Crisscrossed Captions: Extended Intramodal and Intermodal Semantic Similarity Judgments for MS-COCO

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang
Google Research
{zarana, jasonbaldridge, cer, austinwaters, yinfeiy}@google.com

Abstract

Image captioning datasets have proven useful for multimodal representation learning, and a common evaluation paradigm based on multimodal retrieval has emerged. Unfortunately, datasets have only limited cross-modal associations: images are not paired with others, captions are only paired with others that describe the same image, there are no negative associations and there are missing positive cross-modal associations. This undermines retrieval evaluation and limits research into how inter-modality learning impacts intra-modality tasks. To address this gap, we create the Crisscrossed Captions (CxC) dataset, extending MS-COCO with new semantic similarity judgments for 247,315 intra- and inter-modality pairs. We provide baseline model performance results for both retrieval and correlations with human rankings, emphasizing both intra- and inter-modality learning.

1 Introduction

The meaning of words and expressions such as blue, chair, hot dog, and garden path have strong visual components, yet the standard approach to creating computational representations of such terms is usually done using only text corpora. Prior work such as Picturebook (Kiros et al., 2018) derives representations of words using images connected to them and shows improvements for tasks such as word similarity ranking and image-text retrieval. On the flip side, the query-based training of the Graph-RISE image embedding models (Juan et al., 2019) is a powerful demonstration of the power of words to produce stronger image representations.

We hypothesize that learned representations in multimodal contexts (Baltruaitis et al., 2019) can improve retrieval and discrimination within as well as across modalities. However, there are no datasets ideally suited for this at present. Image captioning data sets such as Flickr8k (Rashtchian et al., 2010), Flickr30k (Young et al., 2014), its multilingual extension Multi30k (Elliott et al., 2016), MSCOCO (Lin et al., 2014), and Conceptual Captions (Sharma et al., 2018) are incomplete because the only known relationships are between images and textual captions created for them. This misses many valid relationships between unassociated images and captions and those from captions to other captions and from images to other images.

To address this gap, we create Crisscrossed Captions (CxC, exemplified in Figure 1), a dataset with denser and more fine-grained annotations for relationships between and among captions and images in the MS-COCO evaluation splits of Karpathy and Fei-Fei (2015) (with 25,000 captions and 5,000 images each). CxC extends the existing image-caption pairs with continuous (0-5) semantic similarity ratings for those pairs and new pairs. Ratings
criteria are extensions of those used for Semantic Textual Similarity (Agirre et al., 2012). New intramodal pairs are sampled for annotation via an indirect sampling scheme that computes similarity for captions using corresponding images, and vice versa. In total, CxC contains ratings for $247k$ pairs, a massive extension in scale and detail to the 50,000 binary pairings in the original data.

CxC validates the strong semantic alignment between image and their captions, with an average similarity of 4.85. However, we also find that co-caption pairs of the same image have an average of just 3.0. This calls into question the use of such pairs in paraphrase generation (Gupta et al., 2018) and reinforces the need for images as context for human evaluation in paraphrasing (Wang et al., 2019).

We evaluate intra- and inter-modality models to demonstrate CxC’s utility for characterizing different modeling choices—including the mere fact of using both modalities rather than just one. CxC allows us to measure Spearman’s correlations of model and human rankings, which we compute via a bootstrap measure over diverse sub-samples of the paired ratings. CxC also allows us to extend the standard recall@k measures common in cross-modal retrieval experiments with new positive pairs and thereby address some of the gap identified by Ilharco et al. (2019) due to missing valid associations in cross-modal retrieval. Furthermore, CxC supports recall@k for intramodal retrieval. The fact that these evaluations are all possible on a common set of images and captions makes them far more valuable for understanding intermodal learning, compared to disjoint sets for caption-image, caption-caption, and image-image associations.

2 Dataset Collection

We seek graded similarity associations within and across modalities. We need this primarily for evaluation as there is plentiful material for training across and with modalities, so we extend the MS-COCO evaluation splits rather than constructing a new dataset. MS-COCO has five captions for each image, split into 410k training, 25k development, and 25k test captions (82k/5k/5k for images) for the splits defined in Karpathy and Fei-Fei (2015). An ideal extension would include ratings for every possible pair. However, this is infeasible and most pairs are dissimilar anyway. For retrieval evaluations in particular, we need new pairs with reasonably high similarity, so we introduce a biased sampling scheme to select pairs for rating.

The data is collected in two phases. First, we define an indirect sampling scheme that uses model-based similarities from the co-modality items to select intramodality pairs. We use these items and their human ratings to select intermodality pairs for annotation. We also annotate all existing intermodal pairs and a large sample of co-captions (captions associated with the same image).

2.1 Intramodality Selection and Annotation

Two images of a man and a dog can be described very differently, and two similar sentences about a man and a dog can describe dissimilar images. For example, in Figure 2, caption 1 focuses on a visual description while caption 2 gives a description of the event depicted. Divergences also occur because the creators of two captions perceived the scene differently: caption 3 describes the room and caption 4 focuses on the dog seated on the sofa. This semantic gap between images and their caption creates an opportunity for sampling intramodal pairs with varying similarities. The key idea is to use model-based similarities for images for biased sampling of caption pairs, and vice versa—using the paired relationships as to pivot between the modalities. This selects image pairs that are different in appearance but similar in what they depict based on the corresponding descriptions, and vice versa.
Let the known images and their captions be denoted \( V (v_1...v_n) \) and \( C (c_1...c_m) \), with latter representing co-caption groups of five captions each. Each item is encoded using an off-the-shelf intramodality model and cosine similarity between items is computed to build two symmetric matrices: \( S^C \) (pairwise caption similarities) and \( S^V \) (pairwise image similarities). The diagonals of both are set to zero to avoid sampling identical item pairs.

We encode images using the Graph-RISE model (Juan et al., 2019). Computing cosine similarity between these representations then provides the image-based similarity for each bag of co-captions. We encode captions with Universal Sentence Encoder (USE) (Cer et al., 2018) and average bag of words (BoW) based on GloVe embeddings (Pennington et al., 2014). The five co-captions for each image are grouped by averaging them to create a single representation. We compute cosine similarity on these vectors to obtain two variants of textual similarity for every possible image pair. Even though there are two caption-based similarity matrices from each encoding method, we refer to them as one for describing the selection process below.

We use \( S^C \) to select image pairs and \( S^V \) for caption pairs. Because of the cross-modal semantic gap and diversity and size of the underlying data, these pairs exhibit a wide range of similarity. Selecting the five most similar items (according to model-based \( S^V \) and \( S^C \)) thus produces good representation of varying amounts of similarity as judged by people.\(^3\) Note since the \( S^V \) matrix is only a similarity measure between co-caption sets, one caption is randomly chosen from each caption set. To further increase diversity, only one caption

---

\(^3\)Other strategies selected mostly dissimilar pairs.
Intramodal Sampling refers examples selected using other-modality selection, Same Example refers to examples originating from an original MS-COCO pair, All refers to all the examples for the task.

Table 3: Number of annotations per task (STS, SIS, SITS) and per split (val, test).

| Split      | Task | STS  | SIS  | SITS | Total (per split) |
|------------|------|------|------|------|-------------------|
| Validation |      | 43,898 | 42,767 | 34,924 | 121,589         |
| Test       |      | 44,044 | 46,719 | 34,962 | 125,725         |
| Total (per task) | | 87,942 | 89,486 | 69,886 | 247,314         |

2.2 Intermodality Selection and Annotation

We select intermodal candidates $C^2 V$ based on the ratings assigned to pairs in $V^2 V$ and $C^2 C$. In this case, we primarily seek positive associations to identify really challenging cross-modal examples as well as likely positive associations that were not explicitly indicated in the dataset (such as those identified by the annotators in Ilharco et al. (2019)).

We rank all items in $C^2 C$ and $V^2 V$ based on their average STS and SIS scores, respectively; then half of the $C^2 V$ pairs are selected based on $C^2 C$ rankings and the other half by $V^2 V$ rankings (skipping pairs already selected from $C^2 C$). Additionally, we select all existing pairs for annotation (25k in validation and 25k in test) to obtain image-text similarity ratings for the known items.

As with SIS, we extend STS to define Semantic Image-Text Similarity (SITS, Table 2). Raters provide a continuous score from 0 to 5 using an interface similar to that described for STS and SIS. Each $C^2 V$ pair is score by five raters, and the averaged is used as the final SITS score.

3 Crisscrossed Captions Dataset

Based on the pair selection and annotation methodology given in the previous section, we obtained similarity ratings for 247,314 text-text, image-image, and image-text pairs (corresponding to 1,216,570 total individual judgments). Table 3 gives annotations counts for each task and split. Figure 3 shows the distribution of CxC ratings for the CxC validation set. Figure 4 shows the distribution of the number of positive pairs of annotations per MS-COCO example for each task. Positive pairs (text pairs for STS, image pairs for SIS and image-text pairs for SITS) are the ones with an annotation score of 3.0 or above.

ST5 For the $C^2 C$ text pairs selected based on $S^V$, the majority of the ratings fall in the negative side (with ratings in [0, 3)). This validates the observation that captions are not perfect paraphrases of each other, and the STS rating rubric captures this aspect well for MS-COCO. Nevertheless, the approach produces 2392 positive pairs.
Table 4: A comparison of CxC STS annotation scores and cosine similarity scores using GloVe BoW embeddings and Universal Sentence Encoder (USE) for five example MS-COCO caption pairs.

| Caption 1 | Caption 2 | STS  | BoW  | USE  |
|-----------|-----------|------|------|------|
| A man standing on a tennis court holding a racquet. | A man standing on a tennis court holding a tennis racquet. | 5.0 | 0.98 | 0.99 |
| A man riding a skateboard off the side of a ramp. | A man riding up the side of a skateboard ramp. | 4.2 | 0.99 | 0.94 |
| A yellow tray topped with a cup of coffee and a donut. | A white plate topped with donuts sitting on a stove top. | 3.1 | 0.94 | 0.39 |
| A bird sitting on top of a park bench. | An empty park bench sitting in front of trees. | 2.2 | 0.90 | 0.69 |
| An old car sitting on top of a lush green field. | A couple of motorcycles parked next to each other. | 1.3 | 0.85 | 0.21 |
| A man sanding next to an orange frisbee. | A couple of swans swimming in a pond next to two people. | 0.2 | 0.84 | 0.11 |

Figure 5: Plot of annotation and BoW cosine similarity scores for a sample of caption pairs data from STS.

scores and cosine similarity scores using GloVe BoW embeddings and Universal Sentence Encoder. Though there is broad agreement, the annotated semantic similarity is not fully captured by either of the similarity scores. USE provides a broader range, but scores the third pair lower than the fourth. BoW scores are bunched within a high similarity band\(^4\) that aligns well with these five examples. Overall, there is a positive but weak correlation between BoW and STS scores, as shown in Figure 5, which plots average BoW cosine similarity scores vs STS for 1000 randomly sampled caption pairs.

Figure 6 shows a pair of captions (and corresponding images) selected by the other-modality strategy with higher STS compared to their respective co-captions. For co-caption pairs, STS scores are more positive but many are still negative (Figure 3, left). Thus, combining both approaches to generate caption pairs leads to a more representative distribution overall. The number of negative pairs from co-captions underscores the problem with assuming captions of the same image are paraphrases.

**SIS** All image pairs \(V^2V\) are selected using other-modality with \(S^C\). This plus the stringent criteria for SIS rating of 5 means very few examples are rated above 4. Nevertheless, there are many pairs with SIS \(\geq 3\), indicating there are many images depicting similar scenes and events. Figure 7 gives example pairs for each SIS rating level.

**SITS** As shown in Figure 4, there are many more pairs with 4-5 SITS ratings, compared to text-text and image-image pairs. This is by design, as the inter-modality strategy selects pairs based on decreasing STS/SIS scores. This ensures we can capture as many possible intra-modality positive associations and augment the existing validation and test splits. Since these pairs are missing from the existing data, they are among the examples that inappropriately penalize a model that identifies them correctly in image-text retrieval. The SITS ratings collected for known pairs support new evaluations using Spearman’s correlation and Rank-Biased Recall, as discussed in the next section.

\(^4\)BoW scores fall mostly in the range .8 to 1.0 over all possible pairs; STS scores fall mostly in \(1 (.2)\) to \(4 (.8)\).
| Score | Image Pairs | Image Text Pairs |
|-------|-------------|-----------------|
| 5     | ![Image](image1) ![Image](image2) ![Image](image3) ![Image](image4) ![Image](image5) | A man poses with a surfboard on a beach. |
| 4     | ![Image](image1) ![Image](image2) ![Image](image3) ![Image](image4) ![Image](image5) | A couple of birds that are walking on some sand. |
| 3     | ![Image](image1) ![Image](image2) ![Image](image3) ![Image](image4) ![Image](image5) | A man is riding a surfboard at the beach. |
| 2     | ![Image](image1) ![Image](image2) ![Image](image3) ![Image](image4) ![Image](image5) | Three people stand on an empty beach watching a bird in the sky. |
| 1     | ![Image](image1) ![Image](image2) ![Image](image3) ![Image](image4) ![Image](image5) | A man in a hat rides an elephant in a river. |
| 0     | ![Image](image1) ![Image](image2) ![Image](image3) ![Image](image4) ![Image](image5) | Long road with a sign titled Jackson River Rd and East Main St. |

Figure 7: Examples for each annotation score of SIS (left) and SITS (right) tasks.
4 Tasks and Evaluation Metrics

With the collected intra- and inter-modality annotations, CxC can be used to measure model performance on many new tasks. Here, we focus on semantic similarity and retrieval.

Semantic Similarity Semantic similarity tasks measure the degree of similarity given two inputs (Cer et al., 2017). This task has been adopted to the Semantic Textual Similarity and Visual Semantic Textual Similarity (de Lacalle et al., 2020). Typically, the models are evaluated based on the Pearson correlation with the gold labels. This is valid when there is training data that can be used to calibrate model scores to the human ratings. With CxC, we do not have such training data, so we instead use Spearman’s $r$ to assess whether a model ranks pairs similarly to human raters. However, we cannot measure Spearman’s $r$ over all pairs because CxC’s dense annotation means that the scores between many pairs are themselves correlated. To mitigate this, we use a sampled bootstrap correlation instead. For each correlation estimate, we sample half of the queries (to increase diversity across samples) and for each selected query, we choose one of the items for which CxC supplies a paired rating. We compute Spearman’s $r$ between the CxC scores and the model scores for the selected pairs. The final correlation is the average over 1000 of these bootstrap samples.

Retrieval Karpathy and Fei-Fei (2015) first used MS-COCO for image-to-text and text-to-image retrieval. We extend the binary paired associations with positive CxC pairs, and also include new retrieval tasks for text-to-text and image-to-image using the same dataset, for a total of four retrieval tasks. To the best of our best knowledge, this is the first work that enables the evaluation of image-to-image retrieval with an image captioning dataset. Following Karpathy and Fei-Fei (2015), we evaluate using Recall@K (R@K), computed as “the fraction of times a correct item was found among the top K results, and median rank (med. r), which is “the closest ground truth result in the list.”

5 Evaluated Models

We provide baseline results for both correlation and retrieval evaluation. We focus on representation learning based approaches as they can be applied to both measuring semantic similarity and large-scale retrieval with fast Approximate Nearest Neighbor (ANN) search.

5.1 Dual Encoder Models

We consider several neural baseline models, all of which are dual encoders (Gillick et al., 2018; Yang et al., 2019a) that encoding left and right inputs separately, as illustrated in Figure 8.

Dual encoders are often trained using an in-batch sampled softmax loss, as this has been observed to converge quickly and perform well on retrieval tasks (Gillick et al., 2018; Yang et al., 2019a). We employ the bidirectional in-batch sampled softmax loss (eq. 1). It encourages the score of a correct pair $S(l_i, r_j)$ to be higher than scores for non-matching input pairs from the batch $S(l_i, r_j)$ where $i \neq j$:

$$
\mathcal{L} = -\frac{1}{K} \sum_{i=1}^{K} S(l_i, r_i) - \log \sum_{j=1, j \neq i}^{K} e^{S(l_i, r_j)} - \frac{1}{K} \sum_{i=1}^{K} S(r_i, l_i) - \log \sum_{j=1, j \neq i}^{K} e^{S(r_i, l_j)}
$$

where $S(i, j)$ is a simple dot product of embeddings of example $i$ and $j$. Unlike full cross-attention models, this architecture enables retrieval through approximate nearest neighbor search, and thereby scales well to large-scale retrieval problems. See Gillick et al. (2018) for more discussion on dual encoders for retrieval.

In-batch sampled softmax loss performs best when computed over a large number of negative pairs (Gillick et al., 2018). In our distributed training setup, each replica computes $l$ and $r$ for its local mini-batch and broadcasts them to all others. Training with $N$ replicas thus allows the loss to be computed over the global batch of $(R \cdot K)^2$ pairs.

We train dual encoder models for image-text and text-text tasks, as well as a multitask dual encoder model combining the image-text and text-text tasks. Details of each model are described below.

The Image-Text dual encoder employs Resnet-152 (He et al., 2016) as the image encoder on the
left and a text encoder on the right. For text, we extract token-level features from a pretrained BERT-Base model (Devlin et al., 2018), feed them into a trainable transformer layer, and use features at the 0th token position as the caption representation.

Both inputs of the Text-Text dual encoder use a single, shared text encoder with the same configuration as in the image-text dual encoder.

We also consider a Multitask dual encoder in which the model is trained on a combination of multiple dual-encoder tasks (Chidambaram et al., 2019). Its architecture is the same as the image-text dual encoder, and it is trained in the same way, except that the final loss is a weighted sum of image-text (i2t) and text-text (t2t) losses:

$$\mathcal{L} = \mathcal{L}_{i2t} + c \times \mathcal{L}_{t2t}$$  \hspace{1cm} (2)

where c is a scalar to control the weights of losses from the two tasks. Note this model only has one text encoder, shared between all retrieval tasks.

5.2 Pretrained Text / Image Representations

We also benchmark pre-trained text and image representations for STS and SIS, respectively.

Universal Sentence Encoder is a sentence level representation model trained on a variety of tasks (Cer et al., 2018). It has shown the state-of-the-art performance on the STS benchmark (Cer et al., 2017) among all representation based approaches. We use the multilingual transformer version from TensorFlow Hub (Yang et al., 2019b).5

InceptionV3 and ResNet-152 are deep convolutional image classification models (Szegedy et al., 2016; He et al., 2016) trained on the ImageNet dataset. We extract 2048-dimensional image-level representations on a central crop containing 87.5% of the original image area. The pretrained models are accessed via TensorFlow Hub.6

6 Experiments

Experimental Setup The image encoder and text encoder are both pre-trained in all experiments. ResNet-152 pretrained on ImageNet is used in the image encoder. It yields a vector representation of size $d = 2048$. We employ the BERT_Base model (Devlin et al., 2018) architecture in the text encoder, which has 12 transformer layers with 12 attention heads and a hidden dimension of 3072; it outputs features of dimension 768. The additional transformer layer has 8 attention heads, a hidden dimension of 3072, and also outputs 768-dimensional token-level features. We use the features at the 0th token position as the caption representation and project it to 2048 dimensions to match the image representation size. A word piece model with vocabulary size 28,996 is used. BERT parameters are initialized from the public BERT checkpoint7 and frozen during training. The additional transformer parameters are trainable and randomly initialized. Preliminary results indicate freezing BERT parameters has minor impact on performance, but makes the training much faster.

The Conceptual Captions dataset (Sharma et al., 2018) contains 3.3 million pairs of images and captions, much larger in scale compared to MS-COCO. Following Ilharco et al. (2019), we use this dataset during pre-training with image-to-caption and caption-to-image losses for all our dual encoder models. Pre-training uses the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) and a learning rate that starts at 1e-4 and decays by 0.1% every 1000 steps. We stop pre-training after 30k steps and select the model checkpoint that maximizes R@10 on a held-out set. We then fine-tune this checkpoint on MS-COCO using the same hyper parameters, except for a smaller learning rate of 5e-6.

The models are trained on 32-core slices of Cloud TPU V3 pods, with a per-replica batch size of 64 for pre-training and fine-tuning. As discussed in Section 5.1, the loss is computed on item pairs aggregated from all replicas; in this case, over the global batch of 2048 examples.

6.1 Semantic Similarity Results

Table 5 shows Spearman’s R bootstrapped correlation for all models on the Semantic Image-Text Similarity (SITS), Semantic Textual Similarity (STS), and Semantic Image Similarity (SIS) tasks (Cer et al., 2018). It has shown the state-of-the-art performance on the STS benchmark (Cer et al., 2018). It has shown the state-of-the-art performance on the STS benchmark (Cer et al., 2018). It has shown the state-of-the-art performance on the STS benchmark (Cer et al., 2018).

Table 5: Spearman’s R Bootstrap Correlation ($\times 100$) on MSCOCO 5k test set with CxC ratings.

| Model                  | STS  | SIS  | SITS |
|------------------------|------|------|------|
| USE                    | 71.4±0.4 | -     | -    |
| Inception V3           | 19.6±1.9 | -     | -    |
| ResNet-152             | 59.2±1.3 | -     | -    |
| DE_T2T                 | 72.9±0.4 | -     | -    |
| DE_T2T                 | 62.1±0.5 | 73.1±1.1 | 56.2±1.5 |
| MultitaskDE_T2T        | 73.3±0.4 | 68.8±1.2 | 55.7±1.6 |

5universal-sentence-encoder-multilingual-large/1
6imagenet/inception_v3/feature_vector/4 and imagenet/resnet_v1_152/feature_vector/4, respectively
7bert_en_uncased_L-12_H-768_A-12/2
Table 6: Image ↔ Text retrieval performance on MSCOCO 5k test set.

| Model            | Image → Text |          |          |          | Text → Image |          |          |          |
|------------------|--------------|----------|----------|----------|--------------|----------|----------|----------|
|                  | R@1 | R@5 | R@10 | med r | R@1 | R@5 | R@10 | med r |
| **MS-COCO Retrieval** |     |      |       |       |     |      |       |       |
| DE_{I2T}         | 40.1 | 70.6 | 80.7  | 2     | 28.1 | 57.2 | 69.9  | 4     |
| Multitask_{I2T+T2T} | 40.0 | 69.4 | 81.0  | 2     | 28.1 | 57.2 | 70.0  | 4     |
| VSE++            | 41.3 | 71.1 | 81.2  | 2     | 30.3 | 59.4 | 72.4  | 4     |
| **CxC Retrieval** |     |      |       |       |     |      |       |       |
| DE_{I2T}         | 41.9 | 73.1 | 83.6  | 2     | 30.1 | 60.5 | 73.2  | 3     |
| Multitask_{I2T+T2T} | 41.7 | 72.0 | 83.4  | 2     | 30.2 | 60.7 | 73.2  | 3     |

Table 7: Text ↔ Text and Image ↔ Image retrieval performance on MSCOCO 5k test set with CxC annotations.

Table 6: Image ↔ Text retrieval performance on MSCOCO 5k test set.

Table 7: Text ↔ Text and Image ↔ Image retrieval performance on MSCOCO 5k test set with CxC annotations.

6.2 Retrieval Results

**Intermodal Retrieval** Table 6 summarizes intermodal retrieval performance on both the original MSCOCO annotations and CxC. The dual encoder models perform competitively with the state-of-the-art VSE++ model on the original annotations (Faghri et al., 2017). New positive items added by CxC annotations show improved retrieval performance as they identify missing positives that are incorrectly penalized when using only original pairs (as noted in Ilharco et al. (2019) for Flickr8k).

**Intramodal Retrieval** Table 7 summarizes intramodal retrieval performance. USE_{mling-large} provides a strong baseline for Text → Text, while InceptionV3 and ResNet-152 are used as baselines for Image → Image. The dual encoder model DE_{T2T} achieves the strongest performance on Text → Text. DE_{I2T} provides the best performance on Image → Image. For both modalities, the best performing DE model outperforms the baseline by a sizable margin. Multitask_{I2T+T2T} also achieves much better retrieval performance than the baselines for both Text → Text and Image → Image. However, multitask training performs slightly worse than the more specialized DE models for each modality.

Figure 9 shows three examples of images retrieved for caption queries. The CxC annotations capture missing examples in the first two cases, and the last shows that there are still more positive pairs that remain unassociated in the data. Figure 10 shows the same for captions retrieved from image queries, again showing that many examples are captured in CxC that were missing before. In the case of the clock tower, all of the top retrieved items would have been counted as misses in retrieval evaluation, but are correctly counted as hits with CxC.

7 Conclusion

We introduce the Crisscrossed Captions dataset and provide baselines for both correlation and retrieval tasks. We hope to motivate further exploration in this area with joint learning of intra- and intermodality tasks and possibly extend it to a multilingual setting.

Acknowledgments

We would like to thank Julia Hockenmaier and Yuan Zhang for their feedback, and the Google Data Compute team for completing the annotations.
| Caption                                                                 | Ranked Images                                | SITS |
|------------------------------------------------------------------------|----------------------------------------------|------|
| People at the beach sitting in the sand and under umbrellas.           | ![Image 1](image1.png) ![Image 2](image2.png) ![Image 3](image3.png) | 4.58 |
| MS-COCO image rank: 2                                                   |                                              | 5.0  |
|                                                                       |                                              | 4.83 |
| A jet airliner is in the air with a cloudy grey sky.                   | ![Image 4](image4.png) ![Image 5](image5.png) ![Image 6](image6.png) | 4.87 |
| MS-COCO image rank: 2                                                   |                                              | 5.0  |
|                                                                       |                                              | N/A  |
| Numerous trains are parked at a train yard.                            | ![Image 7](image7.png) ![Image 8](image8.png) ![Image 9](image9.png) | N/A  |
| MS-COCO image rank: 3                                                   |                                              | N/A  |
|                                                                       |                                              | 4.72 |

Figure 9: Text → Image Retrieval Results from MS-COCO test set using the \( DE_{12T} \) model. The retrieved images are ranked from left to right and the respective scores (wherever available) are listed from top to bottom. The rank of the image from the MS-COCO annotations has been mentioned with each caption.

| Image                                                                 | Ranked Captions                                                                 | SITS |
|-----------------------------------------------------------------------|--------------------------------------------------------------------------------|------|
| ![Image 10](image10.png)                                               | Five motorcycles are parked on the sidewalk in front of a building.             | 5.0  |
|                                                                       | A row of motorcycles parked in front of a building.                            | 5.0  |
|                                                                       | **A number of motorcycles parked near each other near a building.**            | 4.97 |
|                                                                       | **Motorcycles parked in a row outside of metropolitan building.**             | 4.77 |
|                                                                       | **A bunch of motorcycles parked beside a building.**                          | 4.94 |
| Black and white photograph of two children holding hands.              | 4.98                                                                          |      |
|                                                                       | A black and white photograph of boys posing for a school picture.              | 4.89 |
|                                                                       | Two kids standing in front of a counter together.                             | 5.0  |
|                                                                       | **A girl is standing next to a boy and looking ahead.**                       | 4.53 |
|                                                                       | Two little kids going toward the school bus.                                  | 4.94 |
| A wooden clock tower with a weather vain on top.                       | 5.0                                                                          |      |
|                                                                       | The top of a brick clock tower surrounded by birds.                           | 4.94 |
|                                                                       | Large stone clocktower with a giant clock on it.                              | 4.94 |
|                                                                       | A clock tower on top of a church with a weather vein.                        | 4.92 |
|                                                                       | An old wooden steeple with a clock on it.                                    | 5.0  |

Figure 10: Image → Text Retrieval Results from MS-COCO test set using the \( DE_{12T} \) model. The retrieved captions are ranked from top to bottom. The MS-COCO annotated pairs are marked in **bold** text.
References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

T. Baltruaitis, C. Ahuja, and L. Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. *SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation*. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yanshuang Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoded model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy. Association for Computational Linguistics.

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

Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *arXiv preprint arXiv:1811.08008*.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. 2019. Large-scale representation learning from visually grounded untranscribed speech. *arXiv preprint arXiv:1909.08782*.

Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2019. Graph-rise: Graph-regularized image semantic embedding. *arXiv preprint arXiv:1902.10814*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933.

Oier Lopez de Lacalle, Ander Salaberria, Aitor Soroa, Gorka Aizkune, and Eneko Agirre. 2020. Evaluating multimodal representations on visual semantic textual similarity. In *Proceedings of the Twenty-third European Conference on Artificial Intelligence (ECAI)*, Santiago Compostela, Spain.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon’s mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 139–147. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hyponymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
Su Wang, Rahul Gupta, Nancy Chang, and Jason Baldridge. 2019. A task in a suit and a tie: Paraphrase generation with semantic augmentation. pages 7176–7183.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. arXiv preprint arXiv:1902.08564.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019b. Multilingual universal sentence encoder for semantic retrieval. arXiv preprint arXiv:1907.04307.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2:67–78.