
Contrast and Classify: Alternate Training for Robust VQA

Yash Kant¹, Abhinav Moudgil¹, Dhruv Batra^{1,2}, Devi Parikh^{1,2}, Harsh Agrawal¹
¹Georgia Institute of Technology, ²Facebook AI Research

Abstract

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 the augmented data, we build on the recent success in contrastive learning. We propose a novel training paradigm (ConCAT) that alternately optimizes 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 the representations for answer classification. VQA models trained with ConCAT achieve higher consensus scores on the VQA-Rephrasings dataset as well as higher VQA accuracy on the VQA 2.0 dataset.

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 a 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?” as their semantic meaning is same. While being accurate, VQA models remain brittle to question paraphrases.

To make VQA systems more robust, existing approaches [1, 2] train state-of-the-art VQA systems [3] on augmented data which includes different variations of the input question. For instance, VQA-CC [1] use a visual question generation (VQG) model to generate paraphrased question given an image and an answer. VQA models are trained by minimizing the standard cross-entropy loss on augmented data. 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 [4] 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 answers. Since we operate in a supervised setting, we choose Supervised Contrastive Loss (SCL) [5] which extends InfoNCE to utilize label information by bringing samples from the same class together. We introduce a variant of the SCL loss which emphasizes on rephrased image-question pairs (defined in Section 3.3) and alternately minimize it with cross-entropy loss 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.

We show the efficacy of our training paradigm across two data-augmentation strategies – 1) Back Translation and 2) Visual Question Generation (VQG) model from [1]. On the VQA-Rephrasings

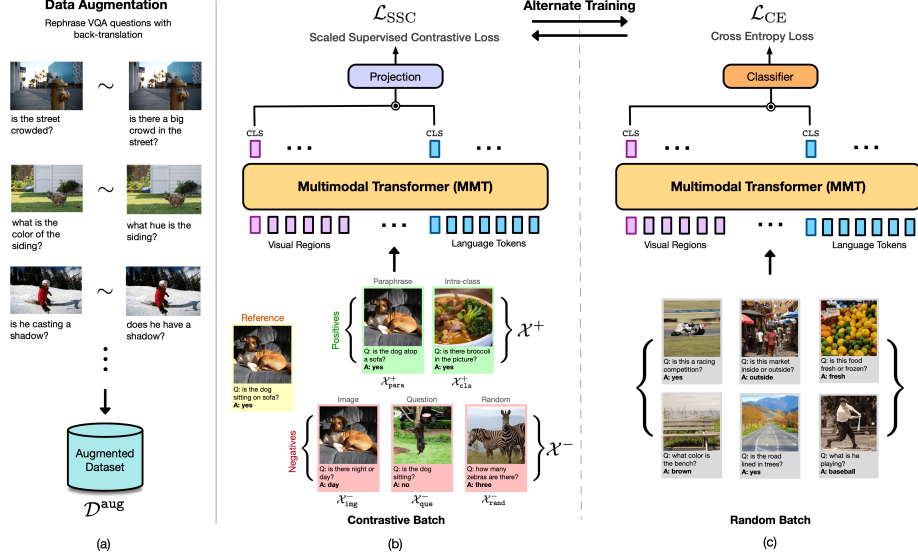


Figure 1: **Overview of ConCAT.** (a) We augment the VQA dataset with back-translated paraphrases. (b) We curate a contrastive batch by sampling different types of positives and negatives and minimize the contrastive loss \mathcal{L}_{SSC} . (c) Cross Entropy loss \mathcal{L}_{CE} is alternately optimized with \mathcal{L}_{SSC} .

benchmark which measures the model’s answer consistency across several rephrasings of a question, ConCAT improves Consensus Score [1] 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.

2 Related Work

Robust Models for VQA. Our work focuses on robustness to question paraphrases introduced in VQA-Rephrasings [1] dataset collected from human annotators. In the same work the authors trained a Visual Question Generation (VQG) model to generate question paraphrases to augment the train data. We show that this augmented data can be better utilized via ConCAT to build more robust and accurate VQA models.

Paraphrase Generation in NLP. Neural Machine Translation (NMT) models have been used to generate paraphrases in a self-supervised fashion via Back-Translation [6, 7]. Similarly, we use state-of-the-art NMT models from HuggingFace [8] to generate paraphrases for visual questions with Back Translation.

Contrastive Learning. There has been recent surge in the use of contrastive losses for learning visual representations in a self-supervised manner [9, 10, 11, 12, 13, 14]. To utilize label information, [5] proposed Supervised Contrastive Loss (SCL) for learning *visual* representations. We use a variant of the SCL loss to learn join V+L representations.

3 Preliminaries and Approach

3.1 Background

VQA Training. Visual Question Answering (VQA) [15, 16] 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 [3, 17, 18] take image and question as input and outputs a joint vision and language (V+L) representation $\mathbf{h} \in \mathcal{R}^{d_h}$ using a multi-modal network f , such that $\mathbf{h} = f(v, q)$. The V+L representation \mathbf{h} is then used to predict a probability distribution over the answer space \mathcal{A} with a softmax classifier $f^c(\mathbf{h})$ learned by minimizing the cross-entropy:

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

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

Contrastive Learning. Recent works in vision [12] 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). In this, the representation \mathbf{h} obtained from an image encoder is projected into a d_z -dimensional hyper-sphere using a projection network g such that $\mathbf{z} = g(\mathbf{h}) \in \mathcal{R}^{d_z}$.

Supervised Contrastive Loss. Given a mini-batch of size K , the image representation \mathbf{h} can be learned by minimizing the Supervised Contrastive Loss (SCL) [5] loss which operates on multiple pairs of positive and negative samples. 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:

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

Overall, $\mathcal{L}_{\text{SC}}^i$ tries to bring the representation of samples in $\mathcal{X}^+(x_i)$ closer together.

3.2 Augmented Dataset with Back Translation

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 $\mathcal{X}_{\text{para}}^+(x)$. As shown in Figure 1(a), we augment the VQA dataset \mathcal{D} with multiple paraphrased samples of a given question, denoted as \mathcal{D}^{aug} :

$$\mathcal{X}_{\text{para}}^+(x) = \{(v, q', a) \mid q' \in \mathcal{Q}(q)\} \quad \text{and} \quad \mathcal{D}^{\text{aug}} = \mathcal{D} \bigcup_{x \in \mathcal{D}} \mathcal{X}_{\text{para}}^+(x) \quad (2)$$

More details on how the paraphrased questions are generated are put in Supplementary Section A.

3.3 Scaled Supervised Contrastive Loss \mathcal{L}_{SSC} for VQA

We want to map joint vision and language (V+L) representations of the reference and paraphrased sample closer to each other. Since we operate in a supervised setting, following SCL [5] 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 $\mathcal{X}^+(x)$.

$$\mathcal{X}^+(x) = \{(\hat{v}, \hat{q}, \hat{a}) \in \mathcal{D}^{\text{aug}} \mid \hat{a} = a\} \quad \text{and} \quad \mathcal{X}_{\text{cls}}^+(x) = \mathcal{X}^+(x) - \mathcal{X}_{\text{para}}^+(x) \quad (3)$$

Note that $\mathcal{X}_{\text{para}}^+(x) \subset \mathcal{X}^+(x)$ 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 $\mathcal{X}_{\text{cls}}^+(x)$ as *intra-class* positives and set $\mathcal{X}_{\text{para}}^+(x)$ as *paraphrased* positives w.r.t. x as depicted in Figure 1(b).

Following Eq. (3.1), all the samples in $\mathcal{X}^+(x_i)$ in \mathcal{L}_{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 loss in Eq. (4) which assigns higher weight to paraphrased positives $\mathcal{X}_{\text{para}}^+(x)$ over intra-class positives $\mathcal{X}_{\text{cls}}^+(x)$. We introduce a scaling factor α_{ip} in the SCL loss (Eq. (3.1)) for a sample x_i as follows:

$$\mathcal{L}_{\text{SSC}}^i = - \sum_{p=1}^{|\mathcal{X}^+(x_i)|} \alpha_{ip} \cdot \log \frac{\exp(\Phi(\mathbf{z}_i \cdot \mathbf{z}_p)/\tau)}{\sum_{k=1}^K \mathbb{1}_{k \neq i} \cdot \exp(\Phi(\mathbf{z}_i \cdot \mathbf{z}_p)/\tau)} \quad \text{and} \quad \mathcal{L}_{\text{SSC}} = \sum_{i=1}^K \frac{\mathcal{L}_{\text{SSC}}^i}{\sum_p \alpha_{ip}} \quad (4)$$

The scaling factor α_{ip} assigns a higher weight $s > 1$ to positive samples corresponding to question paraphrases against other intra-class positives which allows \mathcal{L}_{SSC} to penalize the model strongly if it fails to bring the representations of a question and its paraphrase closer. We define α_{ip} as:

$$\alpha_{ip} = \begin{cases} s & \text{if } x_p \in \mathcal{X}_{\text{para}}^+(x_i), \\ 1, & \text{otherwise} \end{cases} \quad (5)$$

3.4 Alternate Training

Given N total training iterations, we update our model with \mathcal{L}_{SSC} after every $N_{ce} - 1$ updates with \mathcal{L}_{CE} , where N_{ce} is a hyper-parameter. Our overall loss for the n^{th} iteration is:

$$\mathcal{L}_n = \mathbb{1}_{n \pmod{N_{ce}}=0} \cdot \mathcal{L}_{\text{SSC}} + \mathbb{1}_{n \pmod{N_{ce}} \neq 0} \cdot \mathcal{L}_{\text{CE}} \quad \text{for } n \in [1, N] \quad (6)$$

Figure 1 (b),(c) parts depict our training strategy (ConCAT). We put the exact training algorithm and motivation for ConCAT in Supplementary Algorithm 1 and Section B respectively.

3.5 Negative Types and Batch Creation for \mathcal{L}_{SSC}

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

$$\mathcal{X}^-(x) = \{(\bar{v}, \bar{q}, \bar{a}) \in \mathcal{D}^{\text{aug}} \mid \bar{a} \neq a\} \quad (7)$$

We classify the negatives in three different categories – Image, Question and Random based on their similarity w.r.t. to reference sample. Negative samples of each category are depicted in Figure 1(b). Details on negatives and batch creation for \mathcal{L}_{SSC} are put in Supplementary Sections C, D.

4 Experiments and Results

Model	DA	Scaling	N-Type	Consensus Score(k)				VQA Score		
				k=1	k=2	k=3	k=4	val	test-dev	test-std
1 BAN + CC [1]	-	-	-	65.77	56.94	51.76	48.18	66.77	69.87	-
2 MMT + CE	-	-	-	67.74	59.82	55.10	51.82	66.46	-	-
3 MMT + CE	VQG	-	-	66.53	59.26	54.92	51.85	64.50	-	-
4 MMT + ConCAT	VQG	✓	RQI	66.49	59.55	55.33	52.31	64.74	-	-
5 MMT + CE	BT	-	-	67.58	60.04	55.53	52.36	66.31	69.51	69.22
6 MMT + (SCL→CE)	BT	✗	R	65.34	57.39	52.63	49.20	64.21	-	-
7 MMT + (CE + SCL)	BT	✗	R	66.95	59.70	55.32	52.20	65.10	-	-
8 MMT + ConCAT	BT	✗	R	68.35	60.97	56.49	53.30	66.73	-	-
9 MMT + ConCAT	BT	✗	RQI	68.20	60.90	56.49	53.36	66.60	-	-
10 MMT + ConCAT	BT	✓	RQI	68.62	61.42	57.08	53.99	66.98	69.80	70.00

Table 1: **Evaluation on VQA-Rephrasings and VQA 2.0 dataset.** DA denotes the source of augmented data from either Back Translation (BT) or Visual Question Generation (VQG). N-Type defines the type of negatives used from Image (I), Question (Q) and Random (R). Scaling denotes whether scaling factor α (defined in Eq. 5) was used.

We use the VQA v2.0 [16] and the VQA-Rephrasings [1] datasets for experiments. Consensus Score (CS) [1] metric was introduced to quantify the agreement of VQA models across multiple rephrasings of the same question. Definition of consensus score and experiment details is provided in Supplementary Sections F, G. We ablate each component of ConCAT, and compare it with existing approaches (BAN+CC) in Table 1. We use a multi-modal transformer (MMT) [18] for experiments.

Alternate Training. We find that alternate training (ConCAT) with contrastive loss (\mathcal{L}_{SC}) and cross-entropy (\mathcal{L}_{CE}) (Row 8) performs better on both Consensus Scores scores and VQA Accuracy compared to training with just \mathcal{L}_{CE} (Row 5). Following [5], we pretrain the model with \mathcal{L}_{SC} and then finetune the model on \mathcal{L}_{CE} (Row 6) and find that alternate training works better. Also, joint-training with both the losses ($\mathcal{L}_{SC} + \mathcal{L}_{CE}$) together (Row 7) performs worse than just using \mathcal{L}_{CE} (Row 5).

Scaled Supervised Contrastive Loss (\mathcal{L}_{SSC}) and Negative Sampling. Compared to Supervised Contrastive Loss \mathcal{L}_{SC} [5] (Row 9), we also see improvement on both VQA Accuracy and Consensus Scores when using our proposed variant Scaled Supervised Contrastive Loss \mathcal{L}_{SSC} (Rows 10). Furthermore, we find that our proposed negative sampling strategy (Algorithm 2) to curate batches for \mathcal{L}_{SSC} loss (Row 9) helps improve CS(3,4) over random negative sampling (Row 8).

ConCAT with paraphrases from VQG model. Similar to improvements seen by using ConCAT with BT data, we see improvements when using the VQG model from [1] to generate paraphrases. Although, we find the quality and diversity of paraphrases generated by the VQG module to be poor, we show that using ConCAT with VQG rephrasings (Row 4) leads to gains on both VQA and Consensus Scores over using data-augmentation with \mathcal{L}_{CE} (Row 3).

Comparison with existing methods. We compare with existing state-of-the-art approaches from [1], BAN+CC (Row 1) find that ConCAT (Row 10) outperforms it on all Consensus Scores by large margin and performs on-par on VQA scores.

5 Conclusion

We propose ConCAT which alternately optimizes Contrastive and Cross-entropy losses to learn robust and accurate VQA models.

References

- [1] 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.
- [2] Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. Semantic equivalent adversarial data augmentation for visual question answering. 2020.
- [3] 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.
- [4] A. Oord, Y. Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.
- [5] 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.
- [6] 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.
- [7] John Wieting, Jonathan Mallinson, and Kevin Gimpel. Learning paraphrastic sentence embeddings from back-translated bitext. *arXiv preprint arXiv:1706.01847*, 2017.
- [8] 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.
- [9] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*, 2018.
- [10] 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.
- [11] 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.
- [12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.
- [13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [14] 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.
- [15] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2015.
- [16] 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.
- [17] 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.

- [18] 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.
- [19] 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. URL <http://www.aclweb.org/anthology/P18-4020>.
- [20] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.
- [21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.
- [22] Adam Paszke, S. Gross, Francisco Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga, Alban Desmaison, Andreas Köpf, E. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, B. Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *ArXiv*, abs/1912.01703, 2019.
- [23] 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.
- [24] 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.
- [25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. 2015.
- [26] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, 2020.

Supplementary

A Augmented Data

Back Translation: We use 88 different MarianNMT [19] back translation model pairs released by Hugging Face [8] to generate question paraphrases. We use Sentence-BERT [20] to filter out paraphrases that cosine similarity of ≥ 0.95 with the original question and choose three unique paraphrases randomly from the filtered set. After filtering duplicates we end up with 2.89 paraphrases per original question on average.

VQG: We use the VQG model introduced by previous work [1] that takes as input the image and answer to generate a paraphrased question. 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. After filtering duplicates we end up with only 0.96 paraphrases per original question on average. We generally find the quality of VQG paraphrases worse as compared to Back-Translated ones.

B ConCAT Training

Alternately training the model with the two losses simplifies the optimization procedure compared to two-stage training (pretraining-then-finetuning as in [5]) which requires double the hyper-parameters and training iterations. We also experiment with joint-training using a linear combination of the \mathcal{L}_{SC} and \mathcal{L}_{CE} losses. The key drawback of this paradigm is that it does not allow for loss-specific curation of batches. Due to this, \mathcal{L}_{CE} is forced to operate with batches created for \mathcal{L}_{SC} (built with positives and negatives). Empirically, we show that alternate training works better than both joint-training as well as pretraining-then-finetuning approach.

Algorithm 1 ConCAT with \mathcal{L}_{SSC} and \mathcal{L}_{CE}

```
input: steps  $N$ ; constant  $N_{ce}$ ; data  $\mathcal{D}^{aug}$ ; networks  $f, g$ 
for all  $i \in \{1, \dots, N\}$  do
     $\mathcal{B} = \phi$ 
    if  $i \pmod{N_{ce}} = 0$  do
        # sscl iteration
         $\mathcal{B} = \text{CURATE}(N_r, \mathcal{D}^{aug}, w); \mathcal{L} = \mathcal{L}_{SSC}$  # see CURATE in Algorithm 2
    else do
        # ce iteration
         $\mathcal{B} \sim \mathcal{D}^{aug}; \mathcal{L} = \mathcal{L}_{CE}$ 
        update  $f(\cdot), g(\cdot)$  networks to minimize  $\mathcal{L}$  over  $\mathcal{B}$ 
return network  $f(\cdot)$ ; throw away  $g(\cdot)$ 
```

C Negative Types

We carefully curate batches for \mathcal{L}_{SSC} by sampling different types of negatives, we classify the negatives in three different categories – 1.) Image 2.) Question 3.) Random based defined as following:

- **Image Negatives, $\mathcal{X}_{img}^-(x)$:** Image negatives are samples that have the same image as the reference (x) but different answer. Since VQA dataset has multiple questions (~ 5.4) per image, finding image negatives is trivial.
- **Question Negatives, $\mathcal{X}_{que}^-(x)$:** Question negatives are samples with similar questions and different answers. We measure the similarity between the questions by computing their cosine distance in the vector space of the Sentence-BERT [20] model.

- **Random Negatives**, $\mathcal{X}_{\text{rand}}^-(x)$: 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 1(b).

D Batch Curation for \mathcal{L}_{SSC}

To create mini-batches for \mathcal{L}_{SSC} , as described in Algorithm 2, we start by filling our batch with triplets of reference x_i , an 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 $\text{Cat}(\mathcal{T}|\mathbf{w})$ where $\mathbf{w} = (w_{\text{img}}, w_{\text{que}}, w_{\text{rand}})$ are the probability weights of selecting different types. 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. When sampling all the three types of negatives we use $\mathbf{w} = (w_{\text{img}}, w_{\text{que}}, w_{\text{rand}}) = (0.25, 0.25, 0.5)$. For \mathcal{L}_{CE} , we randomly sample mini-batch from the dataset \mathcal{D}^{aug} .

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

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

Now that we understand how the batches are curated for scaled supervised contrastive loss \mathcal{L}_{SSC} , we can understand the role of scaling factor better which is explained in the following Section E.

E 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 positive sample, we call this loss Scaled Supervised Contrastive Loss (\mathcal{L}_{SSC}).

F Consensus Score and VQA-Rephrasings

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.

VQA-Rephrasings [1] 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 $\text{CS}(\mathbf{k})$ measures the fraction of subsets in which *all* the answers have non-zero VQA-Score. For a set of paraphrases Q , the consensus score $\text{CS}(\mathbf{k})$ is

defined as:

$$\mathbf{CS}(\mathbf{k}) = \sum_{Q' \subset Q, |Q'|=k} \frac{\mathcal{S}(Q')}{\binom{n}{k}} \quad (8)$$

$$\mathcal{S}(Q') = \begin{cases} 1 & \text{if } \forall q \in Q', \text{ VQA-Score}(q) > 0, \\ 0 & \text{else} \end{cases} \quad (9)$$

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

G Experiment Details

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 [1] is never used for training and used only for evaluation.

Training Details. All the models have $\sim 100\text{M}$ trainable parameters. We train our models using Adam optimizer [21] with a linear warmup and with a learning rate of $1\text{e-}4$ 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 augmented train dataset \mathcal{D}^{aug} on 4 NVIDIA Titan XP GPUs and use a batch-size of 420 when using \mathcal{L}_{SSC} and \mathcal{L}_{CE} both and 210 otherwise. We use the PyTorch [22] for all the experiments. The hyperparameters are summarized in Table B. We train our models using Adam optimizer [21] with a linear warmup and with a learning rate of $1\text{e-}4$ 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}_{SSC} and \mathcal{L}_{CE} both and 210 otherwise. We use the PyTorch [22] for all the experiments. We set number of references $N_r = 70$, the scaling factor $s = 20$, the similarity threshold $\epsilon = 0.95$ and $N_{ce} = 4$.

Table B: Hyperparameter choices for models.

#	Hyperparameters	Value	#	Hyperparameters	Value
1	Maximum question tokens	23	2	Maximum object tokens	101
3	$\mathcal{L}_{CE}:\mathcal{L}_{SSC}$ iterations ratio	3:1	4	Number of TextBert layers	3
5	Embedding size	768	6	Number of Multimodal layers	6
7	Multimodal layer intermediate size	3072	8	Number of attention heads	12
9	Negative type weights (\mathbf{w})	(0.25, 0.25, 0.5)	10	Multimodal layer dropout	0.1
11	Similarity Threshold (s)	0.95	12	Optimizer	Adam
13	Batch size	210/420	14	Base Learning rate	$2\text{e-}4$
15	Warm-up learning rate factor	0.1	16	Warm-up iterations	4266
17	Vocabulary size	3129	18	Gradient clipping (L-2 Norm)	0.25
19	Number of epochs	5/20	20	Learning rate decay	0.2
21	Learning rate decay steps	10665, 14931	22	Number of iterations	25000
23	Projection Dimension (\mathcal{R}^{dz})	128	24	Hidden Dimension (\mathcal{R}^{dh})	3129

VQA Model. For f , we use a multimodal transformer (MMT) inspired from [18], 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 [23] encoder which is fine-tuned along with the multimodal transformer. Object regions are encoded by extracting features from a frozen ResNeXT-152 [24] based Faster R-CNN model [25]. The projection module g consists of two linear layers and a L-2 normalization function.

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 [1]. 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.

Existing state-of-the-art methods. Previous work [1] 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 [3] and BAN [17] as base VQA models. We treat these approaches as baselines for our experiments.

Evaluation: During training we evaluate our models using the back-translated rephrasings on a subset of questions from validation set which do not overlap with VQA-Rephrasings [1] dataset.

H Ablations with Joint Training

In the joint training experiment (Table 1, Row 10) we use a weighing parameter (β) to combine the \mathcal{L}_{SC} and \mathcal{L}_{CE} losses. We ablate on the choice of weight (β) used, and we represent the overall loss in this experiment as:

$$\mathcal{L}_{joint} = \beta \mathcal{L}_{SSC} + (1 - \beta) \mathcal{L}_{CE}$$

We find that increasing β leads to decreasing gains on Consensus Scores as shown in Table A. We also find that the VQA-Accuracy hits a sweet-spot at $\beta = 0.5$ and we use this configuration as our baseline.

	Model	β	CS(4)	VQA Score val
1	MMT + CE	0.25	52.97	66.14
2	MMT + CE	0.50	52.36	66.31
3	MMT + CE	0.75	48.53	61.34
4	MMT + CE	0.90	40.68	51.03
5	MMT + ConCAT	-	53.99	66.98

Table A: Ablations on the choice of our hyper-parameter β for joint training.

I Gradient Surgery of \mathcal{L}_{SSC} and \mathcal{L}_{CE}

To know whether the gradients of both the losses (\mathcal{L}_{SSC} and \mathcal{L}_{CE}) are aligned with each other during training, we follow the gradient surgery setup of [26] for multi-task learning. During joint-training, we take the dot-products of gradients from both the losses and plot them to see how well they are aligned *i.e.* whether the dot product is positive or negative. In Figure A we plot the un-normalized dot product between the gradients corresponding to \mathcal{L}_{CE} and \mathcal{L}_{SSC} losses. We find that except for initial few steps the gradients of both the losses are aligned (dot product is positive) and thus the updates are complementary with respect to each other.

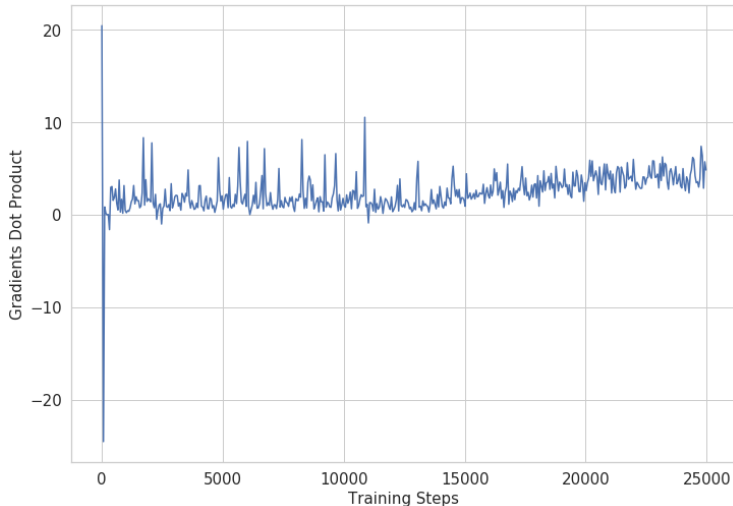


Figure A: Gradient Alignment between the \mathcal{L}_{SSC} and \mathcal{L}_{CE} losses. The dot-product is positive indicating that the gradients from the two losses are aligned.



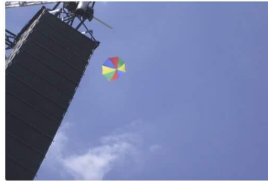






					
Q Who is sniffing who?	dog	What color is the clock?	orange	What color is the umbrella?	blue
Q1 Somebody is being sniffed by who?	dog	Can you tell me what color the clock is?	orange	The color of the umbrella is what?	rainbow
Q2 Who is doing the sniffing?	yes	What color is the pictured clock?	yes	The umbrella's color is what?	rainbow
Q3 What is doing the sniffing?	dog	What is the color of the clock?	orange	The umbrella is what color?	blue
GT dog	dog	orange	orange	rainbow	rainbow
Avg. CS 0.16 / 1.00	0.37 / 1.00	0.16 / 1.00	0.16 / 1.00	0.16 / 1.00	0.16 / 1.00
					
Q How many skis are on the ground?	0	Which corner of the table is in the frame?	top	Is the cat able to access the toilet water?	no
Q1 What is the number of skis on the ground?	0	What corner of the table is shown?	left	Is toilet water something the cat has access to?	yes
Q2 What is the ski count on the ground?	0	Which corner of this table is visible?	left	Does the cat have access to the toilet water?	yes
Q3 What's the exact number of skis on the ground?	1	Which corner of the table is in the shot?	top right	Is accessing the toilet water something the cat can do?	no
GT 0	0	left	left	no	no
Avg. CS 0.16 / 1.00	0.00 / 1.00	0.00 / 1.00	0.16 / 1.00	0.16 / 1.00	0.16 / 1.00
					
Q Are these fruit or vegetable?	food	Is this food sweet?	yes	How many sheep are there?	2
Q1 Are these classified as fruit or vegetable?	vegetable	Does this food have a sweet taste?	yes	What is the total of sheep?	sheep
Q2 Would these be called fruit or vegetable?	carrots	Does this food taste sweet?	yes	What is the number of sheep?	2
Q3 Do these count as fruit or vegetable?	vegetables	Is this a food that tastes sweet?	yes	How many sheep?	2
GT vegetable	vegetable	yes	no	2	2
Avg. CS 0.00 / 1.00	0.16 / 1.00	0.37 / 1.00	0.16 / 1.00	0.37 / 1.00	0.16 / 1.00
<div><div>Ours</div><div>Baseline</div></div>					

Figure B: Qualitative Examples. Predictions of ConCAT and MMT+CE baseline on several image-question pairs and their corresponding rephrased questions. Average Consensus Scores (k=1-4) are also shown at the bottom (higher the better).

J Qualitative Samples

Figures B, C, D show many more qualitative samples comparing the baseline and ConCAT.



Figure C: Qualitative Examples. Predictions of ConCAT and MMT+CE baseline on several image-question pairs and their corresponding rephrased questions. Average Consensus Scores (k=1-4) are also shown at the bottom (higher the better).

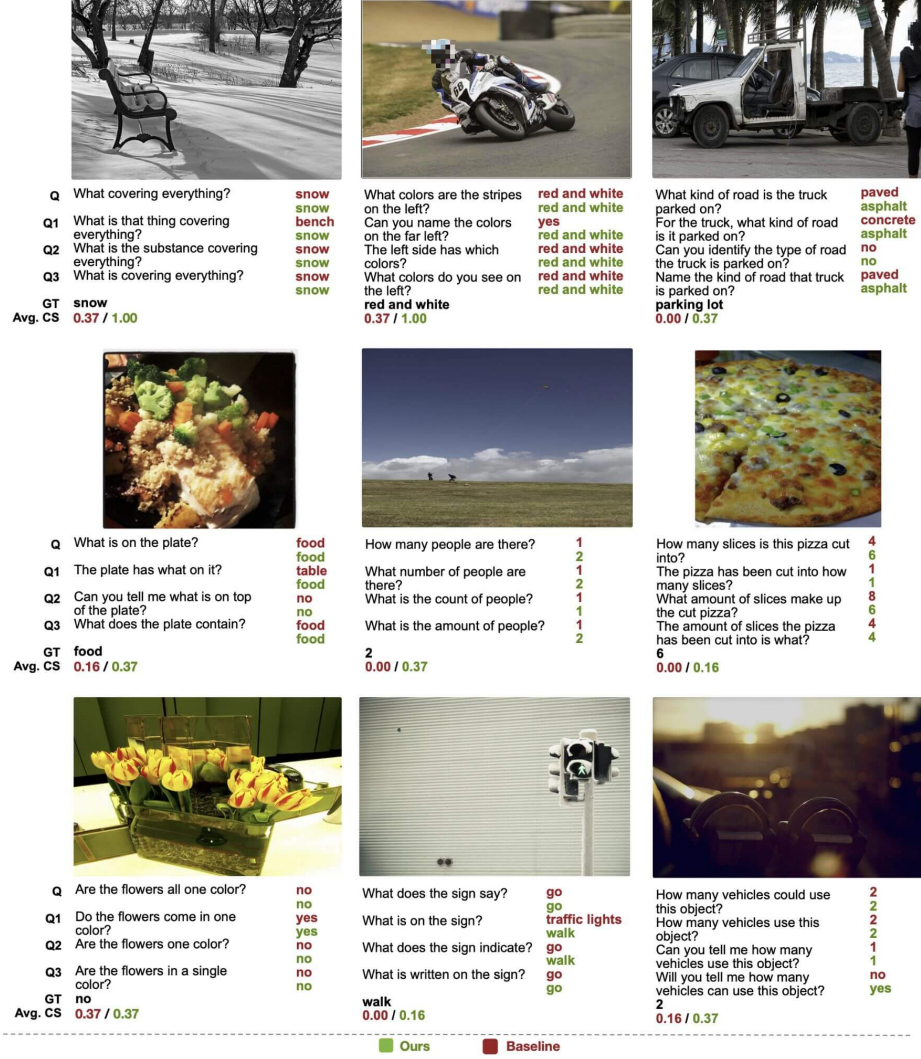


Figure D: Qualitative Examples. Predictions of ConCAT and MMT+CE baseline on several image-question pairs and their corresponding rephrased questions. Average Consensus Scores (k=1-4) are also shown at the bottom (higher the better).