Barlow constrained optimization for Visual Question Answering

Abhishek Jha[1*] Badri Patro[1*] Luc Van Gool[1,2] Tinne Tuytelaars[1]
[1]ESAT-PSI, KU Leuven, [2]CVL, ETH Zürich
firstname.lastname@esat.kuleuven.be

Abstract

Visual question answering is a vision-and-language multimodal task, that aims at predicting answers given samples from the question and image modalities. Most recent methods focus on learning a good joint embedding space of images and questions, either by improving the interaction between these two modalities, or by making it a more discriminant space. However, how informative this joint space is, has not been well explored. In this paper, we propose a novel regularization for VQA models, Constrained Optimization using Barlow's theory (COB), that improves the information content of the joint space by minimizing the redundancy. It reduces the correlation between the learned feature components and thereby disentangles semantic concepts. Our model also aligns the joint space with the answer embedding space, where we consider the answer and image+question as two different 'views' of what in essence is the same semantic information. We propose a constrained optimization policy to balance the categorical and redundancy minimization forces. When built on the state-of-the-art GGE model, the resulting model improves VQA accuracy by 1.4% and 4% on the VQA-CP v2 and VQA v2 datasets respectively. The model also exhibits better interpretability. Code is made available: `https://github.com/abskjha/Barlow-constrained-VQA`

1. Introduction

Visual question answering (VQA) [4] is a challenging vision-and-language task. It involves reasoning about a visual scene based on a free-form natural language question. Answering the question requires learning semantic associations between concepts across the two modalities. As different questions and images referring to the same kind of query and scene should yield a similar answer, learning semantics in the individual modalities and their cross-modal interactions is essential for solving VQA. Many recent works approach this by learning a joint embedding space [37, 30, 14] or by modeling an attention mechanism [43, 13, 25, 44] in one modality conditioned upon the other. Another line of work tries to improve the discriminant power [24] of the joint embedding space to improve the answering performance. These have been important contributions.

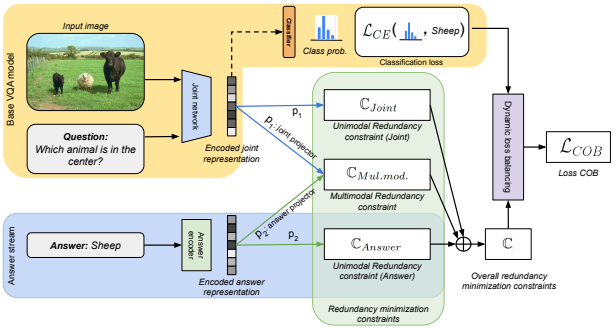

Figure 1: **COB**: we propose a set of redundancy minimization constraints $\mathbb{C}$ (green region) which are applied along with a cross-entropy loss $\mathcal{L}_{CE}$ (yellow region). The final COB loss $\mathcal{L}_{COB}$ is the weighted sum of constraint loss with cross-entropy loss, where the weight is dynamically updated using a loss balancing module.

However, high discriminant power of a feature space does not imply high information content [23]. While a highly discriminant space may yield better performance on a loss-specific task by modelling the most discriminant features for a given data distribution, it may be more susceptible to changes in that data distribution. A discriminant space subjected to an additional information preserving constraint, on the other hand, may yield a richer feature space that can generalize better to previously unseen data.

In this paper, we propose a redundancy reduction constraint, inspired by Barlow's third hypothesis [5] on sensory message transformations, to incorporate more information in the joint feature space. This third hypothesis (*Redundancy-reducing hypothesis*), states *"that sensory relays recode sensory messages so that their redundancy is reduced but comparatively little information is lost."*

Redundancy in a feature space arises when multiple feature components cover the same semantic concept. Taking into account the fixed dimensionality of the feature space, this causes the overall information content of the feature space to be suboptimal. A less redundant feature space can model the same information with fewer feature dimensions,

*Equal contribution.

or more information with the same number of feature dimensions. This results in a more informative embedding space that model multimodal concepts better and thereby provide a superior VQA performance.

To address this challenge for VQA, we propose an additional decorrelation loss term besides the categorical loss for predicting the answers. This additional loss term encourages decorrelation across feature components and thereby improves the information content of the embedding space. Recently for the self-supervised representation learning task, Zbontar *et al*. [46], with their Barlow twins model, have shown that a decorrelation loss, modelled according to Barlow's Redundancy reduction hypothesis, when applied to two views of the same data encoded by the twins model, can act as a good supervisory signal to learn visual features. Here, we use a similar decorrelation formulation as Barlow twins [46], but reformulated for two multimodal views of the data. We pose that the information to be extracted from the image+question input ideally corresponds to the information present in the answer. In other words, image+question and answer can be considered as two different 'views' of the same content. When computing the correlation, we therefore not only consider the auto-correlation in the joint image+question space, but also the cross-correlation between answer and joint space, as well as the auto-correlation in answer space. As an additional advantage, this brings in information about the semantic similarity between answers via the word embedding used for the answer space. Our full pipeline, combining categorical loss with redundancy reduction is shown in Figure 1.

We also found that directly applying the decorrelation minimization loss (Barlow loss) to a randomly initialized embedding space yields a very high loss. As a result, naively adding a Barlow loss, next to the cross-entropy loss, results in inferior VQA results. On the other hand, when applying the Barlow loss to the already aligned (pre-trained by cross-entropy) embedding space, this issue does not occur (see Section 4.2). Based on this empirical evidence, we formalize a parametric constrained optimization policy to balance the two forces. This results in a more informative and discriminant embedding space, leading to an improvement in the answering accuracy. In summary, our contributions are as follows:

(i) We propose the COB regularization which focuses on redundancy reduction between the joint embedding space of questions and images and the answer embedding space, to improve the information content of a VQA model.

(ii) We propose a policy to balance the categorical and redundancy reduction forces to train the model.

(iii) We improve the state-of-the-art performance on the challenging VQA v2 [15] and VQA-CP v2 [2] datasets.

(iv) Our proposed method improves the interpretability of the VQA model it builds on.

2. Related work

Visual question answering: VQA has taken up momentum after the introduction of a standard dataset VQA [4] and early multimodal techniques to solve this problem [31, 4, 20]. Initial approaches [37, 30, 14] jointly analyze visual and question feature embeddings by concatenating or correlating both features. In later works [43, 13, 25, 44], it was observed that attending to specific parts in the images and questions helps to better reason and answer. The subsequent discovery of language bias in the standard VQA dataset led towards less biased datasets and more robust models. Agrawal *et al*. [2] proposed VQA-CP v1 and VQA-CP v2 to overcome the language and distributional bias of the VQA v1 [4] and v2 [15] datasets. A critical reasoning-based method proposed by Wu *et al*. [42] ensured the correct answers match the most influential visual regions to overcome the dataset bias. Various authors such as Ramakrishnan *et al*. [36] proposed an adversarial-based method, and Jing *et al*. [22] decomposed a linguistic representation technique to overcome language prior in VQA. Clark *et al*. [9] proposed an ensemble based method to avoid known dataset bias, while Han *et al*. [16] proposed a gradient ensemble method to overcome both shortcut bias and distributional bias in the dataset. Hence, most methods focus on regularisation techniques to overcome language bias. In this paper, we focus on a regularisation technique to reduce redundancy in the VQA model, and show this further improves its performance.

Redundancy reduction: Dimensionality reduction is one way of reducing redundancy of a feature space, i.e. by minimizing the number of feature components required to model the data. Linear dimensionality reduction techniques like Principal Component Analysis (PCA) [35] for a single modality provide a mapping between the original feature space and the space spanned by principal components. In this new space low energy principal components can be dropped with a minimal loss of information. Similarly for multiple modalities, Canonical Correlation Analysis (CCA)-like techniques [19, 17] provide a linear mapping between individual modalities and a smaller joint embedding space. After CCA, the projections of the modalities are highly correlated, but they are decorrelated across the resulting feature dimensions. Our proposed method promotes the learning of decorrelated feature components similar to PCA and CCA. However, unlike PCA and CCA, the learned projection between the original features and the decorrelated component space is non-linear.

Recently, Kalantidis *et al*. [23] proposed a twin-loss similar to the Barlow twin loss of Zbontar *et al*. [46] to learn a non-linear dimensionality reduction, as an alternative to PCA. They train a twin encoder-decoder architecture with a decorrelation optimization between the output projections of the nearest neighbors in the input space. Our method is

similar to [23] in the way our constraint is motivated, however it is not the primary objective function in our model. We optimize cross-entropy to maximize the answering accuracy, with the decorrelation as an optimization constraint.

Decorrelation loss: Decorrelation losses are often used in recent representation learning methods [46, 6, 23] by using a shared twin encoder-decoder architecture on two views of the same samples coming from a unimodal space, while minimizing the distance between an identity matrix and the correlation matrix of the output representations. This forces the feature components in the output embedding space to be orthogonal. In our case, the inputs come from two different modalities, and hence it differs from the twins formulation. The hypothesis behind the use of our proposed constraint on two different modalities is motivated by the assumption that image-question pairs and their answers should be related to the same underlying concept.

Stabilizing losses: Optimizing networks for different objectives requires balancing or weighting the loss gradients, especially for the objectives that are non-complementary [45, 1, 27, 18], as they force the feature space to sway in two different directions [27, 18]. Improper balancing of such objectives can lead to trivial solutions [18, 38], and hence the loss weighting factor is an important hyperparameter. Rezende and Viola [38] propose a generalized ELBO loss with constrained optimization (GECO), a learnable weighting scheme for balancing KL divergence and the reconstruction loss in the context of training variational auto-encoders [27]. We propose a similar constrained optimization formulation for the cross-entropy loss in our approach, that assigns a dynamic weight to the constraint. Unlike GECO, our objective function and constraint do not have similar scales, with the initial constraint loss being orders of magnitude larger than the main objective function.

3. Method

3.1. Preliminaries

VQA formulation: The VQA task with cross-entropy loss can be defined as modelling the categorical distribution over a fixed answer vocabulary given a set of image and question pairs. For a data distribution $\mathcal{D}$ for this problem with an instance $d_k = \{v_k, q_k, a_k\} \in \mathcal{D}^{VQA}$, the task is to predict an answer $a_k \in \mathcal{D}^A$, given an image $v_k \in \mathcal{D}^V$ based on a question $q_k \in \mathcal{D}^Q$. Contemporary methods [4, 37, 30, 14] solve this task by first encoding each of the two modalities using pre-trained encoders e_v, e_q, and then learning a joint representation over them. Each instance pair (v_k, q_k) can then be represented by a point $m_k^f \in \mathcal{D}_M$ in this joint representation space:

$$m_k^f = f_{\theta_J}\big(e_v(v_k), e_q(q_k)\big) \tag{1}$$

$$m_k^l = l_{\theta_L}(m_k^f) \tag{2}$$

where f_{θ_J} is the joint network with parameters θ_J, and l_{θ_L} with parameters θ_L is the logistic projection, which along with the softmax non-linearity, is used to predict the probability distribution over the answer space $\mathcal{D}^A$. A cross-entropy loss ($\mathcal{L}_{CE}$) between the resulting probability scores and the ground truth answer a_k is then computed. For a batch: (V, Q, A, M^f, M^l) consisting of n_b number of samples $(v_k, q_k, a_k, m_k^f, m_k^l)$, the cross-entropy loss can be defined as:

$$\mathcal{L}_{CE}(M^l, A) = -\frac{1}{n_b}\sum_k log\bigg(\frac{\exp(m_k^l[a_k])}{\sum_{a' \in \mathcal{D}^A}\exp(m_k^l[a'])}\bigg) \tag{3}$$

where $m_k^l[a_k]$ is the logit corresponding to the answer a_k. The resulting gradient is then used to train the parameters of the VQA network.

Barlow twins formulation: In order to reduce redundancy among the feature components, Zbontar *et al*. [46] propose a distance minimization loss between an identity matrix ($I \rightarrow \mathbb{R}^{N_B \times N_B}$) and the correlation matrix ($C \in \mathcal{D}^B \times \mathcal{D}^B$) computed between the non-linear projections $b_{\theta_B}(.)$ of the encoded representation $e_s(.)$ of the two augmented views $(s_{k|1}, s_{k|2})$ of the same input $s_k \in \mathcal{D}^S$. For a batch $S = \{s_k\}_{k=1}^{n_b}$ of n_b such samples, and its two augmented views S_1 and S_2, the Barlow projections are:

$$S_1^b = b_{\theta_B}(e_s(S_1)); S_2^b = b_{\theta_B}(e_s(S_2)) \tag{4}$$

$$C(S_1^b, S_2^b) = Norm(S_1^b) \otimes Norm(S_2^b) \tag{5}$$

where e_s is the modality specific feature encoder, b_{θ_B} is the non-linear projector from the encoded feature space to a N_B dimensional Barlow optimization space $\mathcal{D}^B$, while $Norm(.)$ is a normalization function along the batch [21]. Each element of the correlation matrix $C^{\mathcal{S}} = C(S_1^b, S_2^b)$ can be indexed by (i, j), as C_{ij}^S:

$$C_{ij}^{\mathcal{S}} = \frac{\sum_k s_{k|1}^b[i] s_{k|2}^b[j]}{\sqrt{\sum_k (s_{k|1}^b[i])^2}\sqrt{\sum_k (s_{k|2}^b[j])^2}} \tag{6}$$

$$\mathcal{L}_B^{\mathcal{S}} = \sum_i (1 - C_{ii}^{\mathcal{S}})^2 + \gamma \sum_i \sum_j (C_{ij}^{\mathcal{S}})^2 \tag{7}$$

where $1 \leq i, j \leq N_B$ indexes the feature components of the k^{th} sample ($s_{k|1}^b, s_{k|2}^b \in \mathcal{D}^B$) in the projected batch (S_1^b, S_2^b). The first term in equation 7, minimizes the distance between the two projected representations while the second term promotes decorrelation across the feature components, with γ a positive hyperparameter to weight the two loss terms. Our goal is to learn a discriminant space $\mathcal{D}^M$, that minimizes $\mathcal{L}_{CE}$ while reducing the redundancy, by reformulating the unimodal barlow decorrelation loss $\mathcal{L}_B^{\mathcal{S}}$ for a multimodal input space $(\mathcal{D}^M, \mathcal{D}^A)$.

3.2. Objective function formulation

A typical classification based VQA task can be modelled with equations 1 to 3. Different methodological improvements have emerged either in the base encoders (e_v, e_q), the multimodal interaction between vision and language (f_{θ_J}), or the reasoning network (l_{θ_L}) over the joint embedding. Here, we use Greedy Gradient Ensemble (GGE) [16] as our

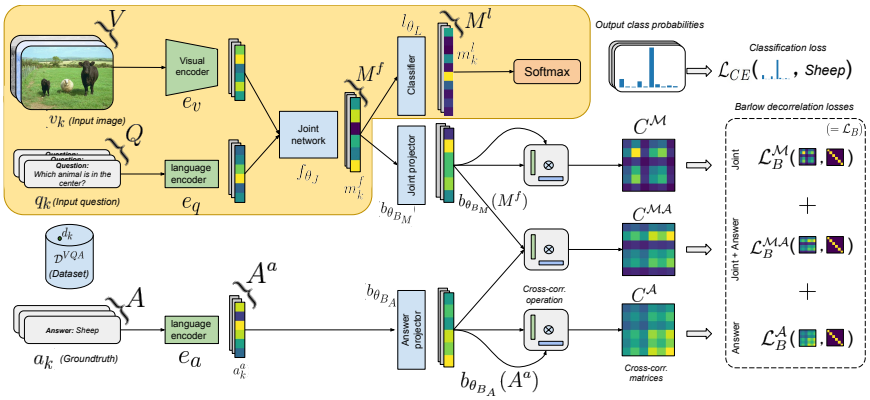

Figure 2: **Overall model**: We present the overall COB model, with both classification loss and Barlow redundancy reduction constraint. We explain notations and corresponding components in detail, in section 3. We also provide a glossary of all the used notations in the supplementary. All the θ parameters are learned, while the encoders $\{e_v, e_q, e_a\}$ are pre-trained models. During evaluation we only use the classification stream (in yellow) and drop the joint and answer projectors.

baseline and use it as our backbone VQA model. The GGE-DQ method optimizes both distribution bias and question short-cut bias. It first optimizes a loss between the logit value of a question-only model with the gradient of distributional bias, and then in a second stage, it obtains a loss between the answer logit of the VQA model with a gradient of both distribution bias and question short-cut bias, as discussed in eq. 16 in [16]. The joint network of the GGE model can be approximated as f_{θ_J}. Hence, our objective function to optimize is the cross-entropy ($\mathcal{L}_{CE}$) in eq. 3.

3.3. Baseline model with Barlow loss

First, we combine the cross-entropy objective function $\mathcal{L}_{CE}$ with a decorrelation loss, see Fig. 2. For a set of encoded question and image representations, $e_q(Q)$ and $e_v(V)$, we obtain a joint representation M^f using eq. 1. This joint representation M^f becomes one of the two modalities which we want to decorrelate. The second modality is the answer space encoded by $A^a = e_a(A)$. We then compute three decorrelation losses: unimodal joint embedding loss $\mathcal{L}_B^{\mathcal{M}}$, unimodal answer embedding loss $\mathcal{L}_B^{\mathcal{A}}$ and a multimodal embedding loss $\mathcal{L}_B^{\mathcal{MA}}$:

$$C^{\mathcal{M}} = C(b_{\theta_{B_M}}(M^f), b_{\theta_{B_M}}(M^f)) \tag{8}$$

$$C^{\mathcal{A}} = C(b_{\theta_{B_A}}(A^a), b_{\theta_{B_A}}(A^a)) \tag{9}$$

$$C^{\mathcal{MA}} = C(b_{\theta_{B_M}}(M^f), b_{\theta_{B_A}}(A^a)) \tag{10}$$

$$\mathcal{L}_B^{\mathcal{O}} = \left\{ \sum_i (1 - C_{ii}^{\mathcal{O}})^2 + \gamma \sum_i \sum_j (C_{ij}^{\mathcal{O}})^2 \right\}_{\mathcal{O} \in \{\mathcal{M}, \mathcal{A}, \mathcal{MA}\}} \tag{11}$$

$$\mathcal{L}_B = \mathcal{L}_B^{\mathcal{M}} + \mathcal{L}_B^{\mathcal{A}} + \mathcal{L}_B^{\mathcal{MA}} \tag{12}$$

where C(.) is defined in eq. 5. Hence, the overall loss $\mathcal{L}_{all_{base}}$ for our baseline model becomes:

$$\mathcal{L}_{all_{base}} = \mathcal{L}_{CE} + \mathcal{L}_B \tag{13}$$

Here, the first loss term is to enforce the discriminative property on the joint features m_k^f, while the second term reduces correlation between the feature components in both projected answer space and the joint image-and-question space. The gradient of the loss term $\mathcal{L}_B$ in eq. 13, is back-propagated to update f_{θ_J}, which optimizes its parameters to learn the joint representations m_k^f to become less redundant. This results in a joint embedding space that is discriminant and informative.

3.4. Balancing the two losses

Contrary to our initial expectations, we observed that, when optimizing the overall loss defined in equation 13, the classification performance actually decreases, (see Section 4.2). We conjecture that this decrease in performance occurs because of the difference in the dynamic range of the two loss terms. These losses are non-complementary and promote different properties in the embedding space. While cross-entropy makes the joint embedding space more discriminative, decorrelation tries to make the feature components orthogonal. An optimal weighing of the two loss terms is needed to ensure a rich representation that is discriminative while being informative. We propose two different approaches to achieve this:

a) Align then Barlow (ATB): This is our intermediate model, given to better understand the dynamics between the cross-entropy loss and the decorrelation constraints. In this setup, the VQA network is first pre-trained with the cross-entropy loss for n number of epochs and then fine-tuned with both loss terms, equation 13, till the loss converges. The resulting loss $\mathcal{L}_{all_{ATB}}$ can be written as:

$$\mathcal{L}_{all_{ATB}} = \begin{cases} \mathcal{L}_{CE}, & \text{if } epoch \leq n \\ \frac{1}{2}(\mathcal{L}_{CE} + \mathcal{L}_B), & \text{otherwise} \end{cases} \tag{14}$$

On analysing the Barlow twins [46] evaluation loss curve, we observe that the Barlow loss requires a large number of epochs to converge ($\sim$ 1000). This suggests that the Barlow twins loss surface is flatter requiring more gradient cycles to converge. Therefore a pre-training step to learn a meaningful representation can expedite the convergence as

orthogonalization of learned features can be viewed as rotating them in the representation space. In contrast, for a randomly initialized feature space, the network has to learn meaningful features and perform rotation simultaneously.

b) Constrained optimization using Barlow's theory (COB): The Barlow decorrelation loss on a randomly initialized joint embedding space is orders of magnitude larger than the cross-entropy, as shown in Figure 3. This high imbalance in the losses forces the network to move towards decorrelation optimization, and as discussed before, the decorrelation loss surface is flatter and hence the network does not converge when having a high loss imbalance. However, if the network is pre-trained with cross-entropy loss for certain number of epochs, the Barlow decorrelation loss decreases swiftly. This calls for a dynamic weighing scheme which changes based on the degree of imbalance between the two losses. Inspired by [38], we propose a constrained optimization formulation of equation 13 to dynamically control the weights assignment to the two loss terms:

$$\mathcal{L}_{all_{COB}} = \mathcal{L}_{CE}; \quad \text{subject to} \quad \mathbb{C}^t \leq 0 \tag{15}$$

$$\mathbb{C}^t = \alpha \mathbb{C}^{t-1} + (1-\alpha)(\mathcal{L}_B - \kappa) \tag{16}$$

where $\mathbb{C}^t$ captures the momentum of Barlow constraint $\mathcal{L}_B$ per epoch with α being the momentum factor and κ is a tolerance hyperparameter [38]. The above equation 15 can be rewritten as a non-constrained optimization problem:

$$\mathcal{L}_{all_{COB^\lambda}} = \mathcal{L}_{CE} + \lambda_t \mathbb{C}^t \tag{17}$$

$$\lambda_t \leftarrow \lambda_{t-1} \exp(\mathbb{C}^t) \tag{18}$$

where λ_t is the Lagrange multiplier (λ) at iteration t. The change in λ is directly proportional to the exponential of the magnitude of the Barlow constraint. Here, λ_t is initialized with a small value to bring both the loss terms in a similar range. While $\mathcal{L}_B$ itself consists of three loss terms, equation 12, we use a single value of λ_t to weight all of them, as their values vary in a similar range. This simplifies the overall formulation and reduces the number of non-gradient parameters (λ) to update.

4. Experiments

Evaluation Metric: We use the answering accuracy, the standard evaluation metric for VQA [4], to evaluate all our models. We use another metric Correctly Grounding Difference (CGD) [16], which is the difference of CGR[41] (Correct Grounding for Right prediction) and CGW (Correct Grounding but Wrong prediction) to evaluate the visual grounding of a method. To evaluate our proposed model we conduct experiments on the standard VQA v2 [15] and language-bias sensitive VQA-CP v2 [2] datasets. We discuss more about the datasets in the supplementary.

4.1. Training details

We train our COB model using the classification loss and the Barlow loss in an end-to-end manner. We use GGE-DQ-iter [16] as our base model. To update model parameters, we use the AdaMax [26] optimizer with configured hyperparameter values as follows: {learning rate = 0.001, batch size = 512, beta = 0.95, alpha = 0.99 and epsilon = 1e-8}. To train this COB model, we configure the hyperparameters of the constraint formulation as follows: λ is a learnable parameter, and it updates based on the moving average of the constraint loss as discussed in Section 3.4. We initialize with $\lambda_{init} = 0.0001$. The value of λ updates after every 100 iterations (called step size), based on the Barlow constraint loss value. The constraint loss depends on the previous constraint value and current value with a factor of $\alpha = 0.99$ and $(1-\alpha)$ as shown in equation 16. Initially, the λ value starts increasing, and after the Barlow loss ($\mathcal{L}_B$) reaches the threshold value ($\kappa = 2.63$), it starts decreasing as shown in Figure 3. More details about the model architecture are provided in the supplementary.

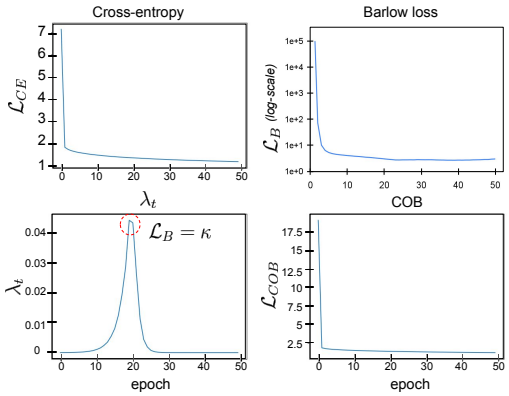

Figure 3: We plot the loss functions for our COB model during training along with Lagrange multiplier λ.

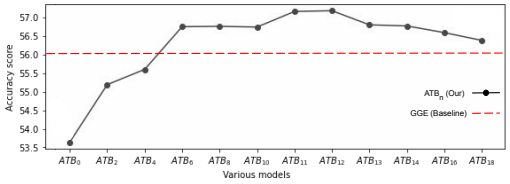

Figure 4: Ablation analysis: Applying Barlow loss after certain epoch. Individual scores for Y/N, Number and "Other" scores are present in supplementary. (In this figure, A_n stands for ATB_n, n is the number of pre-trained epochs)

4.2. Ablation: Epoch analysis for ATB

In this section, we discuss the effect of the pre-training epochs for the ATB model on the final VQA performance. This analysis is critical as it demonstrates that naive addition of the two loss terms, as in equation 13, is not the best training policy. Figure 4 shows the performance of our ATB model at convergence for different pre-training epochs. Without pre-training, a drop in performance of more than 2% can be observed. When the model is fine-tuned on lesser amount of pre-training ($n < 11$), the performance is inferior at convergence. As the initial loss of Barlow decorrelation is orders of magnitude higher, and the

Table 1: SOTA: VQA-CP v2 accuracy on test-set and VQA v2 accuracy on val set. Methods with * use extra annotations (e.g., human attention (HAT) [10], explanations (VQA-X) [34], or object label information). GGE-iter (impl.) is our implementation of GGE-DQ-iter[16] model. We sort Table-1 based on VQA-CP v2 scores.

	VQA-CP v2 [2] test					VQA v2 [15] val			
Models	All	Y/N	Number	Other	CGD	All	Y/N	Number	Other
CSS(UpDn)* [8]	41.16	43.96	12.78	47.48	8.23	59.21	72.97	40.00	55.13
AdvReg.[36]	41.17	65.49	15.48	35.48	-	62.75	79.84	42.35	55.16
RUBi [7]	45.42	63.03	11.91	44.33	6.27	58.19	63.04	41.00	54.43
Hint*[40]	47.50	67.21	10.67	46.80	10.34	63.38	81.18	42.14	**55.66**
GVQE*[28]	48.75	-	-	-	-	**64.04**	-	-	-
LM [9]	48.78	70.37	14.24	46.42	11.33	63.26	81.16	42.22	55.22
DLP [22]	48.87	70.99	18.72	45.57	-	57.96	76.82	39.33	48.54
SCR* [42]	49.45	72.36	10.93	48.02	-	62.20	78.8	41.6	54.4
LMH[9]	52.73	72.95	**31.90**	47.79	10.60	56.35	65.06	37.63	54.69
CF-VQA[33]	53.69	**91.25**	12.80	45.23	-	63.65	**82.63**	**44.01**	54.38
GGE-iter[16]	57.12	87.35	26.16	**49.77**	**16.44**	59.30	73.63	40.30	54.29
GGE-iter (impl.)	56.08	86.64	22.15	**49.38**	15.92	58.92	72.00	40.13	53.95
COB(ours)	**57.53**	**88.36**	28.81	49.27	**16.89**	**63.80**	**81.36**	**43.30**	**55.86**
CSS(LMH)*[8]	**58.21**	83.65	**40.73**	48.14	8.81	53.15	61.20	37.65	53.36

two loss terms are non-complementary, the resulting gradient for cross-entropy loss is relatively weaker to learn good discriminative features. We also observe that the accuracy increases with increase in pre-training epochs, this happens as the loss for Barlow decorrelation for a pre-trained feature space converges faster. Since for a pre-trained feature space, decorrelation is analogous to rotating the feature components towards their orthogonal principal axis, the Barlow decorrelation loss finds it easier to converge. This results in gradients for both cross-entropy loss and Barlow decorrelation to be comparable, and hence results in learning a richer feature space. Finally, we see a drop in performance, for a larger pre-training epoch ($n > 12$). For a larger number of pre-training epochs, the validation cross-entropy loss starts to overfit and the non-complementary Barlow decorrelation loss no longer improves the performance.

Table 2: Ablation analysis of our approach

Method	$\mathcal{L}_{CE}$	$\mathcal{L}_B^{\mathcal{M}}$	$\mathcal{L}_B^{\mathcal{M},\mathcal{A}}$	$\mathcal{L}_B^{\mathcal{A}}$	All	Y/N	Number	Other
GGE	✓				56.08	86.64	22.15	49.38
$COB^{\mathcal{M}}$	✓	✓			57.03	87.17	26.67	49.57
$COB^{\mathcal{M},\mathcal{A}}$	✓		✓		56.77	86.84	24.83	49.75
$COB^{\mathcal{M},\mathcal{M},\mathcal{A}}$	✓	✓	✓		57.49	86.57	**30.12**	**49.77**
COB	✓	✓	✓	✓	**57.53**	**88.36**	28.81	49.27

4.3. Ablation of the proposed approach

Our constraint formulation $\mathcal{L}_B$ consists of three loss terms $\mathcal{L}_B^{\mathcal{M}}$, $\mathcal{L}_B^{\mathcal{A}}$ and $\mathcal{L}_B^{\mathcal{M},\mathcal{A}}$, equation 12. To understand the importance of each of these loss terms, we ablate them individually in the constraint and re-train the COB model. For the model with only $\mathcal{L}_B^{\mathcal{M}}$ loss, i.e. $COB^{\mathcal{M}}$, the answering accuracy is 57.03%, better than the baseline GGE model, as shown in Table 2. This shows that increasing information content (or minimizing the redundancy) of the joint features helps VQA performance. $COB^{\mathcal{M},\mathcal{A}}$, that contains the constraint term $\mathcal{L}_B^{\mathcal{M},\mathcal{A}}$, forces the model to learn an alignment between the answer and the joint features in the projected Barlow space while maintaining the decorrelation between the feature components. The gradients from $\mathcal{L}_B^{\mathcal{M},\mathcal{A}}$ provide an additional supervision along with $\mathcal{L}_{CE}$ to help the underlying joint embedding space m_k^f to learn features relevant to the answer, resulting in an answering performance of 56.77%, Table 2. Combining the these two constraint terms, $\mathcal{L}_B^{\mathcal{M}}$ and $\mathcal{L}_B^{\mathcal{M},\mathcal{A}}$, in $COB^{\mathcal{M},\mathcal{M},\mathcal{A}}$ results in an increased performance of 57.49%. Finally, COB model contains all three loss terms, the additional $\mathcal{L}_B^{\mathcal{A}}$ improves the information content of the answer embedding. This further assists the $\mathcal{L}_B^{\mathcal{M},\mathcal{A}}$ loss to learn a better alignment between the less redundant joint and the answer embedding spaces, outperforming the other three ablated baselines. This ablation analysis shows that each of the three loss terms in our constraint provides a different supervision to the model and thereby improves the underlying joint representations.

4.4. Comparison with state-of-the-art

We provide performance results on two datasets, challenging VQA-CP v2[2] that has a less language bias and a standard VQA v2[15] dataset in Table 1. CSS[8] & CF-VQA[33] use counterfactual examples to overcome bias, AdvReg[36] uses regularisation techniques, HINT[40] & SCR[42] use grounding techniques, RUBi[7], LM[9] and GGE[16] use ensemble methods, GVQE [28] & DLP[22] use new encoder based method to overcome language and dataset bias. Some methods use extra annotations to improve debiasing performance, but our method does not use any extra annotations and performs better than most current state-of-the-art (SOTA) methods with better explainability in the results (see Section 5.1). Our implementation of GGE model performance is 56.08% and 58.92% on VQA-CP v2 and VQA v2 datasets respectively. In comparison, our COB model, built upon the base GGE model, obtains a performance of 57.53% and 63.80%, which is an improvement of 1.4% and 4.9% respectively. We also outperform the official GGE [16] performance. Our COB model out-

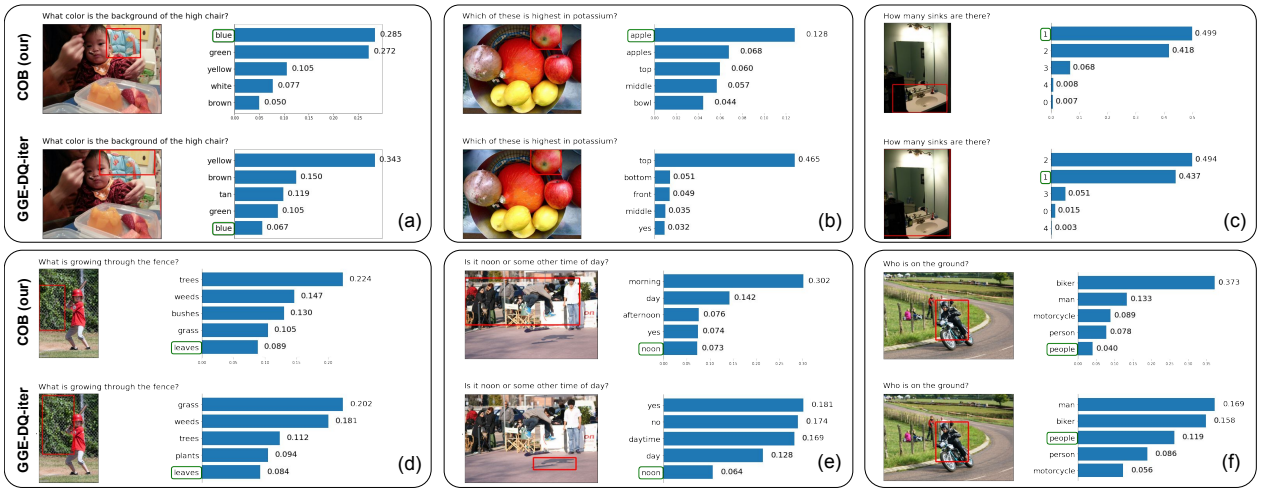

Figure 5: **Qualitative results:** Each set of images show the input image-question pair and top-5 predictions for our proposed COB model compared against the baseline GGE-DQ-iter model. Red bounding box shows the maximal attention region in each image. Answers within the green boxes are the ground truths. We see that COB performs better with higher prediction score to the ground truth answer in comparison to the baseline method (a)-(e). For negative results (d)-(f) as well, the predicted classes are semantically relevant. This is further analyzed in the context of the explainability in section 5.1. We provide more qualitative results in supplementary.

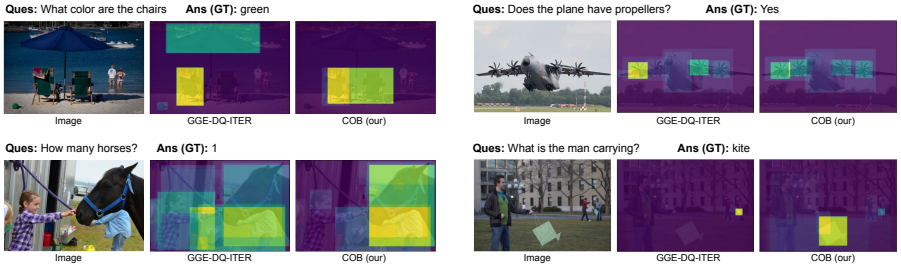

Figure 6: **Explainability of the models:** Given an image and a question, we show the class activation maps for the samples in the joint embedding space m_k^j corresponding to the answer. We observe that the COB model's Grad-CAM outputs are better localized in the salient regions for answering the question. More examples are shown in supplementary.

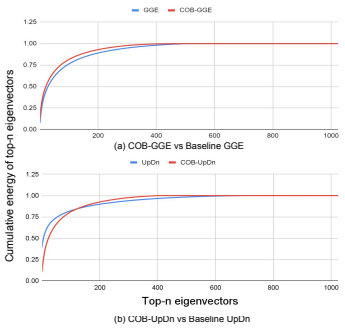

Figure 7: **PCA analysis COB vs baselines**

performs other SOTA methods, that do not use extra annotations, for overall answer prediction task on both VQA-CP v2 and VQA v2 datasets, as presented in Table 1. We also improve overall CGD score by 0.45 units, which shows that our model is able to learn a better grounding between vision-and-language modalities.

4.5. Qualitative results

Figure 5 illustrates top-5 answers and probability scores for a few examples. We compare our qualitative results with the most recent state-of-the-art method GGE-DQ [16]. In the first and third examples (part a and c of Figure 5), our COB model attends a more precise salient region leading to a correct answer as compared to GGE-DQ model whose attention region extends over a larger non-salient region, thus answering incorrectly. For the second image, both the models focus on the same region, however COB assigns a higher probability score to the correct answer. These results indicate that the more informative latent features provide better reasoning, improving the localization and the probability scores for the correct answers compared to the baseline method. Similarly, we show results for various combinations of attention and answer prediction results.

5. Analysis and Discussion

5.1. Explainability: Grad-Class activation maps

Reasoning is an essential part of question answering, and is directly influenced by the quality of the joint representation space. Hence it is crucial to study what the model has learned and how it processes the input data. This interpretability of a model is even more important for failure cases, in order to understand the cause of failure and model shortcomings. We use Grad-CAM[39] as an indicative of model interpretability by computing the saliency for an im-

	Cumulative energy for top-k PCA components (in%)			
N_B	k=512	k=256	k=128	k=64
512	100	99.8	99.3	94.1
1024	99.9	99.5	98.1	80.4
2048	99.7	98.5	91.1	65.0
4096	98.8	95.7	85.2	59.9

Table 3: **Projector dimensionality (N_B) selection:** PCA energy for top-k components for different Barlow projection (b_{θ_B}) dimensions.

age given the question and the ground truth answer. We analyze our COB and the SOTA baseline model [16], trained on VQA-CP v2, in the context of model interpretability in Figure 6. We observe that our model produces more interpretable regions compared to the baseline GGE model, which also indicates the reason for a higher CGD score in Table 1. For both examples, our model focuses on correct regions that are salient for the answer prediction.

5.2. Redundancy, information and VQA:

COB aims to reduce redundancy in the Barlow space, and in turn makes the joint representation (i.e. the output of the joint projection layer on top of the fixed encoders) less redundant. A less redundant joint projector would capture the least redundant information from the output space of fixed encoders and project them into the joint representation space. We perform PCA analysis in this joint projection space for GGE and COB-GGE models. We observe that for COB-GGE, top-350 eigenvectors amounts to 99% energy, against top-440 for GGE, Figure 7. For another base model, UpDn [3], that uses top-556 eigenvectors to capture 99% energy, our COB-UpDn variant uses top-349 eigenvector. This shows COB forces the joint space to capture least redundant information from the fixed encoder spaces. Hence, only most informative features are captured. It also means the effective remaining capacity of the joint space is increased. In other words, more information (additional data; a possible future research direction) can be modelled in the same number of feature dimensions or the same amount of information (fixed encoder output space; i.e. our case) can be modelled in a lesser number of feature dimensions.

5.3. Projector dimensionality selection

Zbontar *et al*. [46] show that increase in the projector's (b_{θ_B}) output dimensions (N_B) improves the input self-supervised feature space. However, for our Barlow decorrelation constraint we found that larger projection spaces, $\mathcal{D}^B \rightarrow \mathbb{R}^{N_B}$ for $N_B \in \{1024, 2048, 4096\}$, have more redundant components. To analyse this, we compute the PCA eigenvalues for the representations for higher projection spaces, as shown in Table 3. We observe that top-512 components can preserve $\sim 99\%$ of total energy of the embedding space and hence we choose $N_B = 512$ as the projection dimension for the Barlow projectors (i.e. $b_{\theta_{B_M}}, b_{\theta_{B_A}}$). In supplementary, we provide more ablation analysis and pseudo-code for our methods.

Methods	(test)
MAML[12]	59.6
MEVF[32]	62.7
MMQ[11]	67.7
QCR[47]	71.6
COB-QCR	**71.9**

Table 4: VQA-Rad dataset

Methods	GQA(testdev)
MAC	41.2
COB-MAC	**42.1**

Table 5: GQA dataset. (Base repository, MAC: https://github.com/ronilp/mac-network-pytorch-gqa)

5.4. Generalizability

Here, we evaluate our COB method on two more datasets: a real world visual reasoning dataset (GQA[20]) and a dataset of clinically generated VQA about radiology images (VQA-Rad[29]). We compare the COB method on QCR model on VQA-Rad dataset and on MAC model for GQA dataset as shown in the Table 4 and 5 respectively. COB works well in both. In sections 4.2 and 4.5, we show that ATB and COB models built upon the base GGE outperforms it by learning a more informative latent space, Figure 6. GGE [16] model is the SOTA for VQA, and hence while improving it validates our proposed models, it also raises the question if the improvement in the results only comes due to the better latent features of the base GGE model; i.e. does the improvement in results is dependent on the better quality of the base model. To study this we apply ATB and COB constraints on the UpDn [3] model, which itself is the base of GGE model. The resulting ATB-UpDn and COB-UpDn models outperform (answering accuracy: 47.36% and 48.24% respectively) the base UpDn model (39.38%) by a significant margin on VQA-CP v2. This shows that our constraint formulation, despite being limited by the quality of the base model, imposes a regularization on the latent features to be more informative, resulting in an improved performance over the corresponding baseline.

6. Conclusion

We propose a new VQA regularization scheme called COB that optimizes cross-entropy loss while subjected to a redundancy minimization constraint. Cross-modal Barlow decorrelation loss as the constraint formulation promotes the alignment of the answer with image-and-question modalities while improving the information content of the underlying feature space. We propose two training policies, ATB and COB, to balance these two losses. We show that both ATB and COB outperform the most recent SOTA baseline (GGE), Table 2, on VQA-CP v2 and VQA v2 datasets for the answer prediction task. COB model also either outperforms or provides comparative results against other competing baselines, Table 1 without using additional annotations. Finally, Figure 6 shows that our model focuses more on the salient regions while answering the questions, hence being more interpretable.

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6430–6439, 2019.

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018.

[3] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01), 1961.

[6] Piotr Bielak, Tomasz Kajdanowicz, and Nitesh V Chawla. Graph barlow twins: A self-supervised representation learning framework for graphs. *arXiv preprint arXiv:2106.02466*, 2021.

[7] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32:841–852, 2019.

[8] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809, 2020.

[9] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, 2019.

[10] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *EMNLP*, 2016.

[11] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 64–74. Springer, 2021.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing*, pages 457–468. ACL, 2016.

[14] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304, 2015.

[15] 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.

[16] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1584–1593, 2021.

[17] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[18] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[19] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992.

[20] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[22] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11181–11188, 2020.

[23] Yannis Kalantidis, Carlos Lassance, Jon Almazan, and Diane Larlus. Tldr: Twin learning for dimensionality reduction. *arXiv preprint arXiv:2110.09455*, 2021.

[24] Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. Contrast and classify: Training robust vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1604–1613, 2021.

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

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[28] Gouthaman KV and Anurag Mittal. Reducing language biases in visual question answering with visually-grounded question encoder. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 18–34. Springer, 2020.

[29] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[30] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[31] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[32] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 522–530. Springer, 2019.

[33] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021.

[34] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788, 2018.

[35] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[36] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, 2018.

[37] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2953–2961, 2015.

[38] Danilo Jimenez Rezende and Fabio Viola. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.

[39] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[40] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2591–2600, 2019.

[41] Robik Shrestha, Kushal Kafle, and Christopher Kanan. A negative case analysis of visual grounding methods for vqa. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8172–8181, 2020.

[42] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. *Advances in Neural Information Processing Systems*, 32:8604–8614, 2019.

[43] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[44] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[45] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.

[46] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

[47] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354, 2020.