

Quality Estimation for Image Captions Based on Large-scale Human Evaluations

Tomer Levinboim Ashish V. Thapliyal Piyush Sharma Radu Soricut

Google Research

Venice, CA 90291

{tomerl, asht, piyushsharma, rsoricut}@google.com

Abstract

Automatic image captioning has improved significantly over the last few years, but the problem is far from being solved, with state of the art models still often producing low quality captions when used in the wild. In this paper, we focus on the task of Quality Estimation (QE) for image captions, which attempts to model the caption quality from a human perspective and *without* access to ground-truth references, so that it can be applied at prediction time to detect low-quality captions produced on *previously unseen images*. For this task, we develop a human evaluation process that collects coarse-grained caption annotations from crowdsourced users, which is then used to collect a large scale dataset spanning more than 600k caption quality ratings. We then carefully validate the quality of the collected ratings and establish baseline models for this new QE task. Finally, we further collect fine-grained caption quality annotations from trained raters, and use them to demonstrate that QE models trained over the coarse ratings can effectively detect and filter out low-quality image captions, thereby improving the user experience from captioning systems.

1 Introduction

Image captioning technology produces automatic image descriptions using natural language with the goal of being consumed by end-users that may not be able to directly access the images. This need arises either because the user has a permanent condition (accessibility for visually impaired people), or due to a temporary situation where the user cannot use the visual modality (such as limited bandwidth, or smart voice-assistant). In any of these situations, exposing the end-users to a generated caption that is incorrect negatively impacts user-trust, as it can have undesirable consequences for how they act next (for example, how they comment on a social-media site based on their misguided understanding).

In this paper, we propose to mitigate such risks through Quality Estimation (QE) of image captions. That is, we propose to automatically compute a quality estimation score $QE(image, caption)$ for a generated caption, and use it to control the quality of the captions presented to the user. For example, by filtering out captions with a low QE score (below a carefully chosen threshold), only high scoring captions would be served thereby minimizing the risks associated with low-quality captions.

We emphasize two aspects of QE that have guided us in our design choices: First, the QE task is distinct from the model selection task: model selection measures output similarity to a fixed, ground-truth annotated dataset during training time (with traditional offline solutions such as CIDEr and SPICE). In contrast, a QE model estimates the caption quality with respect to the input image only and does so on *previously unseen samples at prediction time* where ground-truth captions are unavailable. Second, a QE model’s goal is to assess the *caption* as a whole and relate it to the *image* content in a way that $QE(image, caption)$ aligns with human understanding of language and their perception of visual information.

To address these aspects we develop an image-caption evaluation process for collecting vast amounts of human judgements. Specifically, we design the process to elicit only the type of human signal that is required for quality estimation – human annotators are shown the image and asked to evaluate the caption as a whole by simply answering whether it is good or not. This type of high level feedback trades away the ability to understand in what way the caption is wrong, but its simplicity enables scaling up human evaluations to cover many more images, which promotes the generalization of the QE model to unseen images.

The dataset resulting from the evaluation process includes captions generated by various image-captioning model over 16,000 unique images from

the Open Image Dataset (Kuznetsova et al., 2018) for a total of 55,000 *unique* $\langle \text{image}, \text{caption} \rangle$ pairs, over which we collected approximately 600,000 binary human ratings. We denote this dataset as Caption-Quality, provide extensive details on its generation process as well as make it publicly available¹, available.

The following summarizes our contributions:

1. We release the Caption-Quality dataset of roughly 65k human rated image-caption pairs, obtained by collecting approximately 600k binary human ratings in total. By analyzing the collected ratings, we show that they encode a stable and consistent signal about the caption.
2. We establish baseline results on the QE task and demonstrate that the signal encoded in the collected ratings is learnable, yet, cannot be trivially captured by an image-text similarity model trained over a large scale image-captioning dataset.
3. We further test our QE models, trained over the Caption-Quality dataset, and show that they can successfully rank correct-and-helpful captions higher than incorrect or unhelpful ones, even though they were never exposed to such a fine-grained signal. This is done by collecting additional fine-grained caption annotations from trained human raters, over images that are out-of-domain for the QE model.

2 Related Work

Our paper is most similar to work done on evaluation metrics of image captions, where the main difference is that QE does not have access to the ground truth captions.

Quality estimation has more than a decade long history in the Machine Translation (MT) field, from the early work based on feature engineering (Specia et al., 2009; Soricut and Echiabi, 2010), to more recent neural-network-based approaches (Kreutzer et al., 2015; Kim and Lee, 2016; Kim et al., 2017). The QE track at the WMT conference (Specia et al., 2019) has been running for several years, with multiple participants and notable improvements in model performance over the years. However, there are significant differences in the formulation of the QE task between MT and image captioning, most notably the fact that the MT formulation is

uni-modal (text-only alignment). As a result, solutions for QE in the MT context tend to focus on feature-engineering that exploits this aspect (Specia et al., 2013; Kreutzer et al., 2015; Martins et al., 2017; Wang et al., 2018). In contrast, QE for Image Captioning is a bi-modal problem (image-and-text alignment), and therefore better suited to approaches based primarily on deep feature representations and multi-modal feature integration, as we present in this paper.

Beyond quality estimation modeling, the issue of effectively using quality estimators to improve the accessibility use-case for Blind or Visually Impaired (BVI) people has been previously studied (MacLeod et al., 2017). The main question of their study is how to best inform the BVI user about the uncertainty around the generated captions, experimenting with framing the captions using phrases like “I’m not really sure but I think it’s \$CAPTION” or “I’m 98% sure that’s \$CAPTION”. The findings are relevant in that BVI users of this technology have difficulties calibrating themselves into trusting or distrusting \$CAPTION, mostly because there is no alternative form of reference for the image content. Therefore, if the caption provided to them (even accompanied by “I’m not really sure but ...”) is in dissonance with the rest of the context (as it may be available in text form, e.g., as part of a tweet thread as in the study cited above), they tend to resolve this dissonance not by believing that the caption is wrong, but by constructing scenarios or explanations that would somehow connect the two sources of information. To mitigate this problem, we propose a thresholding-based approach that simply decides whether to show a caption or not based on a QE model’s prediction (See section 6.2).

3 Building the Caption-Quality Dataset

The key contribution of this paper is the Caption-Quality dataset, a large collection of binary human judgments on the quality of machine-generated image captions (in English). Below, we describe the dataset generation process, as well as the rating collection process with which we collect approximately 600,000 binary ratings via crowdsourcing. We then provide an analysis of the ratings which shows that they contain a consistent signal about the captions. Note that in the experiments (section 6.2), we further verify that indeed this signal captures the quality of the caption as perceived by

¹<https://github.com/google-research-datasets/image-caption-quality-dataset>

trained humans annotators.

3.1 Image-Caption Generation

The starting point for our dataset is the Open Images Dataset (OID) (Kuznetsova et al., 2018) from which we randomly sample 16,000 images and then, for legal and privacy concerns, filter out those which contain faces². The choice for OID images is driven by their image copyright status (CC BY) and the fact that they are out-of-domain for popular image captioning datasets such as COCO and Conceptual Captions.

To generate a diverse set of captions for annotation, we used several variants of Transformer-based (Vaswani et al., 2017) image-captioning models, trained on the Conceptual Captions dataset (Sharma et al., 2018), which consists of 3.3M training and $\sim 15,000$ validation image-caption pairs. As previous work indicates (Sharma et al., 2018), for out-of-domain images (OID), captions produced by Conceptual Captions trained models tend to have higher quality compared to captions produced by COCO-trained models.

All of the models are trained to minimize the ground-truth caption perplexity; however, they differ on several important aspects (which contributes to caption diversity): the image feature representations, the number of object detection results they use, and the caption decoding procedure. We briefly discuss these differences below; for further details, see (Sharma et al., 2018; Changpinyo et al., 2019).

Global Image Representation Our captioning models use one of the following pretrained image encoders: (1) The Inception-ResNet-v2 model (Szegedy et al., 2016), (2) The Picturebook image encoder (Kiros et al., 2018), or, (3) The Graph-RISE model (Juan et al., 2019), a ResNet-101 model (He et al., 2016) trained for an image classification task at ultra-fine granularity levels.

Object Representations The identification of objects in an image is done using a Faster R-CNN model, training it to predict both 1,600 object and 400 attribute labels in Visual Genome (Krishna et al., 2017), following the Bottom-Up Top-Down setting (Anderson et al., 2018). In terms of featurization for the identified bounding boxes, we use variants that include a ResNet-101 model pretrained on ImageNet (Russakovsky et al., 2015)

²Detected using the Google Cloud Vision API, <https://cloud.google.com/vision/>

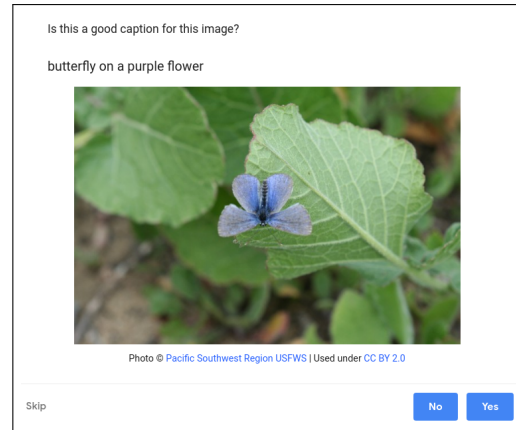


Figure 1: Our caption evaluation interface. Raters indicate whether the caption is good/bad, or, they can skip.

and one pre-trained using the Graph-RISE model (Juan et al., 2019).

Object Labels In addition to object-level representations, we detect object labels over the entire image, using a ResNet object-detection classifier trained on the JFT dataset (Hinton et al., 2015). The classifier produces a list of detected object-label identifiers, sorted in decreasing order by the classifier’s confidence score. These identifiers are then mapped to embeddings o_j using an object-label embedding layer which is pre-trained to predict label co-occurrences in web documents using a word2vec approach (Mikolov et al., 2013).

Decoding To further increase caption variance, we use either greedy decoding or beam search with beam width 5.

3.2 Fast&Simple Human Annotation

Traditional approaches for human evaluation of automatically generated text, such as for image captioning (Vinyals et al., 2016) and machine translation (Banchs et al., 2015), approach the task by collecting human ratings across multiple evaluation dimensions, such as correctness, informativeness and fluency. Such fine-grained evaluations are typically used to expose model deficiencies during

Set	Samples	Unique Images	Unique Captions	Unique Models
Train	58354	11027	34532	11
Dev	2392	654	1832	4
Test	4592	1237	3359	4

Table 1: The Caption-Quality dataset statistics

development and can also assist during model selection. However, obtaining fine-grained rating on a large scale is a slow and costly process because it requires extensive manual labor by professionally trained human annotators. Furthermore, it is not immediately clear how the resulting multi-dimensional ratings can be combined to estimate the *overall* caption quality in a human-like manner.

To avoid these complications we develop an evaluation process that asks the human evaluators to rate the generated text not per dimension, but *as a whole*. The benefits of our approach are threefold: (1) the collected ratings better align with our end goal of quality estimation from a human perspective (2) having a single question accelerates caption evaluation, and (3) it substantially reduces the training and qualification requirements from the raters, which further contributes to the scalability of the evaluation process.

Specifically, we formulate the quality of an image-caption as the binomial probability $p = P(\text{GOOD} | \text{image}, \text{caption})$ that can be estimated from the Bernoulli process in which every trial corresponds to a different rater. We then leverage Google’s crowdsourcing platform³ on which we present (image, caption) pairs and ask **volunteer** raters the following coarse binary question,

“Is this a good caption for the image?”.

The raters can then select YES/NO, or skip to the next sample (SKIP) (see Fig. 1). In adopting this approach we take into account the fact that the platform’s community consists of passionate volunteer raters, who may not have the linguistic background to provide fine-grained annotations. Furthermore, allowing the raters to skip captions reduces the risk of an undecided rater arbitrarily picking YES/NO just to move to the next image.

In order to reliably estimate the quality p we collect a high number of 10 ratings per image-caption sample. Once collected, the human ratings are further processed by: (1) filtering out (image, caption) entries that received more than 2 SKIP ratings (practically, the vast majority of images were kept), and (2) estimating p by averaging the 8 to 10 ratings r_i for each of the remaining (image, caption) pairs, and rounding to the closest score in $\{0, \frac{1}{8}, \dots, \frac{7}{8}, 1\}$, using the equation

$$\hat{p} = \text{round}(\text{mean}(r_i) * 8) / 8,$$

³<https://crowdsource.google.com>

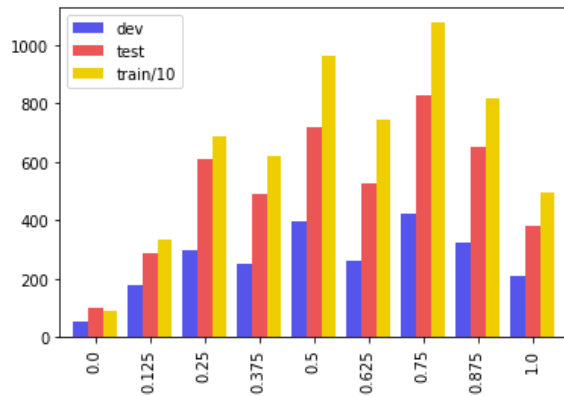


Figure 2: A histogram of the dev, test and train p^* . The train set values were divided by 10 (for scale).

where r_i is 0 for NO answers and 1 for YES.

The resulting dataset, which we call the Caption-Quality v1.0 dataset, is then split into three image-disjoint subsets, used as train, dev and test folds in our experiments. We provide statistics for these subsets in Table 1, as well as histograms of \hat{p} in Fig. 2. Finally, we provide examples from the dev set in Table 2.

3.3 Stability Analysis

As described above, the interpretation of what a “GOOD” caption means is left up to the raters, which could lead to unstable or inconsistent human ratings (Graham et al., 2013). In order to verify the stability of the quality ratings \hat{p} , we study

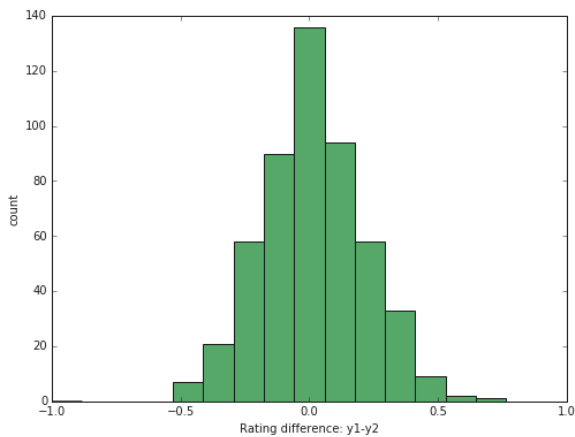


Figure 3: A set of 509 captions were evaluated twice by different sets of 10 raters and 4 weeks apart. The figure shows a histogram of average human score differences $(\hat{p}_1 - \hat{p}_2) \in [-1, 1]$, with scores \hat{p}_i ($i \in \{1, 2\}$) collected during the i -th evaluation. 85% of pairs are within a 0.25 distance, indicating that the evaluation setup produces a consistent and reliable signal.

the degree of agreement between different sets of 10 raters. We ran an evaluation over the same set of 509 image-captions twice, but 4 weeks apart⁴. An analysis of the difference of scores ($\hat{p}_1 - \hat{p}_2$) over these 509 pairs results in an almost zero mean (mean=0.015) as well as low variance (std=0.212). Figure 3 provides a histogram of the differences ($\hat{p}_1 - \hat{p}_2$) which clearly shows a concentration of the difference about 0. Furthermore, repeating this analysis over a different set of image-captions results in similar statistics.

In conclusion, the stability analysis shows that by collecting and averaging 8-10 coarse binary ratings, we obtain consistent and reproducible P(GOOD) estimates \hat{p} that are well-concentrated on a *sample-level*.

4 A Fine-Grained Caption Evaluation

We further collect fine-grained human annotations of image-captions to ascertain that the signal in the Caption-Quality dataset is beneficial for estimating the quality of image captions and filtering out low-quality ones. Specifically, we ask professional human annotators to evaluate image-captions across two specific dimensions: helpfulness and correctness⁵. Fig 4 shows the evaluation interface.

Distinguishing between correctness and helpfulness is particularly crucial for quality estimation, as it helps diagnose models that produce abstract

⁴The evaluation platform roughly guarantees that the ratings are provided by different subsets of raters.

⁵We also evaluate along a fluency dimension, but current captioning models tend to produce overall fluent outputs, which makes this dimension non-discriminative.

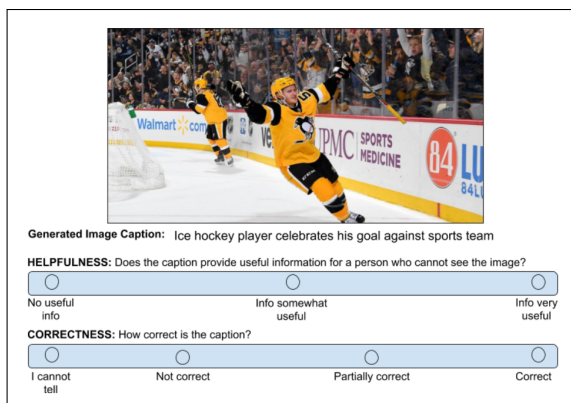


Figure 4: Fine-grained evaluation interface presented to professional raters. The raters determine whether (1) the caption provides a helpful description for a person who cannot see the image, (2) the information in the caption is correct.

or irrelevant captions which, while correct, do not provide useful image descriptions (specifically, for a person who is unable to see the image). For example, consider the correct yet abstract caption “Person in a sport event” compared to the more descriptive caption “Ice hockey player celebrates his goal against sports team” (See Fig 4). Another example of a correct but unhelpful caption is “A view of the game from my living room” because it conveys more information about the camera position rather than the actual image content. While the previously discussed Fast&Simple evaluation may assign all these captions with similar scores, the fine-grained evaluation is capable of capturing such nuanced differences.

We posit that the large-scale annotations obtained by the Fast&Simple approach will enable a model to distinguish between correct-and-helpful captions, and those that are not. We ran the fine-grained evaluation once over 2,700 images, collecting 3 ratings per image. The resulting dataset, denoted Caption-Ext is used for our extrinsic QE evaluations (Sec. 6.2).

5 Models

This section presents a simple bilinear QE model which learns to combine the image and caption features to arrive at a quality estimate $QE(image, caption)$. To construct the bilinear model we rely on expressive image and text representations that are produced by pretrained models that were themselves trained on vast amounts of uni-modal data. Note that aside from building on top of pretrained models, we restrict further modeling to a simple architecture. This was done in order to establish a baseline for our new QE task, as well as to remain focused on providing evidence that the signal in the *Caption – Quality* dataset is both learnable and beneficial for quality estimation of image captions.

5.1 A Bilinear QE model

Our bilinear neural network model relies on three input types: caption, image and object labels. These representations are produced by the following pretrained models:

Global Image Embedding For a global image representation, we used the latest Graph-RISE model version (Juan et al., 2019) which produces a compact image embedding i of dimension $D_i = 64$.







Image	Generated captions	Human rating
	a general view of atmosphere .	0.375
	this is a picture of a yacht .	0.75
	the yacht is a great place to take a rest .	0.875
	person and her husband take a walk .	0.125
	people walking along the beach	0.25
	people walking along the beach	0.5
	cat in the grass with a dog	0.25
	a tiger in the grass	0.25
	cat lying on the grass	0.75
	a police car in the middle of the road	0.125
	automobile model in the rain .	0.5
	vehicles drive through a flooded street	0.875
	plants for sale at the local market	0.625
	a selection of plants in the flower market	0.875
	flowers for sale at the market	1.0
	the team at the opening .	0.375
	the cast performs on stage .	0.5
	the cast of musical film	0.75

Table 2: Samples from the Caption-Quality dataset (dev fold). Images are paired with 3 captions and their corresponding mean human ratings. Repeated captions (which were generated by different captioning models) were rated by different sets of 10 raters, and tend to have similar scores (See stability analysis, cf. Figure 3). As can be seen, the higher scoring captions tend to include more information or contain fewer mistakes.

Using this model enables transfer learning for QE with respect to image representation.

Object Labels Embeddings Objects present in the image (e.g. “cat”, “vehicle”, “flower”) can help assess the correctness and helpfulness of a candidate caption, where the intuition is that the caption should likely mention the more salient objects. We use the object label model mentioned in Sec. 3.1, whose resulting embedding sequence is $O = (o_1, \dots, o_{|O|})$, where each o_j has dimension $D_o = 256$.

Caption Universal Sentence Embedding The caption text is embedded using a pretrained version of the Universal Sentence Encoder (USE) (Cer et al., 2018) into a $D_s = 512$ dimensional vector s . The USE model itself is trained on large amounts of English sources (Wikipedia, web news, discussion forums, etc.) and fine-tuned using supervised labels from the SNLI corpus (Bowman et al., 2015). We have alternatively tried a BERT (Devlin et al., 2019) model as an encoder, but observed it provides no additional gains (Alikhani et al., 2020)

Given these features, the bilinear QE model (illustrated in Figure 5) processes each individual feature using a dense layer with a leaky-ReLU activation

(Xu et al., 2015), and then combines each of the resulting vector pairs using bilinear layers (see below). All bilinear outputs are then concatenated and fed to a dense layer with a sigmoid activation, to produce the quality estimation \hat{y} .

5.1.1 Bilinear Layers

A bilinear layer models the inner product of its two inputs after applying a linear transformation to the second input. This layer is defined as:

$$b(x, y; B) = x^T B y = \langle x, B y \rangle \quad (1)$$

where $x \in R^{D_x}$ and $y \in R^{D_y}$ are input features, and $B \in R^{D_x \times D_y}$ is the learned parameter matrix. Linear and bias terms can be added by appending a constant 1 to each of x and y .

We use three such parameter matrices to capture the interaction between each pair of input-types:

1. $B_{o,i} \in R^{D_o \times D_i}$, applied to each of the object-label embeddings $[o_1, \dots, o_{|O|}]$ and the image embedding i .
2. $B_{o,s} \in R^{D_o \times D_s}$, applied to each of the object-label embeddings $[o_1, \dots, o_{|O|}]$ and the sentence embedding s
3. $B_{i,s} \in R^{D_i \times D_s}$, for the image embedding i and sentence embedding s .

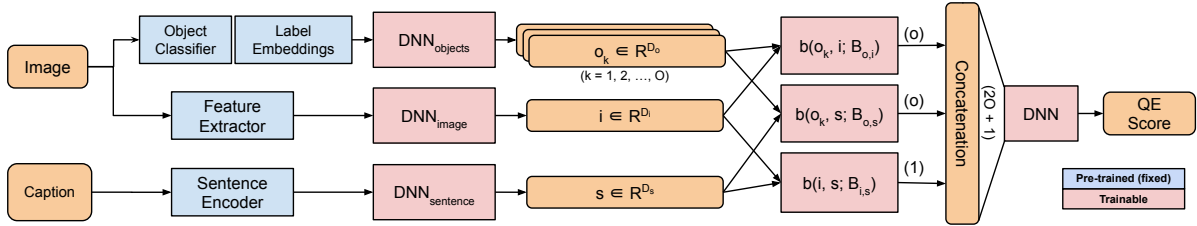


Figure 5: The bilinear QE model: Each input-modality pair has its own dedicated bilinear layer. The inputs to the model are pre-trained embeddings (blue) for the image, caption and object-label input types. The model parameters (pink) may further be warm-started by pretraining the model on an image-text similarity task (see Section 5.2).

5.2 An Image-Text Similarity Baseline

Having the large scale Conceptual Captions dataset (Sharma et al., 2018) opens up the option to pre-train a QE model on an image-text similarity task (Cui et al., 2018) before fine-tuning on the Caption-Quality dataset. We exercise this option by setting up a classification task whose goal is to match each image within a mini-batch with its corresponding ground truth caption. Specifically, we feed the bilinear QE model mini-batches of size 256 and train it to detect the ground-truth caption of each image among the other ground-truth captions in the batch (along the lines of noise-contrastive estimation (Gutmann and Hyvärinen, 2010)). The pre-trained model achieves 62% accuracy over the Conceptual Captions dev set and serves as an image-text similarity baseline. In addition, its parameters serve as a fine-tuning initialization point that is better informed about the relationship between image and text compared to random initialization.

6 Experimental Results

All QE models are trained on the Caption-Quality training set (Section 3). We use Mean Squared Error ($MSE = \sum_{j=1}^B \frac{1}{N} (y_j - \hat{y}_j)^2$) as the loss function, where \hat{y}_j are the predicted scores and y_j the ground-truth human scores. For optimization, we use Adam (Kingma and Ba, 2015) with batch size $B = 256$ and tune the learning rate $lr \in \{1e-4, 1e-5, 1e-6\}$. Dropout rate is set to 0.2, and applied on the inputs of all trainable layers. The following pretrained models are fixed during optimization: the image encoder, the USE caption encoder, and object-label encoder. The number of object-labels is tuned over $\{0, 5, 10, 20\}$, while the pretrained variants were fixed to 16.

Model selection is done by picking the checkpoint that maximizes the dev set Spearman’s correlation $\rho^S(y, \hat{y})$. Specifically, compared to MSE (the

objective), the Spearman-based selection criterion better matches the intended use of the QE model, where at inference time, only images whose QE scores pass some threshold will be served. Since this threshold can be tuned, the absolute value of the predicted scores \hat{y} is not as critical as obtaining a monotonic relationship between the predicted and ground truth scores (using ρ^S as the loss function is less feasible due to non-differentiability).

6.1 Spearman’s ρ Analysis

We present in Table 3 our dev and test Spearman results based on selecting the best-performing model configurations over the dev set.

Rows 1 and 2 show the bilinear model achieves minor improvements given additional 20 object labels. The poor Spearman scores in row 3, which were obtained *without* fine tuning over the Caption-Quality dataset, demonstrate that predicting the human ratings cannot be trivially achieved with an image-text similarity model, even when trained on a large dataset as Conceptual Captions. On the other hand, after fine-tuning it for the QE task (row 4), both dev and test Spearman scores increase substantially by 6-7 Spearman points over the best non-pretrained variant, which demonstrates the effectiveness of bi-modal pretraining for the QE task.

6.2 Extrinsic Evaluation

So far we have shown that the signal in Caption-Quality is both consistent and learnable. In this section, we further show that the collected signal is effective for filtering out low-quality image captions. To do so, we evaluate the performance of Caption-Quality trained QE models over the Caption-Ext dataset, a more challenging setting which contains out-of-domain images (non-OID) and where each caption is annotated by three trained raters for its correctness and helpfulness (Sec. 4). Our analysis

Model	QE training features	learning rate	ρ_{dev}^S	ρ_{test}^S	MSE_{dev}	MSE_{test}
Bilinear	image, caption	1e-5	0.49	0.47	0.055	0.056
Bilinear	+ 20 object labels	1e-5	0.50	0.47	0.055	0.058
Bilinear (Pretrained)	-	1e-5	0.26	0.25	0.075	0.073
Bilinear (Pretrained)	image, caption, 16 labels	1e-5	0.57	0.53	0.053	0.053

Table 3: Spearman’s ρ^S scores on the Caption-Quality dev and test dataset (higher is better). The pretrained and fine-tuned bilinear model exhibits the best Spearman results on the QE task. MSE results show the same trend and are included for completeness.

reveals that QE models trained over the Caption-Quality dataset generalize well to this harder task, having the ability to distinguish between correct-and-helpful image-captions and those that are not, even though these models were never exposed to such fine-grained signal.

Specifically, for a given image, we define a caption as *Ext-Good* (extrinsically good) if a majority of raters agreed that it is at least partially-correct, and, a majority of raters agreed it is at least somewhat-useful. With this definition, we compute the Ext-Good precision and recall statistics of a QE model Q for each threshold $th \in [0, 1]$ using the following equations:

$$precision_{th}^Q = \frac{\sum_s 1_{Ext-Good}^s \cdot 1_{QE(s) > th}}{\sum_s 1_{QE(s) > th}} \quad (2)$$

$$recall_{th}^Q = \frac{\sum_s 1_{Ext-Good}^s \cdot 1_{QE(s) > th}}{\sum_s 1_{Ext-Good}^s} \quad (3)$$

where the indicator variable $1_{Ext-Good}^s$ is on only when s is Ext-Good, and similarly the indicator

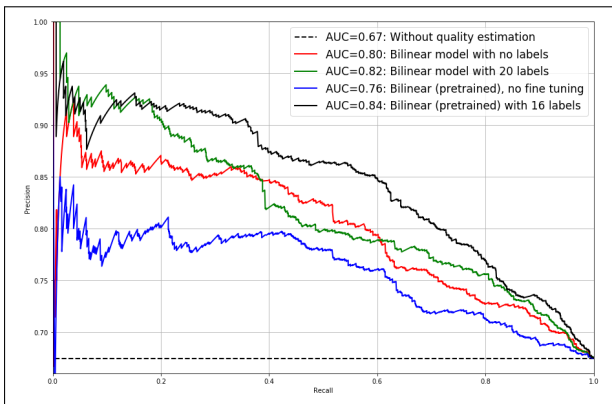


Figure 6: Precision-Recall curves for the various Bilinear models. AUC values are reported in the legend. The pretrained and fine-tuned model (black) attains the highest precision values across almost all recall values.

variable $1_{QE(s) > th}$ is on only when the QE score of sample s is higher than the threshold th .

Figure 6 shows the precision-recall curves and AUC scores for the same models analyzed in the previous section. A visual inspection of this figure shows that the precision of the pretrained and fine-tuned bilinear model (black) dominates the other models across almost all recall values. Indeed, in terms of AUC, the worst performing model is the image-text similarity baseline (blue; AUC=0.76) which has no access to the Caption-Quality dataset and its human ratings. On the other hand, the pretrained and fine-tuned model (which is also the Spearman maximizing model) attains the highest AUC score (AUC=0.84).

Put differently, to achieve precision=0.8 (i.e., 80% of served captions are both correct and helpful), the image-text similarity model would be thresholded to serve only its top 21% scoring image-captions (recall=0.21) while the pretrained and fine-tuned model would serve its top 71% scoring image-captions (recall=0.71, or x3.4 improvement). This analysis clearly demonstrates the usefulness of the Caption-Quality dataset for filtering out image-captions of low quality (where quality is determined by professional human raters).

7 Future Work

Beyond its relevance for the QE task, we expect that the collected signal in the Caption-Quality dataset will find usage in other image captioning tasks, such as (1) fine-grained caption evaluation (that is, caption classifiers that evaluate captions across multiple dimensions) for example, by way of pretraining against our dataset, as well as (2) improving caption generation itself, for example, by means of QE-based caption re-ranking, or by using the ratings in a reinforcement learning setup, as has recently been done by (Seo et al., 2020).

8 Conclusion

In this paper we discussed how low-quality image-captions can negatively impact end-users and proposed a thresholding solution that relies on quality estimation of image captions, where caption quality is defined from a human perspective. To make this solution feasible we developed a scalable human evaluation process with which we annotated a large number of image-captions with their human estimated quality. We provided supporting evidence that the resulting dataset contains a consistent and reliable signal, as well as reported experimental results over professionally labeled fine-grained caption annotations, which verify that QE models trained over the Caption-Quality dataset are effective at filtering out low-quality image captions.

To encourage further research in automatic evaluation of image-captions, we make available our large-scale dataset of human judgments at <https://github.com/google-research-datasets/Image-Caption-Quality-Dataset>.

References

- Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. **Cross-modal coherence modeling for caption generation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of CVPR*.
- R. E. Banchs, L. F. D’Haro, and H. Li. 2015. **Adequacy–fluency metrics: Evaluating mt in the continuous space model framework**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. **Universal sentence encoder for English**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.
- Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. 2019. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. In *EMNLP-IJCNLP*.
- Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge J. Belongie. 2018. Learning to evaluate image captioning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5804–5812.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. **Crowd-sourcing of human judgments of machine translation fluency**. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 16–24, Brisbane, Australia.
- Michael Gutmann and Aapo Hyvärinen. 2010. **Noise-contrastive estimation: A new estimation principle for unnormalized statistical models**. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy. PMLR.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*.
- Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. **Distilling the knowledge in a neural network**. In *NIPS Deep Learning and Representation Learning Workshop*.
- 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**. *CoRR*, abs/1902.10814.
- Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. **Predictor-estimator: Neural quality estimation based on target word prediction for machine translation**. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(1):3:1–3:22.
- Hyun Kim and Jong-Hyeok Lee. 2016. A recurrent neural networks approach for estimating the quality of machine translation output. In *Proceedings of the North American Chapter of the Association of Computational Linguistics*, pages 494–498.

- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.
- 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, Melbourne, Australia. Association for Computational Linguistics.
- Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. 2015. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT)*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73.
- Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Tom Duerig, and Vittorio Ferrari. 2018. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.
- Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people’s experiences with computer-generated captions of social media images. In *CHI*.
- André F. T. Martins, Marcin Junczys-Dowmunt, Fábio Kepler, Ramón Fernández Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252.
- Paul Hongsuck Seo, Piyush Sharma, Tomer Levinboim, and Radu Soricut. 2020. Reinforcing an image caption generator using off-line human feedback. In *Proceedings of AAAI*.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.
- Radu Soricut and Abdessamad Echihabi. 2010. *Trustrank: Inducing trust in automatic translations via ranking*. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Lucia Specia, Frederic Blain, Varvara Logacheva, Ramon Astudillo, and Andre Martins. 2019. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*.
- Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. pages 28–37.
- Lucia Specia, Kashif Shah, Jose Guilherme Camargo de Souza, and Trevor Cohn. 2013. QuEst - A Translation Quality Estimation Framework. In *Proceedings of the 51th Conference of the Association for Computational Linguistics (ACL), Demo Session*, Sofia, Bulgaria. Association for Computational Linguistics.
- Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1–1.
- Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for wmt18 quality estimation task. In *Proceedings of the Third Conference on Machine Translation (WMT)*.
- Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *Deep Learning Workshop, ICML*.