Contrast and Classify: Training Robust VQA Models

Yash Kant[1*] Abhinav Moudgil[1] Dhruv Batra[1,2] Devi Parikh[1,2] Harsh Agrawal[1]
[1]Georgia Institute of Technology [2] Facebook AI Research

Abstract

Recent Visual Question Answering (VQA) models have shown impressive performance on the VQA benchmark but remain sensitive to small linguistic variations in input questions. Existing approaches address this by augmenting the dataset with question paraphrases from visual question generation models or adversarial perturbations. These approaches use the combined data to learn an answer classifier by minimizing the standard cross-entropy loss. To more effectively leverage augmented data, we build on the recent success in contrastive learning. We propose a novel training paradigm (ConClaT) that optimizes both cross-entropy and contrastive losses. The contrastive loss encourages representations to be robust to linguistic variations in questions while the cross-entropy loss preserves the discriminative power of representations for answer prediction.

We find that optimizing both losses – either alternately or jointly – is key to effective training. On the VQA-Rephrasings [44] benchmark, which measures the VQA model's answer consistency across human paraphrases of a question, ConClaT improves Consensus Score by 1.63% over an improved baseline. In addition, on the standard VQA 2.0 benchmark, we improve the VQA accuracy by 0.78% overall. We also show that ConClaT is agnostic to the type of data-augmentation strategy used.

1. Introduction

Visual Question Answering (VQA) refers to the task of automatically answering free-form natural language questions about an image. For VQA systems to work reliably when deployed in the wild, for applications such as assisting visually impaired users, they need to be robust to different ways a user might ask the same question. For example, VQA models should produce the same answer for two paraphrased questions – "What is in the basket?" and "What is contained in the basket?" since their semantic meaning is the same. While significant progress has been made towards building more accurate VQA systems, these models remain brittle to minor linguistic variations in the input question.

*Correspondence to ysh.kant@gmail.com

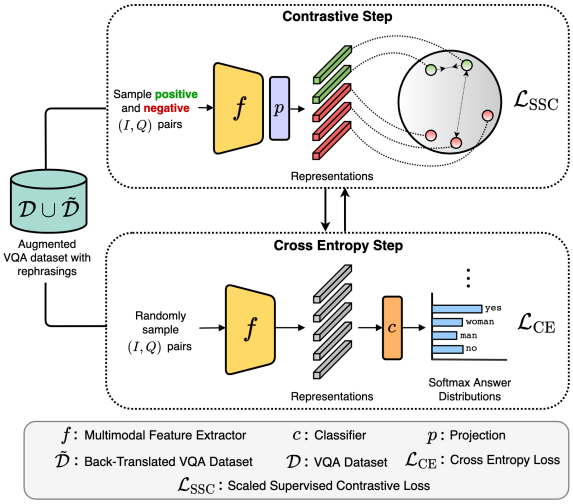

Figure 1: We make VQA model robust to question paraphrases using a training paradigm ConClaT that minimizes contrastive and cross-entropy losses together. Contrastive learning step pulls representations of positive samples corresponding to paraphrased questions closer together while pushing those with different answers farther apart. Cross-entropy step makes these representations discriminative to help model answer visual questions accurately.

To make VQA systems robust, existing approaches [44, 47] have trained VQA systems [24] by augmenting the training data with different variations of the input question. For instance, VQA-CC [44] use a visual question generation (VQG) model to generate paraphrased question given an image and answer. Generally, these models fuse image and question features into a joint vision and language (V+L) representation followed by a standard softmax classifier to produce answer probabilities and are optimized by minimizing the cross-entropy loss. Unfortunately, cross-entropy loss treats every image-question pair independently and fails to exploit the information that some questions in the augmented dataset are paraphrases of each other.

We overcome this limitation by using a contrastive loss InfoNCE [36] that encourages joint V+L (Vision and Language) representations obtained from samples whose questions are paraphrases of each other to be closer while pulling apart the V+L representations of samples with different an-

swers. As we operate in a supervised setting, we choose Supervised Contrastive Loss (SCL) [26] which extends InfoNCE by utilizing the label information to bring samples from the same class (ground-truth answer) together. We introduce a variant of the SCL which emphasizes rephrased image-question pairs over pairs that are entirely different but have the same answer. Our proposed training paradigm, ConClaT (**Con**trast and **Cla**ssify **T**raining), minimizes SCL and cross-entropy loss together to learn better vision and language representations as shown in Fig.1. Minimizing the contrastive loss encourages representations to be robust to linguistic variations in questions while the cross-entropy loss preserves the discriminative power of the representations for answer classification. Instead of pretraining with SCL, then fine-tuning with cross-entropy loss as in [26], we find that minimizing the two losses either alternately or jointly by constructing loss-specific mini-batches helps learn better representations. For contrastive loss, we carefully curate mini-batches by sampling various types of negatives and positives given a reference sample.

We show the efficacy of our training paradigm across two rephrasing (i.e., data-augmentation) strategies. Using rephrasings obtained from a VQG model proposed in [44], our approach outperforms a baseline that simply treats these rephrasings as additional samples and ignores the link between question and its paraphrases. We noticed that the VQG model fails to produce a diverse set of rephrasings for a question. Hence, we use Back-translation to obtain question rephrasings. Back-translation [15] involves translating an input sentence from one language to another and then translating it back into the original language using a pair of machine translation models (e.g. `en-fr` and `fr-en`). We find that Back-translation preserves the semantic meaning of the question while generating syntactically diverse question. Utilizing the publicly available collection of neural machine translation models in HuggingFace [52], we generate numerous rephrasings of every question. Then, we filter poor/irrelevant rephrasings with a sentence similarity model [41] and store 3 rephrasings per original question of VQA v2.0 dataset without any manual supervision.

We extensively ablate ConClaT with alternate [8], joint and pretrain-finetune [26] training schemes, and compare with previously proposed triplet [38] and margin-based losses [58] . We evaluate on the VQA Rephrasings benchmark [44] which measures the model's answer consistency across several rephrasings of a question. ConClaT improves Consensus Score by 1.63% over an improved baseline. In addition, on the standard VQA 2.0 benchmark, we improve VQA accuracy by 0.78% overall. It is also worth noting that VQA models trained using ConClaT perform better than existing approaches across both the aforementioned data-augmentation strategies – Back-translation and VQG.

2. Related Work

Models for VQA. Several models have been proposed for Visual Question Answering which fuse CNN grid features and LSTM features with different forms of attention [34, 55, 16, 23]. Bottom-Up and Top-Down [6] proposed to learn attention over object regions obtained from a pretrained object detector and subsequent works [27, 56, 24] introduced various ways to fuse image and language representations. Recent works [32, 33, 45, 29, 45, 46, 13] use multi-modal transformers to learn visuo-linguistic representations from object detector features and BERT question features [14]. We use the multi-modal transformer architecture similar to UNITER [13] for all our experiments.

Robustness of VQA Models. Robustness of VQA models with respect to multi-modal vision and language input has been studied in great detail. [18, 57] proposed balanced datasets to ensure models don't overfit to language while answering visual questions. C-VQA [4] and VQA-CP [3] datasets were proposed to test robustness against changing question-answer distributions. SQUINT [43] encouraged consistency between reasoning questions and associated sub-questions. Our work focuses on robustness to question paraphrases in VQA-Rephrasings [44] that were collected from human annotators. VQA-CC [44] trained a Visual Question Generation (VQG) model to generate paraphrases of questions to augment the training dataset while VQA-Aug [47] augmented the training dataset by generating paraphrases of questions via back-translation. We show that these data augmentation techniques can be better utilized via ConClaT to build robust and accurate VQA models. Concurrent to our work, Whitehead *et al*. [50] propose a rule-based mechanism to generate question paraphrases for VQA. They constrain their model architecture to be modular [7] and use module-level loss to improve consistency. In contrast, our approach is agnostic to model architecture.

Various works [3, 2, 58, 38] made VQA models robust to language bias (For example, "What is the color of x" will always produce 'blue' irrespective of x). Recent works [48, 1, 9, 37] also studied robustness from counterfactual answering lens – answer should change according to the change in semantic content of the question or image. Our work, on the other hand, focuses on robustness to *syntactic* variations in questions.

Paraphrase Generation in NLP. There has been significant work in the area of Natural Language Processing (NLP) for generating paraphrases of a sentence using LSTM networks [39], Deep Reinforcement Learning [31], Variational Autoencoders [19] and Transformers [49]. However, these works require supervision in the form of paraphrase pairs. In order to mitigate this limitation of labelled data, Neural Machine Translation (NMT) models have been used to generate paraphrases in a self-supervised fashion via back-translation [35, 51]. We build

on top of these works and use state-of-the-art NMT models from HuggingFace [52] to generate paraphrases for visual questions without any supervision.

Contrastive Learning. There has been recent interest in the use of Contrastive Learning for learning visual representations in a self-supervised manner [53, 22, 21, 10, 12, 11, 40]. Going beyond Image Classification, recently, [20] used contrastive learning for phrase grounding. They used the InfoNCE loss [36] to learn a compatibility function between a set of region features from an image and contextualized word representations. In contrast, we want to learn representations which are robust to linguistic variations in the question for VQA.

To utilize label information in contrastive losses, [26] proposed Supervised Contrastive Learning (SCL) loss for learning *visual* representations. We introduce a variant of the SCL which scales the contributions from augmented positive samples (rephrasings in our case) over intra-class positive samples (that have the same answer) using a scaling factor. Morever, our training paradigm optimizes both (cross-entropy and SCL) losses together, whereas [26] follow the pretrain-finetune training scheme. Furthermore, [26] randomly sample positive and negative pairs based on label information, whereas we carefully curate batches by sampling hard-negatives from the dataset. We show how these differences affect performance through a series of ablations in our experiments section.

3. Preliminaries

In this section, we introduce the VQA task and the standard cross entropy training of VQA models. We then recap contrastive methods for learning representations [10] and the recently proposed Supervised Contrastive Learning (SCL) [26] setup. We describe our approach in section 4.

VQA. The task of Visual Question Answering (VQA) [5, 18] involves predicting an answer a for a question q about an image v. An instance of this problem in the VQA Dataset $\mathcal{D}$ is represented via a tuple $x = (v, q, a), \forall x \in \mathcal{D}$. Recent VQA models [24, 6, 13] take image and question as input and output a joint vision and language (V+L) representation $\boldsymbol{h} \in \mathcal{R}^{d_h}$ using a multi-modal network f:

$$\boldsymbol{h} = f(v, q)$$

The V+L representation $\boldsymbol{h}$ is then used to predict a probability distribution over the answer space $\mathcal{A}$ with a softmax classifier $f^c(\boldsymbol{h})$ learned by minimizing the cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp(f^c(\boldsymbol{h})[a])}{\sum_{a' \in \mathcal{A}} \exp(f^c(\boldsymbol{h})[a'])} \tag{1}$$

where $f^c(\boldsymbol{h})[a]$ is the logit corresponding to the answer a.

Contrastive Learning. Recent works in vision [10] have used contrastive losses to bring representations of two augmented views of the same image (called positives) closer together while pulling apart the representations of two different images (called negatives). The representation $\boldsymbol{h}$ obtained from an image encoder is projected into a d_z dimensional hyper-sphere using a projection network g such that $\boldsymbol{z} = g(\boldsymbol{h}) \in \mathcal{R}^{d_z}$. Given a mini-batch of size K, the image representation $\boldsymbol{h}$ is learned by minimizing the InfoNCE [36] loss which operates on a pair of positives $(\boldsymbol{z}_i, \boldsymbol{z}_p)$ and $K-1$ negative pairs $(\boldsymbol{z}_i, \boldsymbol{z}_k)$ such that $i, p, k \in [1, K], k \neq i$ as follows:

$$\mathcal{L}^i_{\text{NCE}} = -\log \frac{\exp(\Phi(\boldsymbol{z}_i, \boldsymbol{z}_p)/\tau)}{\sum_{k=1}^{K} \mathbb{1}_{k \neq i} \exp(\Phi(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}, \tag{2}$$

where $\Phi(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$ computes similarity between $\boldsymbol{u}$ and $\boldsymbol{v}$ and $\tau > 0$ is a scalar temperature parameter.

A generalization of InfoNCE loss to handle more than one positive-pair was proposed by [26] called Supervised Contrastive Loss (SCL). Given a reference sample x, SCL uses class-label information to form a set of positives $\mathcal{X}^+(x)$ that contains samples with the same label as x. $\mathcal{X}^+(x)$ also contains augmented views of the sample because they share the same label as x. For a minibatch with K samples, SCL is defined as:

$$\mathcal{L}^i_{\text{SC}} = -\sum_{p=1}^{|\mathcal{X}^+(x_i)|} \log \frac{\exp(\Phi(\boldsymbol{z}_i.\boldsymbol{z}_p)/\tau)}{\sum_{k=1}^{K} \mathbb{1}_{k \neq i} \cdot \exp(\Phi(\boldsymbol{z}_i.\boldsymbol{z}_p)/\tau)}$$

$$\mathcal{L}_{\text{SC}} = \sum_{i=1}^{K} \frac{\mathcal{L}^i_{\text{SC}}}{|\mathcal{X}^+(x_i)|} \tag{3}$$

Overall, $\mathcal{L}^i_{\text{SC}}$ tries to bring the representation of samples in $\mathcal{X}^+(x_i)$ closer together compared to representations of samples with a different ground-truth label.

4. Approach

We now describe our approach, ConClaT, which uses contrastive and cross-entropy training to learn VQA models robust to question paraphrases.

4.1. Augmented Dataset with Back-translation

We augment the train set with question paraphrases using 88 different MarianNMT [25] Back-translation model pairs released by HuggingFace [52]. We produce 27 *unique* rephrasings per question with cosine similarity of 0.88 on average, the similarity is calculated by first encoding the questions via Sentence-BERT [41]. We only select paraphrases that have ≥ 0.95 similarity with the original question and choose three unique paraphrases randomly from this subset. We use three paraphrases to keep the compute manageable. Overall, our augmented train set consists of $\sim$ 1.6M samples.

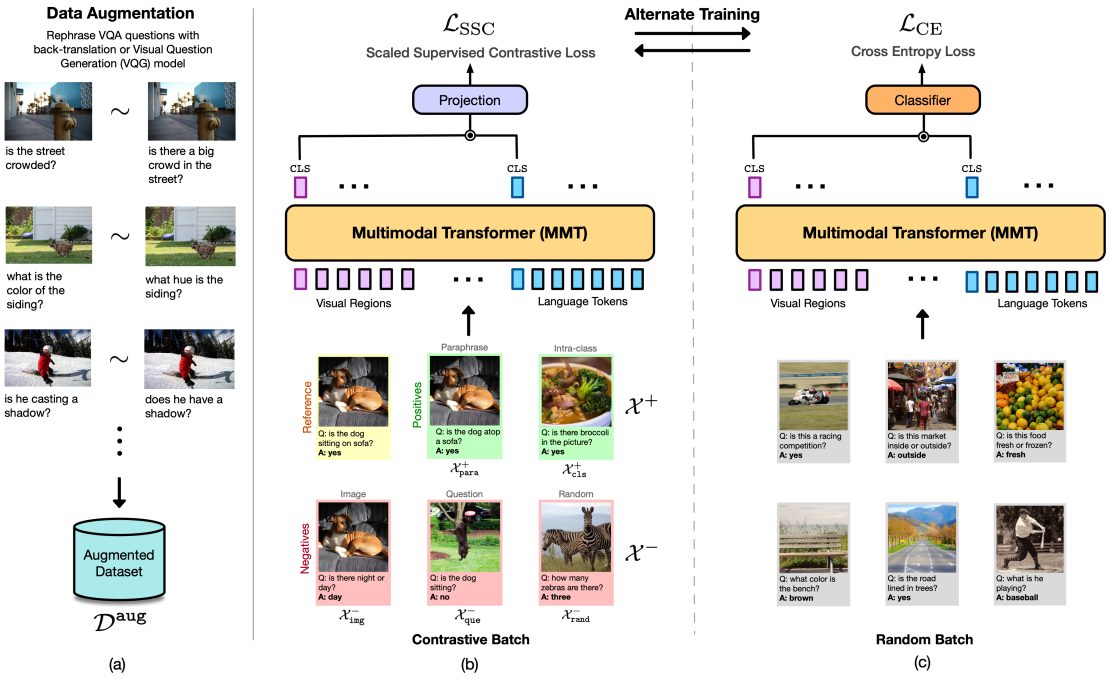

Figure 2: **Overview of ConClaT.** (a) We augment the VQA dataset by paraphrasing every question via Back-translation or Visual Question Generation. (b) We carefully curate a contrastive batch by sampling different types of positives and negatives to learn joint V+L representations by minimizing scaled supervised contrastive loss $\mathcal{L}_{\text{SSC}}$. (c) Cross Entropy loss $\mathcal{L}_{\text{CE}}$ is optimized with $\mathcal{L}_{\text{SSC}}$.

For a sample $x = (v, q, a) \in \mathcal{D}$, let's denote a set of paraphrases for question q by $\mathcal{Q}(q)$ and the corresponding set of VQA triplets as:

$$\mathcal{X}^+_{\texttt{para}}(x) = \{(v, q', a) \mid q' \in \mathcal{Q}(q)\} \tag{4}$$

As shown in Figure 2(a), we augment the VQA dataset $\mathcal{D}$ with multiple paraphrased samples of a given question and denote the augmented dataset $\mathcal{D}^{\texttt{aug}}$ as:

$$\mathcal{D}^{\texttt{aug}} = \mathcal{D} \bigcup_{x \in \mathcal{D}} \mathcal{X}^+_{\texttt{para}}(x) \tag{5}$$

4.2. Scaled Contrastive Loss for VQA

We would like our VQA model to produce the *same and correct* answer for a question and its paraphrase given an input image. This motivates us to map joint vision and language (V+L) representations of an original and paraphrased sample closer to each other. Moreover, since we operate in a supervised setting, following SCL [26] we also pull the joint representations for the questions with the same answer (intra-class positives) closer together while pulling apart the representations of questions with different answers. We define the set of all samples with the same ground truth answer as x by:

$$\mathcal{X}^+(x) = \{(\hat{v}, \hat{q}, \hat{a}) \in \mathcal{D}^{\texttt{aug}} \mid \hat{a} = a\} \tag{6}$$

Note that $\mathcal{X}^+_{\texttt{para}}(x) \subset \mathcal{X}^+(x)$ as all question paraphrases have the same answer for a given image but not all questions with the same answer are paraphrases. We refer to samples in set $\mathcal{X}^+_{\texttt{cls}}(x) = \mathcal{X}^+(x) - \mathcal{X}^+_{\texttt{para}}(x)$ as *intra-class* positives and set $\mathcal{X}^+_{\texttt{para}}(x)$ as *paraphrased* positives w.r.t. x as depicted in Figure 2(b).

Following Eq. (3), all the samples in $\mathcal{X}^+(x_i)$ in $\mathcal{L}_{\text{SC}}$ are treated the same. That is, representations from both the paraphrased positives and intra-class positives are brought closer together. To emphasize on the link between question and its paraphrase, we propose a variant of the SCL in Eq. (7) which assigns higher weight to paraphrased positives $\mathcal{X}^+_{\texttt{para}}(x)$ over intra-class positives $\mathcal{X}^+_{\texttt{cls}}(x)$. We introduce a scaling factor α_{ip} in the SCL (Eq. (3)) for a sample x_i as follows:

$$\mathcal{L}^i_{\text{SSC}} = -\sum_{p=1}^{|\mathcal{X}^+(x_i)|} \alpha_{ip} \cdot \log \frac{\exp(\Phi(\boldsymbol{z}_i.\boldsymbol{z}_p)/\tau)}{\sum_{k=1}^{K} \mathbb{1}_{k \neq i} \cdot \exp(\Phi(\boldsymbol{z}_i.\boldsymbol{z}_p)/\tau)} \tag{7}$$

$$\mathcal{L}_{\text{SSC}} = \sum_{i=1}^{K} \frac{\mathcal{L}^i_{\text{SSC}}}{\sum_p \alpha_{ip}} \tag{8}$$

The scaling factor α_{ip} assigns a higher weight $s > 1$ to positive samples corresponding to question paraphrases compared to other intra-class positives. Intuitively, because of the higher weight, the loss will penalize the model

Algorithm 1 ConClaT with alternate $\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$

input: steps N; constant N_{ce}; data $\mathcal{D}^{\texttt{aug}}$; networks f, g
  for all $i \in \{1, \ldots, N\}$ **do**
    $\mathcal{B} = \phi$
    if $i \pmod{N_{ce}} = 0$ **do**
      # sscl iteration
      $\mathcal{B} = \text{CURATE}(N_r, \mathcal{D}^{\texttt{aug}}, \boldsymbol{w})$; $\mathcal{L} = \mathcal{L}_{\text{SSC}}$
    else do
      # ce iteration
      $\mathcal{B} \sim \mathcal{D}^{\texttt{aug}}$; $\mathcal{L} = \mathcal{L}_{\text{CE}}$
    update $f(.), g(.)$ networks to minimize $\mathcal{L}$ over $\mathcal{B}$
return network $f(.)$; throw away $g(.)$

strongly if it fails to bring the representations of a question and its paraphrase closer. We define α_{ip} as:

$$\alpha_{ip} = \begin{cases} s & \text{if } x_p \in \mathcal{X}^{+}_{\texttt{para}}(x_i), \\ 1, & \text{otherwise} \end{cases} \tag{9}$$

4.3. Training with $\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$

We experiment with various schemes of combining supervision from $\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$ losses. Specifically, we try – alternate (Algorithm 1), joint, and pretrain-finetune [26] training schemes.

Our alternate training scheme is summarized in Algorithm 1. Specifically, given N total training iterations, we update our model with $\mathcal{L}_{\text{SSC}}$ after every $N_{ce}-1$ updates with $\mathcal{L}_{\text{CE}}$, where N_{ce} is a hyper-parameter . In the joint training scheme, we curate loss-specific batches for $\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$ but jointly update the model by accumulating the gradients of these two losses. Training alternately or jointly with the two losses simplifies the optimization procedure compared to two-stage training (pretrain-finetune as in [26]) which requires double the hyper-parameters and longer training iterations. Figure 2 depicts ConClaT training.

4.4. Negative Types and Batch Creation

SCL operates with multiple negative samples. For a given reference sample $x = (v, q, a) \in \mathcal{D}^{\texttt{aug}}$, we define a corresponding set of negatives as samples with ground truth different than the reference x:

$$\mathcal{X}^{-}(x) = \{(\bar{v}, \bar{q}, \bar{a}) \in \mathcal{D}^{\texttt{aug}} \mid \bar{a} \neq a\}$$

We carefully curate batches for $\mathcal{L}_{\text{SSC}}$ by sampling different types of negatives. We classify a negative sample $\bar{x} = (\bar{v}, \bar{q}, \bar{a}) \in \mathcal{X}^{-}(x)$ into one of three negative categories defined below.

- **Image Negatives**, $\mathcal{X}^{-}_{\texttt{img}}(x)$: Image negatives are samples that have the same image ($v = \bar{v}$) as the reference (x) but different answer. Since VQA dataset has multiple questions (~ 5.4) per image, finding image negatives is trivial.

Algorithm 2 Batch Curation Strategy for $\mathcal{L}_{\text{SSC}}$

input: number of references N_r; data $\mathcal{D}$; weights $\boldsymbol{w}$
function CURATE(N_r, $\mathcal{D}$, $\boldsymbol{w}$)
  $\mathcal{B} = \phi, \mathcal{B}_r = \phi$ # initialize batches
  for all $i \in \{1, \ldots, N_r\}$ **do**
    $x_i \sim \mathcal{D}$ # reference
    $\hat{x_i} \sim \mathcal{X}^{+}_{\texttt{cls}}(x_i)$ # intra-class positive
    $t \sim \text{Cat}(\mathcal{T}|\boldsymbol{w})$ # negative type
    $\bar{x}_i \sim \mathcal{X}^{-}_{t}(x_i)$ # negative
    append $\mathcal{B} = \mathcal{B} \cup \{x_i, \hat{x_i}, \bar{x}_i\}$
  for all $i \in \{1, \ldots, |\mathcal{B}|\}$ **do**
    $x'_i \sim \mathcal{X}^{+}_{\texttt{para}}(x_i)$ # paraphrased positive
    append $\mathcal{B}_r = \mathcal{B}_r \cup \{x'_i\}$
return $\mathcal{B} \cup \mathcal{B}_r$

- **Question Negatives**, $\mathcal{X}^{-}_{\texttt{que}}(x)$: Question negatives are samples that have questions similar to the reference but different answer. We measure the similarity between the questions by computing their cosine distance in the vector space of the Sentence-BERT [41] model, i.e. $sim(q, \bar{q}) > \epsilon$, where ϵ is a similarity threshold.

- **Random Negatives**, $\mathcal{X}^{-}_{\texttt{rand}}(x)$: Random negatives are samples that do not fall under either Image or Question negative categories *i.e.* any image and question pair that has a different answer than the reference.

We hypothesize that discriminating between joint V+L representations of above negatives and the reference would lead to more robust V+L representations as it requires the model to preserve relevant information from both modalities in the learnt representation. Negative samples belonging to each of the above types are depicted in Figure 2(b).

Batch Curation. To create mini-batches for $\mathcal{L}_{\text{SSC}}$, as described in Algorithm 2, we start by filling our batch with triplets of reference x_i, a intra-class positive $\hat{x_i}$ and a negative sample $\bar{x_i}$ of type t. The negative type t is sampled from a categorical distribution $\text{Cat}(\mathcal{T}|\boldsymbol{w})$ where $\boldsymbol{w} = (w_{\texttt{img}}, w_{\texttt{que}}, w_{\texttt{rand}})$ are the probability weights of selecting different types of negatives defined by $\mathcal{T} = \{\texttt{que}, \texttt{img}, \texttt{rand}\}$. This procedure is repeated for specified number of times N_r to create a batch $\mathcal{B}$. Finally, for every sample in $\mathcal{B}$ we add a corresponding paraphrased positive x'_i sample. For $\mathcal{L}_{\text{CE}}$, we sample mini-batches randomly from the dataset $\mathcal{D}^{\texttt{aug}}$.

Importance of Scaling Factor. VQA Dataset has a skewed distribution of answer labels and since we sample references for SCL minibatch independently of each other (see Algorithm 2) quite often we end up with multiple intra-class positives but only a single paraphrased positive for given a reference in a minibatch. To balance this trade-off we choose to scale the loss corresponding to paraphrased pos-

itive sample from the intra-class positive samples. We call this loss Scaled Supervised Contrastive Loss ($\mathcal{L}_{\text{SSC}}$).

5. Experiments

5.1. Datasets and Metrics

We use the VQA v2.0 [18] and the VQA-Rephrasings [44] datasets for experiments. VQA contains nearly 443K train, 214K val and 453K test instances. VQA-Rephrasings was collected to evaluate the robustness of VQA models towards human rephrased questions. Specifically, the authors collected 3 human-provided rephrasings for 40k image-question pairs from the VQA v2.0 validation dataset.

Shah *et al*. [44] also introduced Consensus Score (CS) as an evaluation metric to quantify the agreement of VQA models across multiple rephrasings of the same question. Amongst all subsets of paraphrased questions of size k, the consensus score **CS(k)** measures the fraction of subsets in which *all* the answers have non-zero VQA-Score. For a set of paraphrases Q, the consensus score **CS(k)** is defined as:

$$\textbf{CS(k)} \quad = \sum_{Q' \subset Q, |Q'|=k} \frac{\mathcal{S}(Q')}{{}^{n}C_{k}} \tag{10}$$

$$\mathcal{S}(Q') = \begin{cases} 1 & \text{if } \forall q \in Q', \ \text{VQA-Score}(q) > 0, \\ 0 & \text{else} \end{cases} \tag{11}$$

Where ${}^{n}C_{k}$ is number of subsets of size k sampled from a set of size n. **CS(k)** is zero for a group of questions Q when the model answers at least k questions correctly.

When reporting results on the val split and VQA-Rephrasings, we train on the VQA 2.0 train split and when reporting results on the VQA 2.0 test-dev and test-std we train on both VQA 2.0 train and val splits. The VQA Rephrasings dataset [44] is never used for training and used only for evaluation.

5.2. Baselines and Training Details

VQA Model. For f, we use a multimodal transformer (MMT) inspired from [13], with 6 layers and 768-dim embeddings. It takes as input two different modalities. The question tokens are encoded using a pre-trained three layer BERT [14] encoder which is fine-tuned along with the multimodal transformer. Object regions are encoded by extracting features from a frozen ResNeXT-152 [54] based Faster R-CNN model [42]. The projection module g consists of two linear layers and a L-2 normalization function. We choose MMT as representative of current SoTA models [23, 32, 13, 30, 17] in VQA that rely heavily on some form of multi-modal transformer architecture. Also note that our approach (ConClaT) is *agnostic* to the choice of the model.

Question Paraphrases using VQG. Apart from training with question paraphrases generated via Back-translation, we also experiment with generating question paraphrases using the VQG module from [44]. We input the VQG module with 88 random noise vectors to keep the generation comparable with Back-translation approach. For filtering, we use the gating mechanism used by the authors and sentence similarity score of ≥ 0.85 and keep a maximum of 3 unique rephrasings for each question.

Training Details. We train our models using Adam optimizer [28] with a linear warmup and with a learning rate of 1e-4 and a staircase learning rate schedule, where we multiply the learning rate by 0.2 at 10.6K and at 15K iterations. We train for 5 epochs of train + augmented dataset on 4 NVIDIA Titan XP GPUs and use a batch-size of 420 when using $\mathcal{L}_{\text{SSC}}$ and $\mathcal{L}_{\text{CE}}$ both and 210 otherwise. We put remaining hyperparameters in the supplementary.

Existing state-of-the-art methods. Previous work [44] in VQA-Rephrasings trained a VQG model using a cycle-consistent training scheme along with the VQA model. The approach involved generating questions by a VQG model such that the answer for the original and the generated question are consistent with each other. For their experiments, they build on top of Pythia [24] and BAN [6] as base VQA models. We treat these approaches as baselines.

6. Results

In this section, we carefully ablate each component of ConClaT, and also compare results with previous methods (Pythia+CC, BAN+CC) from [44].

6.1. ConClaT

Our baseline architecture MMT without any additional data (Table 2, Row 5) and trained using cross-entropy ($\mathcal{L}_{\text{CE}}$) outperforms previous best (BAN+CC, Table 2, Row 4) by +3.64% on **CS(4)** while being -0.31% worse on VQA 2.0 validation. Training MMT with Back-translated data (Table 2, Row 8) using only $\mathcal{L}_{\text{CE}}$ further improves **CS(4)** by +0.54% while slightly degrading performance on VQA 2.0 by -0.15%, we treat this as our new baseline.

We find that alternate training (ConClaT) (Table 2, Row 9) improves **CS(4)** by +1.63% and VQA Accuracy by +0.67 % on validation. ConClaT outperforms previous SoTA approach BAN+CC by +5.81% on **CS(4)** while being competitive on VQA 2.0 validation (+0.22%) and test-dev (-0.07%) splits. We present this as our main result, which shows that training with both the losses together leads to models that are accurate (higher VQA score) and robust (higher Consensus score).

We ablate the model architecture, and test ConClaT with

	Model	Loss(es)	Scaling	N-Type	Train Scheme	CS(3)	CS(4)	VQA val
1	MMT	$\mathcal{L}_{\text{CE}}$	-	-	-	55.53	52.36	66.31
2	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	✓	R	Alternate	56.53	53.42	66.62
3	MMT	$\mathcal{L}_{\text{SC}}$ & $\mathcal{L}_{\text{CE}}$	✓	RQ	Alternate	56.88	53.77	66.97
4	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	✓	RI	Alternate	56.91	53.79	66.93
5	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	✓	QI	Alternate	57.00	53.90	66.95
6	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	✓	RQI	Alternate	**57.08**	**53.99**	**66.98**
7	MMT	$\mathcal{L}_{\text{SC}}$ & $\mathcal{L}_{\text{CE}}$	✗	RQI	Alternate	56.49	53.36	66.60
8	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	Dynamic (Eq. 12)	RQI	Alternate	57.01	53.92	66.95
9	MMT	$\mathcal{L}_{\text{SSC}}$ & $\mathcal{L}_{\text{CE}}$	✓	RQI	Joint	56.59	53.63	66.23
10	MMT	$\mathcal{L}_{\text{SSC}} \rightarrow \mathcal{L}_{\text{CE}}$ [26]	✗	RQI	Pretrain-Finetune	52.63	49.20	64.21
11	MMT	$\mathcal{L}_{\text{DMT}}$ [58] & $\mathcal{L}_{\text{CE}}$	✗	RQI	Alternate	56.23	53.10	66.59

Table 1: **Ablations Study**. **Scaling** denotes whether scaling factor α (defined in Eq. 9 or Eq. 12) was used. **N-Type** defines the type of negatives used from Image (I), Question (Q) and Random (R). All experiments are run with Back-translation data.

	Model	DA	Consensus Scores		VQA Scores		
			CS(3)	CS(4)	val	test-dev	test-std
1	Pythia [24]	-	45.94	39.49	65.78	68.43	-
2	BAN [27]	-	47.45	39.87	66.04	69.64	-
3	Pythia + CC [44]	-	50.92	44.30	66.03	68.88	-
4	BAN + CC [44]	-	51.76	48.18	66.77	**69.87**	-
4	Pythia [44]	BT	51.76	48.18	66.77	**69.87**	-
5	MMT	-	55.10	51.82	66.46	-	-
6	MMT	VQG [44]	54.92	51.85	64.50	-	-
7	MMT + ConClaT	VQG [44]	55.33	52.31	64.74	-	-
8	MMT	BT	55.53	52.36	66.31	69.51	69.22
9	MMT + ConClaT	BT	**57.08**	**53.99**	**66.98**	69.80	**70.00**
10	Pythia* [24]	BT	51.20	47.87	63.14	-	-
11	Pythia* [24] + ConClaT	BT	56.59	53.63	63.31	-	-

Table 2: ConClaT vs existing methods / baselines on VQA-Rephrasings and VQA 2.0. **DA** denotes the source of augmented data from either Back Translation (BT) or Visual Question Generation (VQG). For test-dev and test-std, we train our model on train+val set of VQA 2.0. Pythia* denotes an in-house implementation without the bells and whistles used in original work.

an in-house implementation of Pythia [24] (Table 2, Rows 10, 11). We find that using ConClaT improves CS(4) by 1.42% and VQA accuracy by .17%,

ConClaT with VQG data. We also experiment by augmenting the data generated from VQG model of [44]. Similar to Back-translation data, we find that using ConClaT (Table 2, Row 7) leads to +0.46% and +0.24% gains on **CS(4)** and VQA 2.0 validation over the baseline (Table 2, Row 8). We attribute the relatively smaller gains from VQG data to the lower quality and lesser quantity of paraphrases generated by the VQG module. We discuss more about the quality of generated data in Supplementary Section 5.

6.2. Ablations

Training schemes. We try three different ways of combining $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{SSC}}$ losses. Training alternately performs the best (Table 1, Row 6), whereas training jointly performs worse by -0.36% and -0.75% on **CS(4)** and VQA validation accuracy respectively (Table 1, Row 9). Following the approach taken in [26], we try pre-training the model with $\mathcal{L}_{\text{SSC}}$ and then finetuning it on $\mathcal{L}_{\text{CE}}$ (Table 1, Row 10) and we find this to perform the worst with -4.79% and -2.77% in **CS(4)** and VQA validation accuracy respectively .

Contrastive vs Triplet Losses. Previous works have explored the use of triplet losses [58, 38] for learning robust VQA models. Specifically, we experiment by replacing our $\mathcal{L}_{\text{SSC}}$ with Dynamic-margin Triplet loss ($\mathcal{L}_{\text{DMT}}$) proposed in [58] for mitigating the tendency of VQA models to ignore the image and rely solely on question for answering (also known as knowledge-inertia) . It is also worth noting that $\mathcal{L}_{\text{DMT}}$ is an improved version of the vanilla triplet loss used in [38]. We find that ConClaT outperforms this ablation (Table 1, Row 11) by +0.89% and +0.39% **CS(4)** and VQA validation accuracy respectively.

Scaling in $\mathcal{L}_{\text{SSC}}$. We see improvement on both VQA validation (+0.56%) and **CS(4)** (+0.35%) when using our proposed variant Scaled Supervised Contrastive Loss ($\mathcal{L}_{\text{SSC}}$) when compared to using unscaled $\mathcal{L}_{\text{SC}}$ (Table 1, Rows 6, 7). Beyond the constant scaling factor defined in Eq. 9, we also experimented with using a dynamic scaling factor defined as follows:

$$\alpha_{ip} = \begin{cases} s.\Phi(\boldsymbol{z}_i.\boldsymbol{z}_p) & \text{if } x_p \in \mathcal{X}^{+}_{\texttt{para}}(x_i), \\ \Phi(\boldsymbol{z}_i.\boldsymbol{z}_p), & \text{otherwise} \end{cases} \tag{12}$$

Where $\Phi(\boldsymbol{u}, \boldsymbol{v}) = 1 - \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$ computes the cosine distance between $\boldsymbol{u}$ and $\boldsymbol{v}$. We did not find significant improvements using dynamic scaling (Table 1, Row 8).

Negative Sampling Strategy. Furthermore, we find that our proposed negative sampling strategy (Algorithm 2) (Table 1, Row 6) helps improve **CS(4)** (+0.57%) and VQA

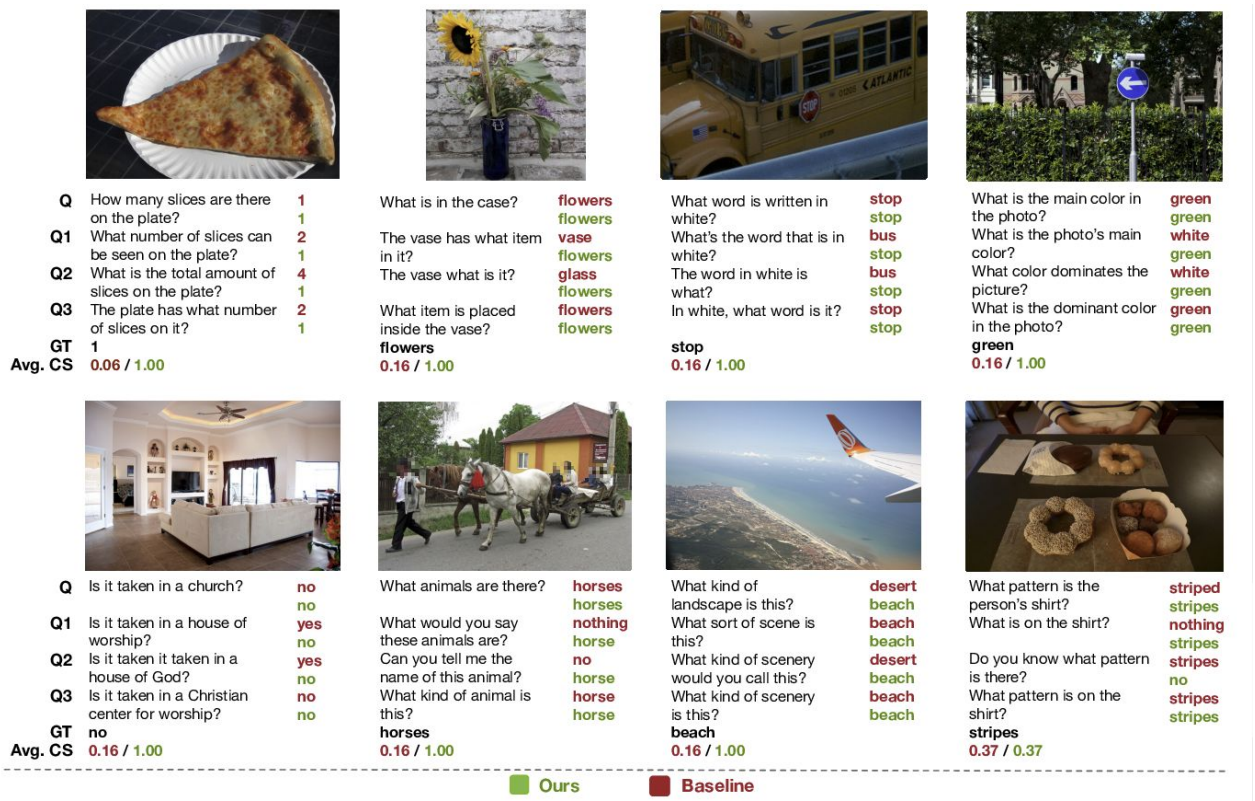

Figure 3: **Qualitative Examples.** Predictions of ConClaT vs baseline (Table 1, Rows 1 vs 6) on several image-question pairs and their corresponding rephrased questions. Average Consensus Scores (k=1-4) are at the bottom (higher the better).

accuracy (+0.36%) over random-sampling (Table 1, Row 2). We find that adding either `que`-type negatives (Table 1, Row 3) or `img`-type negatives(Table 1, Row 4) lead to gains in **CS(4)** and VQA validation accuracy.

6.3. Qualitative Analysis

We qualitatively visualize few samples in Figure 3. We compare ConClaT with our baseline (Table 1, Rows 6 vs 1). ConClaT improves the consistency in answers across the rephrasings. (2,2) shows an interesting example where ConClaT yields a singular answer for one question paraphrase and produces the original plural answer for other paraphrased question. In (2,3), baseline incorrectly answers the original question but correctly answers some of the rephrasings whereas ConClaT gets all the questions right. (2,4) illustrates a failure case where both the approaches fail to answer all the paraphrased questions correctly.

7. Conclusion

To summarize, we have three main contributions. First, we propose a novel training paradigm (ConClaT) that optimizes contrastive and cross-entropy losses to learn joint vision and language representations that are robust to question paraphrases. Minimizing the contrastive loss encourages representations to be robust to linguistic variations in questions while the cross-entropy loss preserves the discriminative power of the representations for answer classification. Second, we introduce Scaled Supervised Contrastive Loss ($\mathcal{L}_{\text{SSC}}$), that assigns higher weight to positive samples associated with question paraphrases over samples that just have the same answer boosting the performance further. Finally, we propose a negative sampling strategy to curate loss-specific batches which improves performance over random sampling strategy. Compared to previous approaches, VQA models trained with ConClaT achieve higher consistency scores on the VQA-Rephrasings dataset as well as higher VQA accuracy on the VQA 2.0 dataset across a variety of data augmentation strategies. We also qualitatively demonstrate that our approach yields correct and consistent answers for VQA questions and their rephrasings.

8. Acknowledgements

The Georgia Tech effort was supported in part by NSF, AFRL, DARPA, ONR YIPs, ARO PECASE, Amazon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government, or any sponsor.

References

[1] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, and A. van den Hengel. Counterfactual vision and language learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10041–10051, 2020. 2

[2] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing, 2020. 2

[3] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering, 2017. 2

[4] Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset, 2017. 2

[5] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2015. 3

[6] 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, 2017. 2, 3, 6

[7] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks, 2015. 2

[8] Hao-Yun Chen, Pei-Hsin Wang, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. Complement objective training, 2019. 2

[9] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering, 2020. 2

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3

[11] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 3

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[13] 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. 2, 3, 6

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 6

[15] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale, 2018. 2

[16] 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. *arXiv preprint arXiv:1606.01847*, 2016. 2

[17] 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, 2016. 6

[18] 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, 2016. 2, 3, 6

[19] Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. A deep generative framework for paraphrase generation. *arXiv preprint arXiv:1709.05074*, 2017. 2

[20] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. 2020. 3

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[22] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. 3

[23] Huaizu Jiang, I. Misra, Marcus Rohrbach, E. Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10264–10273, 2020. 2, 6

[24] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018, 2018. 1, 2, 3, 6, 7

[25] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. 3

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, A. Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *ArXiv*, abs/2004.11362, 2020. 2, 3, 4, 5, 7

[27] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. 2, 7

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. 6

[29] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020. 2

[30] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*, 2020. 6

[31] Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. Paraphrase generation with deep reinforcement learning. *arXiv preprint arXiv:1711.00279*, 2017. 2

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 2, 6

[33] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 2

[34] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering, 2016. 2

[35] Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, 2017. 2

[36] A. Oord, Y. Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 1, 3

[37] Jingjing Pan, Yash Goyal, and Stefan Lee. Question-conditioned counterfactual image generation for vqa, 2019. 2

[38] Badri Patro and Vinay P. Namboodiri. Differential attention for visual question answering, 2018. 2, 7

[39] Aaditya Prakash, Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual lstm networks. *arXiv preprint arXiv:1610.03098*, 2016. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3

[41] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. 2, 3, 5

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2015. 6

[43] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10003–10011, 2020. 2

[44] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6649–6658, 2019. 1, 2, 6, 7

[45] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[46] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[47] Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. Semantic equivalent adversarial data augmentation for visual question answering. 2020. 1, 2

[48] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision, 2020. 2

[49] Su Wang, Rahul Gupta, Nancy Chang, and Jason Baldridge. A task in a suit and a tie: paraphrase generation with semantic augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7176–7183, 2019. 2

[50] Spencer Whitehead, Hui Wu, Yi Ren Fung, Heng Ji, Rogerio Feris, and Kate Saenko. Learning from lexical perturbations for consistent visual question answering, 2020. 2

[51] John Wieting, Jonathan Mallinson, and Kevin Gimpel. Learning paraphrastic sentence embeddings from back-translated bitext. *arXiv preprint arXiv:1706.01847*, 2017. 2

[52] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019. 2, 3

[53] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*, 2018. 3

[54] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[55] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering, 2015. 2

[56] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018. 2

[57] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions, 2015. 2

[58] Yiyi Zhou, Rongrong Ji, Jinsong Su, Xiangming Li, and Xiaoshuai Sun. Free vqa models from knowledge inertia by pairwise inconformity learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9316–9323, Jul. 2019. 2, 7