

# PhotoChat: A Human-Human Dialogue Dataset with Photo Sharing Behavior for Joint Image-Text Modeling

Xiaoxue Zang<sup>1</sup>, Lijuan Liu<sup>1</sup>, Maria Wang<sup>1</sup>, Yang Song<sup>2\*</sup>, Hao Zhang<sup>1</sup>, Jindong Chen<sup>1</sup>

<sup>1</sup> Google Research, <sup>2</sup> Kuaishou Technology

<sup>1</sup>{xiaoxuez, lijuanliu, mariawang, haozhangthu, jdchen}@google.com,

<sup>2</sup> yangsong@kuaishou.com

## Abstract

We present a new human-human dialogue dataset - PhotoChat, the first dataset that casts light on the photo sharing behavior in online messaging. PhotoChat contains 12k dialogues, each of which is paired with a user photo that is shared during the conversation. Based on this dataset, we propose two tasks to facilitate research on image-text modeling: a photo-sharing intent prediction task that predicts whether one intends to share a photo in the next conversation turn, and a photo retrieval task that retrieves the most relevant photo according to the dialogue context. In addition, for both tasks, we provide baseline models using the state-of-the-art models and report their benchmark performances. The best image retrieval model achieves 10.4% recall@1 (out of 1000 candidates) and the best photo intent prediction model achieves 58.1% F1 score, indicating that the dataset presents interesting yet challenging real-world problems. We are releasing PhotoChat to facilitate future research work among the community.

## 1 Introduction

As instant messaging tools gain enormous popularity in the recent decades, sharing photos as an approach to enhance the engagement of an online messaging conversation has become a pervasive routine communicative act (Lobinger, 2016). A survey conducted in 2010 reveals that 74% of teenagers in the US reported messaging a photo or video using their cell phone (Lenhart et al., 2010). In Britain, almost 70% of the internet users shared photos in 2013 (Dutton and Blank, 2013). Considering the proliferation of photo sharing, it's desirable to have an intelligent system that can assist users efficiently engaging in this process, i.e. suggesting the most relevant photos in correct timings. In order to achieve this goal, the intelligent system is expected to not only understand how humans

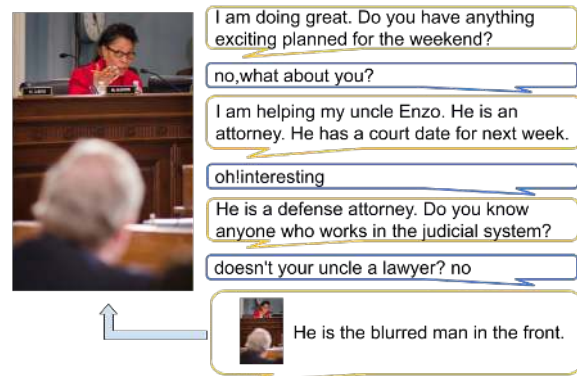


Figure 1: An example of how people share photos in a daily conversation.

communicate with each other, e.g. the natural language human speak, but also perceive images as human do. How to facilitate building such multi-modal system is the goal of this paper.

Though recently many image-text tasks have been proposed and are being actively studied to bridge language and vision, the majority of them are formulated as choosing or composing the text based on the understanding of given images, e.g. image captioning (Anderson et al., 2018), visual question answering (Antol et al., 2015), visual commonsense reasoning (Zellers et al., 2019), and image-grounded dialogue generation (Shuster et al., 2020). Contrary to these tasks, the photo sharing task focuses on the reverse process, i.e. selecting the image based on the understanding of text, as well as proposing different and unique challenges.

Firstly, different from the above popular multi-modal tasks, in photo-sharing task, the dialogue doesn't often explicitly mention the main visible content in the image. Instead of the main object of the photo, sometimes the background story, complemented by human imaginations, can be the focus of the chat. Figure 1 shows such an example, in which the person who shares the photo describes the event location "court" and the occupation "attorney" instead of the main object "lady" in the image. Secondly, the dialogue is not guaranteed

\*Research conducted while working at Google.

to be relevant to the image. For instance, it often contains greetings and chit-chats of other topics, as the first two turns in Figure 1 shows. In order to suggest the relevant photo, a smart system needs to decide which part of the dialogue can be used for suggesting the image. In contrast, in the traditional image-text tasks, the correct text is designed to be highly correlated with the image and has few distracting content. These photo sharing characteristics makes inferring the connection between the image and textual utterances challenging.

To highlight these challenges, we create PhotoChat - a human-human dialogue dataset in which one photo is shared from one person to the other during the conversation<sup>1</sup>. It is, as far as we know, the first dataset that captures the photo sharing activities. We selected images from OpenImage V4 dataset (Kuznetsova et al., 2020) as shared photos and used crowdsourcing plugins to generate 12,286 dialogues with an average of 10 turns per dialogue. During the dialogue collection, the photo is only visible to the side who is instructed to share the photo and then to both sides after it is being shared. Based on the collected dataset, we propose two tasks that are essential for building a photo suggest system: photo-sharing intent prediction task that predicts whether one intends to share the photo in the next conversation turn, and dialogue-based image retrieval task that retrieves the most relevant photo given the dialogue context. For both, we build baseline models, report and analyze their performances. The best photo-sharing intent prediction baseline model achieves 58.1% F1 score with 58.2% precision and 57.9% recall. The best cross-attention image retrieval model achieves 10.4% recall@1 out of 1000 candidates. We also propose a dual-encoder model that leverages object labels to encode image features, which achieves the best performance among all the models w/o cross-attention mechanisms.

In summary, our main contributions are:

- We create the first human-human dialogue with photo sharing acts via crowd-sourcing.
- We propose two new tasks to promote building an intelligent photo suggest system.
- We build baseline models and provide benchmarks for the new tasks. Our proposed image retrieval model outperforms all the prior models w/o cross-attention mechanisms. We im-

plement comprehensive analysis and ablation study to provide more insights.

## 2 Related Work

With the recent advances in deep learning, plenty of image-text datasets have been created and new image-text tasks are proposed based on them. These datasets have greatly stimulated the development of joint image-text models. In this section, we review the widely used image-text datasets and the state-of-the-art (SOTA) approaches for solving the image-text problems.

### 2.1 Image-text Dataset

Image-captioning datasets are first widely used for joint image-text modeling. MSCOCO (Lin et al., 2014) and Flickr30k (Young et al., 2014) that both contain five written caption descriptions for each image are the representative ones used for automated caption generation and cross-modal retrieval tasks. Conceptual Caption (Sharma et al., 2018) is yet another popular image caption dataset but contains an order of magnitude more images than MSCOCO. Because image captions usually only describe the main objects in the image and omit details, to facilitate understanding details of an image along with the reasoning behind them, Antol et al. (2015) introduced VQA which contains three question answer pairs for each image. A further work is VCR (Zellers et al., 2019) that not only requires a model to answer the question derived from the image but also provides a rationale explaining why its answer is right. It was created to teach the model to learn higher-order cognition and commonsense reasoning about the world.

Compared to the work above, Image-Chat (Shuster et al., 2020) and IGA (Mostafazadeh et al., 2017), which focus on the dialogues grounded in the image, are the most related work to ours. IGA includes 4k dialogues where each contains an image with a textual description of it, along with the questions and responses around the image. Due to its small scale, IGA can only be used for evaluation. Image-Chat is a larger scale dataset that consists of 202k image-grounded dialogues. However, both of them were created by asking the crowd workers to talk about a shared image to generate engaging conversation, which is different from the scenario of photo sharing where only one side can access the photo at the start of the conversation. Thus, neither can be used to build a photo-suggest system. In our

<sup>1</sup><https://github.com/google-research/google-research/tree/master/multimodalchat/>

work, we build a new dataset that highlights the challenges of building a photo-suggest system and is the first of its kind to the best of our knowledge.

## 2.2 Image-text Modeling

As the challenge for the photo-suggest system is to retrieve the most relevant image based on the textual utterances, we only review the related work on cross-modal retrieval.

Many models have been proposed for image-caption retrieval where one is required to retrieve the most relevant caption given an image or vice versa. The typical architecture consists of two separate encoders for image and text to first generate visual and textual embeddings. On top of them, a fusion layer, which can simply be a dot product, is used to generate the relevance score for each pair (Frome et al., 2013; Kiros et al., 2014; Parekh et al., 2020; Karpathy and Fei-Fei, 2015; Faghri et al., 2018). Then a triplet ranking loss or cross-entropy loss is employed to learn the latent visual-semantic alignment. VSE++ (Faghri et al., 2018) emphasizes on the hardest negatives by using the max of the hinge loss as the objectives and yielded a significant performance improvement. Stacked Cross Attention Network (SCAN) (Lee et al., 2018) further improves the performance by introducing the cross attention between image regions and word features. Recently, cross-modal transformer based architecture that are pretrained on large-scale image-text datasets via self-supervised learning has shown great advantages in bridging visual and textual embeddings. Multiple concurrent work (Lu et al., 2019; Chen et al., 2020; Li et al., 2019) have refreshed the best records on the benchmark datasets for the image-text retrieval tasks.

## 3 Dataset Creation

We select photos from Open Image Dataset V4 (OID) (Kuznetsova et al., 2020) and collect opened conversations on Amazon Mechanical Turk. Below describes the detailed image filtering, conversation generation, and data verification steps to ensure data quality.

### 3.1 Image-based Filtering

Since OID is large-scale and comprehensive, it contains images that are unlikely to be shared in the daily dialogue, such as images only about remote controls or fire hydrants. To create a dataset that is close to the reality, we filter images based on the

annotated object labels provided with OID.

Based on our investigation of the image-grounded dialogues and daily experiences, photos about four themes are commonly shared: people, food, animal, and product (in the shopping scenario), which are our focus in the dataset creation. From all the 600 object labels that appear in OID, we first enlist the labels that both belong to one of the four themes and have a high chance to appear in the commonly-shared photos. Labels like “traffic light”, “nail”, and “reptile” are excluded and labels like “girl”, “bagel”, and “camera” are included. This process selects 89 object labels (Appendix). We then generate an image pool by selecting those that contain any of the objects in the list. Note that for the objects of the people category, we add another criteria that it must be the main object, i.e. neither positioned in the margin of the image<sup>2</sup> nor extremely small<sup>3</sup> to exclude images that only have people as the background. Images are randomly selected from the image pool to generate conversations in the next step.

### 3.2 Conversation Generation

We randomly assigned two crowd workers to generate a conversation based on a given image. The image comes with an image description which presents the list of objects labels in the image. When the image contains humans, we assign a random name and relationship to one of the humans to help the workers refer to it and unfold the story. They are instructed to imagine talking with their friend. At the start of the task, only one side has access to the image and is instructed to drive the dialogue until it is fit to share the image with the other (website interfaces are shown in the Appendix). It is not restricted that they must message alternatively but the worker with the photo can't share the photo until the total number of the conversation turns reaches five. After sharing the photo, they can continue to chat until they wish to end the conversation and submit the dialogue.

### 3.3 Image&text-based Verification

Lastly, we use another set of in-house professional crowd workers to filter out the invalid dialogues generated in the above step. Dialogues are discarded if the association between the image and the dialogue is in-evident before the photo sharing act

<sup>2</sup>Center of the object is located within 0.1 of the image width/height to the border.

<sup>3</sup>Object width/length  $< 0.3 \times$  (image width/length).




| Good Example   | Good Example   | Bad Example  |
|--|--|--|
|  <p><b>A:</b> hows it going?<br/> <b>B:</b> just got back from vacation!!<br/> <b>A:</b> How was vacation? did you have fun?<br/> <b>B:</b> It was exciting! I took my grand-daughter to Greece and we saw so many beautiful ruins!<br/> <b>A:</b> oh wow! Greece, that's amazing. I bet you got amazing pictures of the ruins<br/> <b>B:</b> Yeah, we saw ancient temples and battlefields<br/> <b>B:</b> <b>Share the photo</b><br/> <b>A:</b> Wow! that's a great photo. you should post it on Insta too.<br/> <b>B:</b> Great idea! Thanks!</p> |  <p><b>A:</b> hey guess what i'm doing now ??<br/> <b>B:</b> What are you up to today?<br/> <b>A:</b> i'm preparing a pizza for the first time i include tomatoes,onions and so on<br/> <b>B:</b> Wow, you must be daring! Whoever taught you should have been confident on your progress.<br/> <b>A:</b> hey..... i'm almost done<br/> <b>B:</b> Must be yummy!<br/> <b>A:</b> wanna see my preparation?<br/> <b>A:</b> <b>Share the photo</b></p> |  <p><b>A:</b> How are you?<br/> <b>B:</b> I'm doing well. I've been watching Netflix because I can't go outside.<br/> <b>A:</b> Yeah, same here. Which show?<br/> <b>A:</b> And actually, I just found this picture of someone who should be a photographer.<br/> <b>B:</b> The office has been my go to.<br/> <b>B:</b> Really? Share the photo to me.<br/> <b>A:</b> <b>Share the photo</b><br/> <b>B:</b> Whoa! You were totally right<br/> <b>A:</b> It's a boy in neon green who I think wants to take photos in academic settings.<br/> <b>B:</b> This photo is so cool</p> |

Figure 2: Examples of PhotoChat dataset. The first two examples are included in the dataset while the last example is excluded in the verification step. **Share the photo** denotes the photo sharing act.

or the content is unnatural, contains inappropriate words, too many typos or broken English. Figure 2 displays examples of qualified and unqualified data. Note that the third unqualified dialogue can happen in a real conversation, yet the content/event of the image is not mentioned until the photo being shared, making it impossible for a model to learn the connection between the dialogue and the images and to suggest a photo in advance. Such dialogues are removed from the dataset in this step.

#### 4 Dataset Statistics

The collected dataset consists of 10,917 unique images and 12,286 dialogues. One image is shared in each dialogue. Based on the object labels of the shared image, we classify the dialogues into four categories: people, food, animals, and daily products. We split the dialogues into 10,086 train, 1,000 dev, and 1,000 test sets while keeping roughly the same distribution of the category across the splits. The detailed statistics of each split and in total are shown in Table 1. Note that the dialogue can have multiple category labels. For instance, if the shared image is about a girl playing with dogs, the dialogue belongs to both people and animals categories. Thus, the sum of the dialogues of each category (people/animal/food/product dial #) exceeds the total number of the dialogues (dial #) in

the table. In addition, some images in the training set are used in multiple dialogues.

Based on the statistics in the table, the average number of turns per dialogue is 12.7 and the average number of tokens per turn is 6.3. Since two sides are not restricted to speak alternatively, if the consecutive turns from the same side are combined as one turn, which is the conventional setting of other dialogue datasets, the average number of turns per dialogue and the average number of tokens per turn become 9.5 and 8.5. On average, people converse for 7 turns before sharing the photo.

#### 5 Task Definition

We decompose the problem of building a smart photo-suggest system into two separate tasks. The first is to detect if the user has the intent to share the photo in the next turn, which we call photo-sharing intent prediction task. The second is to retrieve the photo based on the dialogue context, which we call image retrieval task. Below describes the formal formulation of the problem settings.

Let  $P = \{p_1, p_2, \dots, p_M\}$  be the photo set where each  $p_i = (a_i, l_i)$ ,  $i \in [1, M]$  consists of image  $a_i$  and a list of objects  $l_i$  in it. Given the dialogue  $D = \{t_1, \dots, t_h, p_k, t_{h+1}, \dots, t_N\}$  where two participants speak alternatively,  $t_j$  ( $j \in [1, N]$ ) and  $p_k \in P$  respectively represent the utterance of turn  $j$  and

Table 1: PhotoChat statistics. Table shows the aggregated numbers. From left to right starting from the second column, the name of each column means “the unique number of images”, “the number of dialogues”, “the number of dialogues about people/food/animal/product”, “the number of turns”, “the number of turns when counting consecutive turns of the same speaker as one turn”, and “the number of tokens”. Turns in which photos are shared are excluded in the calculation.

| split | unique img # | dial # | people dial # | food dial # | animal dial # | product dial # | turn #  | turn* # | token # |
|-------|--------------|--------|---------------|-------------|---------------|----------------|---------|---------|---------|
| train | 8,917        | 10,286 | 6,376         | 4,465       | 1,072         | 884            | 130,546 | 97,586  | 827,154 |
| dev   | 1,000        | 1,000  | 606           | 424         | 87            | 109            | 12,701  | 9,533   | 80,214  |
| test  | 1,000        | 1,000  | 615           | 419         | 90            | 108            | 12,852  | 9,590   | 80,847  |
| total | 10,917       | 12,286 | 7,597         | 5,308       | 1,249         | 1,101          | 156,099 | 116,709 | 988,215 |

the shared image.  $t_h$  is the turn immediately before a photo sharing act. We also define the speaker information  $S = \{s_1, s_2, \dots, s_N\}$  where  $s_j$  ( $j \in [1, N]$ ), either 0 or 1, denotes the speaker of turn  $j$ .

**Photo-sharing intent prediction:** The goal of the intent prediction task is to predict whether a photo will be shared in the next turn for any  $t_j$  given all the turns before. In equation, it’s formulated as a binary classification task:

$$\forall j \in [1, h], C(t_{1:j}, s_{1:j}) \in \{0, 1\}, \quad (1)$$

where  $C$  is the intent prediction model taking the utterances and the speaker information of all the previous turns as the input and outputs a binary value. In the above case, it should only predicts 1 when  $j = h$ , otherwise 0. Note that whether the model make use of all the previous turns and the speaker information depends on the model design. We use F1 score, precision, and recall as the evaluation metrics for this task.

**Image retrieval:** Under the same settings, model  $R$  of the image retrieval task is expected to correctly retrieve  $p_k$  from  $P$  given the dialogue:

$$R(t_{1:h}, s_{1:h}, P) \in [1, M]. \quad (2)$$

During training, the candidate pool  $P$  is usually comprised of in-batch images while during evaluation,  $P$  contains all images in the test set. Following Karpathy and Fei-Fei (2015), we use Recall@K (R@K), computed as “the fraction of times a correct item was found among the top K results” as the evaluation metrics. Specifically, we choose R@1, R@5, and R@10, as well as the sum of them which we denote as “sum(R@1, 5, 10)” to evaluate the models.

## 6 Baselines

### 6.1 Photo-sharing Intent Prediction Model

To establish the baselines, we fine-tune three SOTA pretrained models - BERT (Devlin et al., 2018a),

ALBERT (Lan et al., 2020), and T5 (Raffel et al., 2020), as the pretrained models have achieved remarkable performance in many NLP tasks.

To adapt BERT and ALBERT to our settings, we concatenate all the previous turns ( $t_{1:j}$  in Equation 1) by [SEP] and prepend the concatenated text with [CLS] to generate the input to the model. We use the speaker information  $s_{1:j}$  as the segment id of the input. The output of [CLS] token is fed into two fully-connected layers, of which the output dimensions are respectively 128 and 2 to generate the final prediction. To utilize T5, we concatenate  $t_{1:j}$  by [SEP] and prepend the text with “predict share intent:” as the model input. We use cross entropy loss for all three models.

### 6.2 Image Retrieval Model

Our baselines consists of both statistical and neural network-based approaches, as elaborated below:

**Dual encoder:** We built a dual-encoder model similar to Parekh et al. (2020); Gillick et al. (2018), which separately encodes image and text leveraging SOTA pre-trained models. Its entire architecture is shown in Figure 3.

To encode the image, for each  $p_i = (a_i, l_i)$  we first resize the image  $a_i$  to  $224 \times 224$  and feed it into a pretrained ResNet (He et al., 2016) to generate  $A_i$ . A pretrained BERT is used to encode  $l_i$  to achieve the label embedding  $L_i$  which is the output of [CLS] token.  $L_i$  is concatenated with  $A_i$  to generate the image embedding. For encoding the dialogue context, we use a second pretrained BERT (Devlin et al., 2018b). Its input is the concatenation of all the prior utterances of the speaker who shares the photo. The output of [CLS] token is used as the contextual text embedding. Two fully connected layers are then used to separately project image and text embeddings into a joint image-text embedding space of dimension  $H$ . Then, the dot product of the normalized image embedding  $B_i$

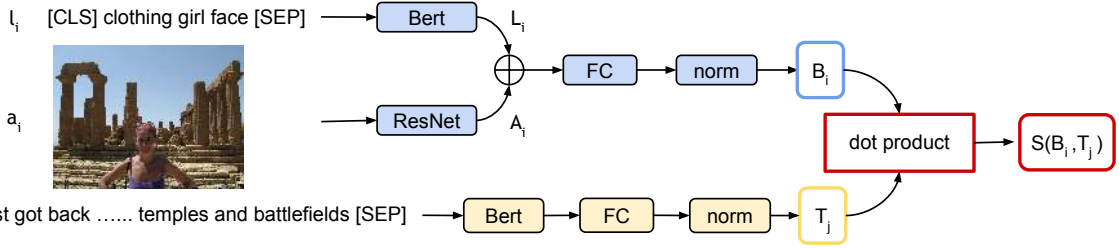


Figure 3: Our dual encoder. The first dialogue in Figure 2 is used as the input example. Image and text are encoded separately to generate their embeddings. The dot product of them is then used to compute the similarity score.

and text embedding  $T_j$  is used as the similarity score  $S(B_i, T_j)$ . Following Young et al. (2014); Gillick et al. (2018), bidirectional in-batch sampled cross entropy loss is employed:

$$l_{sm}(B_i, T_j) = -(S(B_i, T_j) - \log \sum_{\hat{T}_j} e^{S(B_i, \hat{T}_j)}) - (S(B_i, T_j) - \log \sum_{\hat{B}_i} e^{S(\hat{B}_i, T_j)}),$$

where  $\hat{B}_i$  and  $\hat{T}_j$  are the image embeddings and text embeddings of the other examples in the batch.

We also experiment with bidirectional in-batch hinge loss, defined as:

$$l_{sh}(B_i, T_j) = \sum_{\hat{T}_j} [\alpha - S(B_i, T_j) + S(B_i, \hat{T}_j)]_+ + \sum_{\hat{B}_i} [\alpha - S(B_i, T_j) + S(\hat{B}_i, T_j)]_+,$$

where  $\alpha$  is the margin parameter and  $[x]_+ \equiv \max(x, 0)$ . In our preliminary experiments, we observe cross entropy loss works better and implement most experiments with cross entropy loss.

**VSE++:** VSE++ (Faghri et al., 2018) is a simple and effective dual encoder model. It encodes the image and the text, which is the concatenation of all the previous utterances of the person who shares the photo in our case, separately by ResNet152 (He et al., 2016) and GRU (Cho et al., 2014). It is then followed by linear projections to map them into the joint embedding space. Finally, dot products of the normalized embeddings are used to compute the ranking scores. They innovatively make use of the hardest negatives, which are the negatives closest to the query, in the ranking loss function:

$$l_{mh}(B_i, T_j) = [\alpha - S(B_i, T_j) + S(B_i, \hat{T}_j^h)]_+ + [\alpha - S(B_i, T_j) + S(\hat{B}_i^h, T_j)]_+,$$

where  $\hat{T}_j^h = \operatorname{argmax}(S(B_i, \hat{T}_j))$  and  $\hat{B}_i^h = \operatorname{argmax}(S(\hat{B}_i, T_j))$  are the hardest negatives.

**SCAN:** SCAN (Lee et al., 2018) is a full cross attention model that captures the fine-grained interplay between image regions and text tokens to infer image-text similarity. It uses fasterRCNN (Ren et al., 2017) in conjunction with ResNet-101 to compute image region embeddings and bidirectional GRU to achieve text embeddings. Same as VSE++, SCAN uses hard negatives in the triple ranking loss function. Though it beats VSE++ on the image captioning tasks, it doesn't scale well to large-scale retrieval problems due to the high computational cost of cross attention.

**BM25:** BM25 (Amati, 2009) is a probabilistic retrieval function widely used for document retrieval. To adapt it to our settings, we directly utilize the object labels of each image  $l_j, j \in [1, m]$  as the document term. All the utterances before photo is shared are concatenated, tokenized and used as the query term to retrieve the image.

## 7 Experiments

### 7.1 Setup

The maximum sequence length of BERT, ALBERT, and T5 for the photo-sharing intent prediction task is 512. We choose checkpoints that achieve the best F1 score on the dev set for evaluation on the test set.

For our dual encoder model, the maximum sequence length of BERT is 128, the dimension of the joint image-text embedding space  $H$  is 512, and margin parameter  $\alpha$  is 0.2 for all the experiments. All parameters are trainable. We use the Adam optimizer ( $\beta_1 = 0.9, \beta_2 = 0.999$ ) and a learning rate that starts at  $5e-5$  and decays by 0.1% every 1000 steps. The models are trained on 32-core pod slices of Cloud TPU V3 Pod, with a per-replica batch size of 4. The loss is computed on item pairs aggregated from all replicas, which is over the global batch of 128 samples in this case.

For VSE++ and SCAN models, as GRU is not a pretrained encoder, directly training them on Pho-

Table 2: Experimental results of the baseline models for the photo-sharing intent prediction task. All numbers are in percentage.

| Model       | F1 $\uparrow$ | Precision $\uparrow$ | Recall $\uparrow$ |
|-------------|---------------|----------------------|-------------------|
| ALBERT-base | 52.2          | 44.8                 | 62.7              |
| BERT-base   | 53.2          | 56.1                 | 50.6              |
| T5-base     | 58.1          | <b>58.2</b>          | 57.9              |
| T5-3B       | <b>58.9</b>   | 54.1                 | <b>64.6</b>       |

Table 3: Number of negative turns and positive turns in each split of the dataset for the photo-sharing intent prediction task.

| Split | Number of negatives | Number of positives |
|-------|---------------------|---------------------|
| Train | 68,795              | 10,286              |
| Dev   | 6,802               | 1,000               |
| Test  | 6,748               | 1,000               |

toChat yields unpleasant results. As such, we first train them on MSCOCO and finetune them on PhotoChat for 20 epochs. We utilize the same setting as the single models that are reported to perform the best on the image-retrieval task on MSCOCO; more specifically, *VSE++ (ResNet, FT)* and *SCAN t-i AVG ( $\lambda_1 = 9$ )* following the annotations in the original papers.

## 7.2 Results of intent prediction

Table 2 presents model performance on the test set. We observe that T5 outperforms BERT and ALBERT in all metrics. Note that our dataset suffers from class imbalance that the negative examples outnumber the positive examples 3, which we suspect causes the low precision across all the models.

Figure 4 shows examples of the prediction by T5-3B model. Though a few turns are falsely predicted as positive (e.g. “*They were really pretty.*” and the second to last turn in example 2), it’s possible for the speaker to share the photo after this turn in real life, indicating that when to share a photo is subjective and the model may be more viable than the low precision would suggest. We also anticipate if the model has access to the set of photos the speaker can share, the accuracy can be elevated. In this case, the model will be able to infer that the photo in example 1 and 2 of Figure 4 are more likely to follow utterances about food and statues.

## 7.3 Results of image retrieval

Table 4 lists the experimental results on PhotoChat. Our dual encoder model is denoted as *DE*. *DE<sub>img</sub>* and *DE<sub>label</sub>* are the ablation models that only take the image  $a_i$  or image labels  $l_i$  as the input compared to the default architecture in Figure 3. CE, SH, MH represents cross entropy loss, hinge loss,

| Example 1   | Example 2   |
|---|---|
|   |    |
| <p>...</p> <p><b>B:</b> That’s good. I took the day off to spend with Isa.</p> <p><b>A:</b> Wow</p> <p><b>B:</b> It’s our anniversary.</p> <p><b>A:</b> Really needed some-times</p> <p><b>B:</b> We are getting brunch right now. Have you been to the blue herron cafe?</p> <p><b>A:</b> no I haven’t.</p> <p><b>B:</b> <u>They have a beautiful balcony.</u></p> <p><b>A:</b> tell me about it anything to share?</p> <p><b>B:</b> <u>Check out these amazing waffles!</u></p> | <p>...</p> <p><b>B:</b> Pretty good, I spent the day at the beach with my family</p> <p><b>A:</b> that sounds fun where at?</p> <p><b>B:</b> Spain. They had many statues out on the beach</p> <p><b>A:</b> I love the beach wow sounds beautiful!!!</p> <p><b>B:</b> <u>They were really pretty</u></p> <p><b>A:</b> did you take pics?</p> <p><b>B:</b> <u>I think so... There was this one sculpture that was unique... and the birds seemed to like it too haha</u></p> <p><b>A:</b> <u>oh let me see that!</u></p> |

Figure 4: Predictions by T5-3B model for the intent prediction task. Turns with underline are predicted as positive. False positives are marked in red while true positives are marked in blue. Best viewed in color.

and hinge loss using hard negatives. We attempt training *DE* on MSCOCO first and finetuning it on PhotoChat. These models are specially annotated with \*. We also experiment with different image encoders: ResNet-50 and ResNet-152, in combination with different label encoders: Bert-base and Bert-tiny. They are annotated in the brackets after the model names in Table 4. Among all the models, SCAN achieves the best performance with 10.4% R@1, 27% R@5, and 37.1% R@10, which is consistent with the prior work (Lee et al., 2018), demonstrating the power of the bottom-up cross attention. Among all the models that don’t have cross-attention, our model *DE\*(ResNet-152, Bert-tiny)* performs the best and beats a strong prior work VSE++, indicating the effectiveness of using image labels in the retrieval task.

**Ablation study:** By comparing *DE<sub>label</sub>(Bert-base)* and *DE<sub>img</sub>(ResNet-152)*, we find that using image features is more effective than using image label features, which is expected as images contain more information. Compared to the model using only image pixel values (*DE<sub>img</sub>(ResNet-152)*), adding the label features contributes to an increase of 1.3% in sum(R@1, 5, 10) to 66.4% (*DE(ResNet-152, Bert-base)*). Pretraining the model on MSCOCO further boosts it by 3.5% to

Table 4: Experimental results of the baseline models on image retrieval task.  $DE$  stands for our proposing dual encoders.  $DE_{img}$  only uses the image pixel values and  $DE_{label}$  only uses image labels to extract image features.  $DE^*$  is the model pretrained on MSCOCO. All numbers are in percentage.

| Model   | Loss function | R@1 ↑       | R@5 ↑       | R@10 ↑      | Sum(R@1, 5, 10) ↑ |
|---|---------------|-------------|-------------|-------------|-------------------|
| BM25  | -             | 6.6         | 15.4        | 23.0        | 45.0              |
| $DE_{label}$ (Bert-base)                        | CE            | 6.7         | 22.1        | 31.2        | 60.0              |
| $DE_{img}$ (ResNet-50)                          | CE            | 6.7         | 21.9        | 32.3        | 60.9              |
| $DE_{img}$ (ResNet-152)                         | CE            | 6.8         | 24.0        | 34.3        | 65.1              |
| DE(ResNet-152, Bert-base)                       | CE            | 8.1         | 23.7        | 34.6        | 66.4              |
| $DE^*$ (ResNet-152, Bert-base)                  | SH            | 8.0         | 22.0        | 31.0        | 61.0              |
| $DE^*$ (ResNet-152, Bert-tiny)                  | SH            | 7.1         | 23.3        | 33.0        | 63.4              |
| $DE^*$ (ResNet-152, Bert-base)                  | CE            | 8.5         | 26.1        | 35.3        | 69.9              |
| <b><math>DE^*</math>(ResNet-152, Bert-tiny)</b> | CE            | <b>9.0</b>  | <b>26.4</b> | <b>35.7</b> | <b>71.1</b>       |
| VSE++   | MH            | <b>10.2</b> | 25.4        | 34.2        | 69.8              |
| <b>SCAN</b>                                     | MH            | <b>10.4</b> | <b>27</b>   | <b>37.1</b> | <b>74.5</b>       |

69.9% ( $DE^*$ (ResNet-152, Bert-base)).

**Effect of encoders:** We observe that using a smaller model (Bert-tiny) to encode image labels yields better performance regardless of the loss function.  $DE^*$ (ResNet-152, Bert-tiny) improves sum(R@1, 5, 10) by 1.2% compared to  $DE^*$ (ResNet-152, Bert-base) when using cross entropy loss and 2.4% when using hinge loss. The reason might be that labels are a compact list of tokens and thus, using a smaller model alleviate the problem of overfitting. On the other hand, using a larger image encoder ResNet-152 produces better results that  $DE_{img}$ (ResNet-152) beats  $DE_{img}$ (ResNet-50) in sum(R@1, 5, 10) by 4.2%.

**Effect of loss function:** Our dual encoders work significantly better with cross entropy loss than hinge loss and their gap is about 8% in sum(R@1, 5, 10) as we compare the results of  $DE^*$ (ResNet-152, Bert-base) and  $DE^*$ (ResNet-152, Bert-tiny) models under different loss functions.

**Error analysis:** Figure 5 shows the qualitative results of  $DE^*$ (ResNet-152, Bert-tiny) given a text query. In the first example, the model ranks the relevant images of wine glasses and black tea at top instead of the groundtruth image where a man is holding a wine glass, which is easy to be neglected. In the second example, the model fails to distinguish puffins with ducks and infer the background from keyword “atlantic”. It illustrates the challenge of the image retrieval task under the dialogue context that it requires a model to pay attention to the details and the event, as discussed in Section 1. Figure 6 presents more prediction results including some wrong predictions by the model.

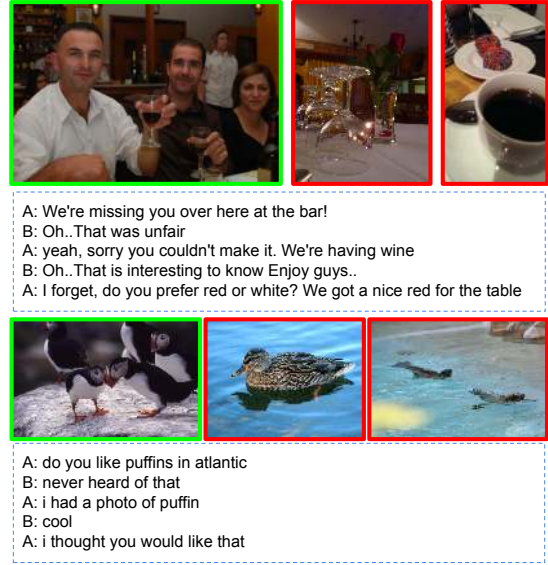


Figure 5: Predictions by  $DE^*$ (ResNet-152, Bert-tiny) for the image retrieval task. For each dialogue query, we show the groundtruth (first image in green) and the top-2 ranked images (in red). Best viewed in color.

## 8 Conclusion

We collected a 12k high-quality dialogue dataset that contains photo sharing activity via crowdsourcing. To facilitate research on building intelligent photo-suggest system, we have introduced two new challenging tasks that aim at improving the photo-sharing experience: photo-sharing intent prediction task and image retrieval task. That is, when given a dialogue, the system should predict whether the user has the intention to share the photo and which photo is suitable to be shared. We built baseline models for both tasks and report their performance with detailed analysis.

Besides the proposed two new tasks, our dataset can potentially be used in other dialogue related tasks, such as dialogue generation in the multi-modal dialogues, as well as inspiring new research





A: heu, how are you?  
 B: Im fine, having a day at the aquarium!  
 A: oh, I love going to the aquarium!  
 B: Next time I will invite you! We are enjoying the sea lions  
 A: Thanks, I would go for sure! What all did you see there?  
 B: The water looks perfectly clean today  
 A: The sea lions are always fun to watch! What was your favorite thing?  
 B: The sea lion swimming there are my favorite I love yo see them enjoying the water  
 A: They have lots of funny antics



A: how are you? I just bought a new jacket  
 B: Ok, trying to keep busy these days.  
 A: makes sense my jacket came in two sizes too small  
 B: Cool - I hope you really like your new jacket  
 A: it's not great I'm going to return it  
 B: Oh, no - can you exchange it?  
 A: no I might get money baack not entirely sure  
 B: I hope you can get your money back  
 A: it's a children's size, but I'm an adult



A: how do you like the school picnic photo ?  
 B: yes i do  
 A: i think alisha and her friends are adorable !  
 B: that is lovely  
 A: I hope to get it copied and framed for each of the parents as well as the teacher  
 B: that great  
 A: how happy they all look , makes me so envious as to being a child .  
 B: smiles



A: My day was pretty uneventful. Just watching tv now  
 B: that is how most of my days go by. any plans for vacation?  
 A: Going to Big Bear for 4th of July, you?  
 B: I want to take my kid to where you went not too long ago with Jordan  
 A: where? I've been lots of places recently  
 B: i think it was disney but not too sure. thats what i wanted to ask you there is a picture of your kid smiling with a chipmunk. Thats where i want to take him  
 A: I went to Disney World last summer. Still need to try Disney Land!

Figure 6: Predictions by  $DE^*(ResNet-152, Bert-tiny)$  for the image retrieval task. For each dialogue query, we show the top-5 ranked images from left to right. The ground-truth image is marked in green while the others are in red. Best viewed in color.

topics, such as composing automatic reply to the photos sent from others. We hope our dataset and modeling work can be beneficial for studies that focus on the interplay between image and dialogue.

## Acknowledgments

We thank Pranav Khaitan and Blaise Aguera y Arcas for the support and assistance; Yinfei Yang, David Bieber for reviewing the draft and providing the feedback; Janel Thamkul and Tulsee Doshi for doing the legal review of the dataset.

## References

- Giambattista Amati. 2009. *BM25*, pages 257–260. Springer US, Boston, MA.
- 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](#).
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [Uniter: Universal image-text representation learning](#).
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using RNN encoder–decoder for statistical machine translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- William Dutton and Grant Blank. 2013. *Cultures of the Internet: The Internet in Britain. Oxford Internet Survey 2013 Report*.
- Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. [Vse++: Improving visual-semantic embeddings with hard negatives](#).
- Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. 2013. [Devise: A deep visual-semantic embedding model](#). In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc.
- Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. [End-to-end retrieval in continuous space](#).
- K. He, X. Zhang, S. Ren, and J. Sun. 2016. [Deep residual learning for image recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- 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.
- A. Karpathy and L. Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137.
- Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. [Unifying visual-semantic embeddings with multimodal neural language models](#).
- Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [Albert: A lite bert for self-supervised learning of language representations](#).
- Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. *arXiv preprint arXiv:1803.08024*.
- A. Lenhart, Rich Ling, S. Campbell, and K. Purcell. 2010. Teens and mobile phones: Text messaging explodes as teens embrace it as the centerpiece of their communication strategies with friends.
- Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2019. [Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training](#).
- 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](#). *Lecture Notes in Computer Science*, page 740–755.
- Katharina Lobinger. 2016. [Photographs as things – photographs of things. a texto-material perspective on photo-sharing practices](#). *Information, Communication & Society*, 19(4):475–488.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. [Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks](#).
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. [Image-grounded conversations: Multimodal context for natural question and response generation](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,

pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2020. [Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco](#).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. [Faster r-cnn: Towards real-time object detection with region proposal networks](#). *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.

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*.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. [Image-chat: Engaging grounded conversations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, Online. Association for Computational Linguistics.

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.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [From recognition to cognition: Visual commonsense reasoning](#). In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

## A Dataset Creation & Details

The website interfaces used to collect dialogues are presented in Figure 7 and 8.

Table 5 shows the 89 object labels that we used to select the photos from Open Image Dataset for generating dialogues.



Figure 7: Website interface of the conversation generation task. It is only visible to the side who shares the photo.

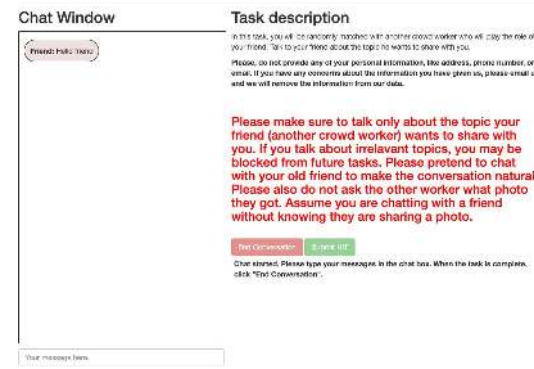


Figure 8: Website interface of the conversation generation task. It is only visible to the side who receives the photo.

Table 5: Object labels we use for image filtering.

| Category | Object labels   |
|----------|---|
| People   | Woman, Man, Girl, Boy, Human body, Face   |
| Food     | Bagel, Baked goods, Beer, Bread, Burrito, Cake, Candy, Cheese, Cocktail, Coffee, Cookie, Croissant, Dessert, Doughnut, Drink, Fast food, French fries, Hamburger, Hot dog, Ice cream, Juice, Milk, Pancake, Pasta, Pizza, Popcorn, Salad, Sandwich, Seafood, Snack, Taco, Tart, Tea, Waffle, Wine, Guacamole  |
| Animals  | Animal  |
| Products | Alarm clock, Backpack, Blender, Banjo, Bed, Belt, Computer keyboard, Computer mouse, Curtain, Guitar, Hair dryer, Hair spray, Harmonica, Humidifier, Jacket, Jeans, Dress, Earrings, Necklace, Fashion accessory, Bicycle, Blender, Calculator, Camera, Food processor, Jug, Mixing bowl, Nightstand, Oboe, Oven, Paper cutter, Pencil case, Perfume, Pillow, Personal care, Pizza cutter, Pressure cooker, Printer, Refridgerator, High heels, Skateboard, Slow cooker, Teddy bear, Teapot, Vase, Wall clock |