (!) for hyperparameter selection and/or model selection [2, 27]. We rather hold out 8k instances from the training set to serve as a validation set as recommended in [67]. They serve for example to monitor training and determine the epoch for early stopping.

4.2. Invariance to equivalent questions (GQA)

Experimental setup. The GQA dataset [34] is a VQA dataset built with images of the Visual Genome project [42] and questions generated from the scene graphs of these images. The questions are generated from a large number of templates and hand-coded rules, such that they are of high linguistic quality and variety. We present experiments using annotations of “equivalent questions” provided with

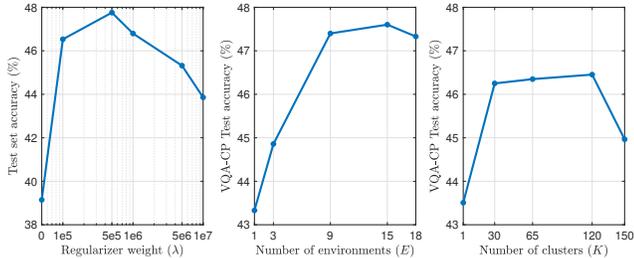


Figure 3. Sensitivity to hyperparameters on VQA-CP, using environments built from question groups (left and middle) or by clustering questions (right). See discussion in Section 4.1.

	VQA-CP v2	
	Val. set	Test set
Baseline	54.74	43.33
Environments: random; rel. var., no alt. opt.	53.34	43.51
Environments: clustered questions; rel. var., no alt. opt.	54.10	46.35
Environments: question groups; rel. var., no alt. opt.	53.87	47.60
+ Alternating optimization (0 warm-up epoch)	54.00	47.71
+ Alternating optimization (2 warm-up epochs)	53.90	47.82
+ Alternating optimization (4 warm-up epochs)	53.98	48.06
+ Alternating optimization (6 warm-up epochs)	53.86	47.38
Without variance regularizer	40.76	39.14
With absolute variance regularizer	51.44	46.17

Table 1. Ablative study on VQA-CP (accuracy in percent, training on ‘Other’ questions only). Our method brings a significant gain over the baseline, both with environments built using the ground truth question types, and with environments built by unsupervised clustering of the questions. As a sanity check, we run the method with random environments, which gives results essentially identical to the baseline, as expected. The alternating optimization scheme brings a additional small improvement, although it is not crucial to the success of the method.

the dataset. These annotations are not used in any existing model, to our knowledge. A small fraction of training questions ($\sim 17.4\%$ in the balanced training set) are annotated with up to three alternative forms. They involve a different word order or represent a different way of asking about a same thing. For example:

- *Is there a fence in the scene ?*
Do you see a fence ?
- *Which size is the green salad, small or large ?*
Does the green salad look large or small ?
- *Are there airplanes or cars ?*
Are there any cars or airplanes in this photo ?

Some alternative forms are already part of the dataset as other training questions, others are not. The straightforward way to use these annotations is by data augmentation, *i.e.* aggregating the equivalent forms with original training set.

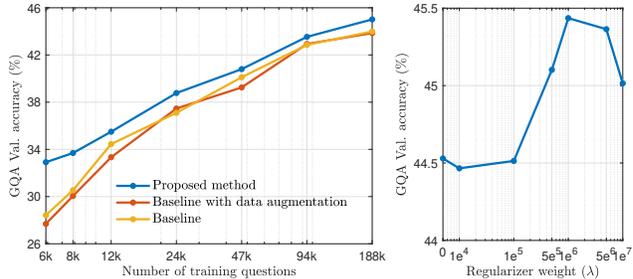


Figure 4. Experiments on GQA using equivalent questions to build environments. Our method provides consistent gains over the baseline, especially in the low-data regime. The improvement diminishes as more data is available, and is essentially imperceptible when the model is training on the full 2M training examples ($\sim 10\times$ than shown on this plot). A naive use of the equivalent questions for data augmentation has a negative effect because it shifts the distribution of the training set away from the test set.

Training environments with equivalent questions. We use our method to help learning invariance to the linguistic patterns of equivalent questions. We use $E=4$ environments where we replace, in each, a question by its e^{th} equivalent form if available, or the original form otherwise. Each environment thus contains a single form of each question.

Results. We compare in Fig. 4 (left) the accuracy of our method with same model trained on the standard training set, and with the data augmentation baseline described above. The data augmentation does not help despite the additional training examples, because it modifies the distribution of training examples away from the distribution of test questions. Our method, in comparison, brings a clear improvement. For a fair comparison, we made sure that the data augmentation uses the exact same questions (original and equivalent forms) in every mini-batch, such that the improvement is strictly brought on by the architectural differences of our method. The improvement with our method is greatest with low amounts of training data (we use random subsets of the full training set). The full dataset provides a massive 14M examples (about 1M in its balanced version), at which point the impact of our method is imperceptible. The training set then essentially covers the variety of linguistic forms and concepts exhaustively enough such that there is no benefit from the additional annotations.

It is worth noting that all improvements brought by our method come from only a small fraction of questions being annotated with equivalent forms. It would therefore be realistic to annotate a real VQA dataset with similar equivalent forms, and investigate possible gains with our method, which we hope to do in the future.

In Fig. 4 (right) we plot the accuracy as a function of the regularizer weight. The clear optimum confirms again that the regularizer is a crucial component of the method.

	VQA-CP v2, Test set				VQA v2, Validation set			
	Overall	Yes/no	Numbers	Other↓	Overall	Yes/no	Numbers	Other
SAN [74]	24.96	38.35	11.14	21.74	52.02	–	–	–
GVQA [2]	31.30	57.99	13.68	22.14	48.24	–	–	–
Ramakrishnan <i>et al.</i> , 2018 [55]	42.04	65.49	15.87	36.60	62.75	79.84	42.35	55.16
Grand and Belinkov, 2019 [27]	42.33	59.74	14.78	40.76	51.92	–	–	–
RUBi [13]	47.11 ± 0.51	68.65	20.28	43.18	61.16	–	–	–
Teney <i>et al.</i> , 2019 [68]	46.00	58.24	29.49	44.33	–	–	–	–
Product of experts [17]	40.04	43.39	12.32	45.89	63.21	81.02	42.30	55.20
Clark <i>et al.</i> , 2019 [17]	52.01	72.58	31.12	46.97	56.35	65.06	37.63	54.69
Our baseline model	37.87 ± 0.24	41.62	10.87	44.02	61.09 ± 0.26	80.23	42.25	53.97
Proposed method	42.39 ± 1.32	47.72	14.43	47.24	61.08 ± 0.12	78.32	42.16	52.81
Our baseline model ($\times 4$ ensemble)	39.30	40.72	11.18	46.44	64.26	82.07	44.56	56.33
Proposed method ($\times 4$ ensemble)	43.37	47.82	14.35	49.18	63.47	81.99	43.07	55.21

Table 2. Comparison with existing methods designed to improve generalization on VQA-CP (accuracy in percents). The evaluation on ‘yes/no’ and ‘number’ questions is highly unreliable (see Section 4.1 and [67]). On the ‘Other’ questions however, our method surpasses all others. Our improvements on VQA-CP come only with a slight decrease in performance when trained and evaluated on the standard splits of VQA v2 (right columns). Reassuringly, the benefits of our method are cumulative with those of an ensemble (obtained by averaging the predictions of four models trained independently). The proposed method evaluated here uses environments built with question groups, $E=15$ environments, the relative variance regularizer, and no alternating optimization.

4.3. Multi-dataset training (VQA v2 and VG QA)

Experimental setup. These experiments apply our method to the training of a model on multiple datasets simultaneously. The VQA v2 dataset has previously been aggregated with Visual Genome QA (VG) [42] as a simple way to use more training data. The datasets contain similar types of questions, but it is reasonable to assume that they have slightly different distributions. We use $E=2$ environments, the first one containing the VQA v2 training data, the second the VG data.

Results. In Table 3, we compare our method with a model trained on VQA v2, another trained on VG, and one trained on the aggregation of the two datasets. The improvement is small but was verified over multiple training runs. We also ruled out explanation of the improvement as merely an ensembling effect, by comparing an ensemble of the baseline with one of the proposed method. The benefits of our method are cumulative with those of an ensemble, which suggests that our method should also apply to higher-capacity models. A number of such models have been described with a higher performance on VQA v2 [15, 23, 24, 45, 46, 61, 75] and it will be interesting to combine them with our method in the future.

5. Conclusions

We presented a method to train a deep models to better capture the mechanism of a task of interest, rather than blindly absorbing all statistical patterns from a training set. The method is based on the identification of correlations that are stable across multiple training environments, *i.e.*

	VQA v2, Validation set					VG
	Overall	Overall	Yes/no	Numbers	Other	Val.
	Ens. $\times 4$	Single model				
Baseline model						
Trained on VQA v2	64.86	63.07 ± 0.23	81.40	42.09	54.21	49.67
Trained on VG	28.48	27.58 ± 0.22	0.11	36.03	47.11	60.17
Trained on Aggregated data	65.47	63.32 ± 0.35	82.27	40.99	55.98	61.20
Proposed method						
Without variance reg.	64.33	62.18 ± 0.27	78.95	41.68	54.42	59.68
With variance reg.	65.73	63.80 ± 0.17	81.00	42.35	55.97	60.54

Table 3. Multi-dataset training with VQA v2 and Visual Genome. The standard practice is to aggregate the two datasets. Our method treats them as two distinct training environments. The improvement is very small, but it comes at zero extra cost, and it was verified over multiple runs (mean and standard deviation are reported), as well as in an ensemble (first column). It was also verified on two different implementations of the baseline model (not in table).

subsets of the training data. We described several strategies to build these environments using different forms of prior knowledge and auxiliary annotations. We showed benefits in various conditions including out-of-distribution test data, low-data training, and multi-dataset training.

An exciting challenge in computer vision is to design models solving tasks rather than datasets. Our strong results on VQA, which is known for its challenges in generalization and data scarcity, give us confidence that suitable tools like this method are emerging to make progress in this direction.

The proposed method is related to a series of works on IRM that appeared concurrently with the preparation of this paper [8, 3, 14, 16, 35]. Much remains to be done in terms of theoretical and empirical comparisons of these methods and their application to real (non-toy) data beyond VQA.

References

- [1] Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M Pohl. Bias-resilient neural network. *arXiv preprint arXiv:1910.03676*, 2019. **1, 3**
- [2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4971–4980, 2018. **1, 2, 3, 5, 6, 8, 15**
- [3] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. *arXiv preprint arXiv:2002.04692*, 2020. **3, 8**
- [4] Arjun R Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. Words aren’t enough, their order matters: On the robustness of grounding visual referring expressions. *arXiv preprint arXiv:2005.01655*, 2020. **1, 2**
- [5] 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 vqa. *CVPR*, 2018. **14**
- [6] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018. **1**
- [7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015. **1**
- [8] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. **2, 3, 4, 5, 8, 13**
- [9] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 9448–9458, 2019. **1, 2**
- [10] Elias Bareinboim, JD Correa, Duligur Ibeling, and Thomas Icard. On pearl’s hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl (provisional title)*, 2020. **13**
- [11] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *arXiv preprint arXiv:1705.08841*, 2017. **3**
- [12] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug 1996. **3**
- [13] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. *arXiv preprint arXiv:1906.10169*, 2019. **2, 3, 5, 8**
- [14] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi S Jaakkola. Invariant rationalization. *arXiv preprint arXiv:2003.09772*, 2020. **3, 8**
- [15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. **5, 8**
- [16] Yo Joong Choe, Jiyeon Ham, and Kyubyong Park. An empirical study of invariant risk minimization. *arXiv preprint arXiv:2004.05007*, 2020. **3, 8**
- [17] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019. **2, 3, 5, 6, 8, 15**
- [18] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *ICML Workshop on Uncertainty and Robustness*, 2020. **3**
- [19] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *CVPR*, 2017. **1**
- [20] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004. **1, 3**
- [21] Shi Feng, Eric Wallace, and Jordan Boyd-Graber. Misleading failures of partial-input baselines. *arXiv preprint arXiv:1905.05778*, 2019. **1**
- [22] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. **2**
- [23] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6639–6648, 2019. **5, 8**
- [24] Peng Gao, Haoxuan You, Zhanpeng Zhang, Xiaogang Wang, and Hongsheng Li. Multi-modality latent interaction network for visual question answering. In *Proc.*

- IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5825–5835, 2019. 5, 8
- [25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. *arXiv preprint arXiv:1612.00837*, 2016. 2, 5, 6
- [26] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 1, 2
- [27] Gabriel Grand and Yonatan Belinkov. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. *arXiv preprint arXiv:1906.08430*, 2019. 1, 2, 3, 5, 6, 8
- [28] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 2
- [29] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yibing Liu, Yinglong Wang, and Mohan Kankanhalli. Quantifying and alleviating the language prior problem in visual question answering. *arXiv preprint arXiv:1905.04877*, 2019. 1, 2, 3, 5
- [30] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017. 3, 4
- [31] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018. 3
- [32] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. *arXiv preprint arXiv:1905.04405*, 2019. 5
- [33] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018. 5
- [34] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. 2, 3, 5, 6, 15
- [35] Daksh Idnani and Jonathan C Kao. Learning robust representations with score invariant learning. 3, 8
- [36] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. 2016. 2
- [37] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017. 1, 2
- [38] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*, 2016. 3
- [39] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019. 4
- [40] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217, 2020. 5
- [41] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *Proc. Eur. Conf. Comp. Vis.*, 2012. 1, 2
- [42] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yanis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 2, 5, 6, 8
- [43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Advances in Neural Inf. Process. Syst.*, 2012. 2
- [44] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017. 1, 2
- [45] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*, 2019. 5, 8
- [46] Bei Liu, Zhicheng Huang, Zhaoyang Zeng, Zheyu Chen, and Jianlong Fu. Learning rich image region representation for visual question answering. *arXiv preprint arXiv:1910.13077*, 2019. 5, 8
- [47] Rabeeh Karimi Mahabadi and James Henderson. Simple but effective techniques to reduce biases. *arXiv preprint arXiv:1909.06321*, 2019. 2, 3, 5
- [48] Tom M Mitchell. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research . . . , 1980. 1

- [49] Sepideh Pashami, Anders Holst, Juhee Bae, and Sławomir Nowaczyk. Causal discovery using clusters from observational data. In *FAIM'18 Workshop on CausalML, Stockholm, Sweden, July 15, 2018*, 2018. [3](#)
- [50] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000. [5](#)
- [51] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv:1801.04016*, 2018. [13](#)
- [52] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014. [14](#)
- [53] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *J. of the Royal Statistical Society*, 2016. [4](#), [13](#)
- [54] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. The MIT Press, 2017. [5](#), [13](#)
- [55] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pages 1541–1551, 2018. [2](#), [3](#), [5](#), [8](#)
- [56] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021. [13](#)
- [57] Eleni Sgouritsa, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Identifying finite mixtures of nonparametric product distributions and causal inference of confounders. *arXiv preprint arXiv:1309.6860*, 2013. [3](#)
- [58] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Proc. Advances in Neural Inf. Process. Syst.*, 2020. [2](#)
- [59] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. [14](#)
- [60] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*, 2020. [5](#)
- [61] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. [5](#), [8](#)
- [62] Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. Semantic equivalent adversarial data augmentation for visual question answering. In *Proc. Eur. Conf. Comp. Vis.*, 2020. [14](#)
- [63] Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987. [2](#)
- [64] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. On incorporating semantic prior knowledge in deep learning through embedding-space constraints. *arXiv preprint arXiv:1909.13471*, 2019. [3](#)
- [65] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*, 2020. [3](#), [4](#)
- [66] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *CVPR*, 2018. [2](#), [5](#), [14](#)
- [67] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart’s law. In *Proc. Advances in Neural Inf. Process. Syst.*, 2020. [3](#), [5](#), [6](#), [8](#), [13](#)
- [68] Damien Teney and Anton van den Hengel. Actively seeking and learning from live data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [8](#)
- [69] Antonio Torralba, Alexei A Efros, et al. Unbiased look at dataset bias. In *CVPR*, volume 1, page 7, 2011. [1](#)
- [70] Vladimir Vapnik. *Statistical learning theory*. john wiley&sons. Inc., New York, 1998. [1](#)
- [71] M Alex O Vasilescu and Demetri Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Proc. Eur. Conf. Comp. Vis.* Springer, 2002. [3](#)
- [72] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Proc. Advances in Neural Inf. Process. Syst.*, 2019. [2](#)
- [73] Tianlu Wang, Jieyu Zhao, Kai-Wei Chang, Mark Yatskar, and Vicente Ordonez. Adversarial removal of gender from deep image representations. *arXiv preprint arXiv:1811.08489*, 2018. [3](#)
- [74] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked Attention Networks for Image Question Answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. [8](#)
- [75] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual