

ENGAGING IMAGE CAPTIONING VIA PERSONALITY

Anonymous authors

Paper under double-blind review

ABSTRACT

Standard image captioning tasks such as COCO and Flickr30k are factual, neutral in tone and (to a human) state the obvious (e.g., “a man playing a guitar”). While such tasks are useful to verify that a machine understands the content of an image, they are not engaging to humans as captions. With this in mind we define a new task, PERSONALITY-CAPTIONS, where the goal is to be as engaging to humans as possible by incorporating controllable style and personality traits. We collect and release a large dataset of 201,858 of such captions conditioned over 215 possible traits. We build models that combine existing work from (i) sentence representations (Mazaré et al., 2018) with Transformers trained on 1.7 billion dialogue examples; and (ii) image representations (Mahajan et al., 2018) with ResNets trained on 3.5 billion social media images. We obtain state-of-the-art performance on Flickr30k and COCO, and strong performance on our new task. Finally, online evaluations validate that our task and models are engaging to humans, with our best model close to human performance.

1 INTRODUCTION

If we want machines to communicate with humans, they must be able to capture our interest, which means spanning both the ability to understand and the ability to be engaging, in particular to display emotion and personality as well as conversational function (Jay & Janschewitz, 2007; Jonczyk & Jończyk, 2016; Scheutz et al., 2006; Kampman et al., 2019).

Communication grounded in images is naturally engaging to humans (Hu et al., 2014), and yet the majority of studies in the machine learning community have so far focused on function only: standard image captioning (Pan et al., 2004) requires the machine to generate a sentence which factually describes the elements of the scene in a neutral tone. Similarly, visual question answering (Antol et al., 2015) and visual dialogue (Das et al., 2017) require the machine to answer factual questions about the contents of the image, either in single turn or dialogue form. They assess whether the machine can perform basic perception over the image which humans take for granted. Hence, they are useful for developing models that understand content, but are not useful as an end application unless the human cannot see the image, e.g. due to visual impairment (Gurari et al., 2018).

Standard image captioning tasks simply state the obvious, and are not considered engaging captions by humans. For example, in the COCO (Chen et al., 2015) and Flickr30k (Young et al., 2014) tasks, some examples of captions include “a large bus sitting next to a very tall building” and “a butcher cutting an animal to sell”, which describe the contents of those images in a personality-free, factual manner. However, humans consider engaging and effective captions ones that “avoid stating the obvious”, as shown by advice to human captioners outside of machine learning.¹ For example, “If the bride and groom are smiling at each other, don’t write that they are smiling at each other. The photo already visually shows what the subject is doing. Rephrase the caption to reflect the story behind the image”. Moreover, it is considered that “conversational language works best. Write the caption as though you are talking to a family member or friend”.² These instructions for human captioners to engage human readers seem to be in direct opposition to standard captioning datasets.

In this work we focus on image captioning that is engaging for humans by incorporating personality. As no large dataset exists that covers the range of human personalities, we build and release a new dataset, PERSONALITY-CAPTIONS, with 201,858 captions, each conditioned on one of 215

¹<https://www.photoup.net/how-to-write-more-engaging-photo-captions/>

²<https://www.poynter.org/news/6-tips-writing-photo-captions>



Standard captioning output: A plate with a sandwich and salad on it.

Our model with different personality traits:

<i>Sweet</i>	That is a lovely sandwich.
<i>Dramatic</i>	This sandwich looks so delicious! My goodness!
<i>Anxious</i>	I'm afraid this might make me sick if I eat it.
<i>Sympathetic</i>	I feel so bad for that carrot, about to be consumed.
<i>Arrogant</i>	I make better food than this
<i>Optimistic</i>	It will taste positively wonderful!
<i>Money-minded</i>	I would totally pay \$100 for this plate.

Figure 1: Comparison of a standard captioning model compared to our TransResNet model’s predictions on the same image conditioned on various personality traits. Our model is trained on the new PERSONALITY-CAPTIONS dataset which covers 215 different personality traits. The standard captioning system used for comparison is the best COCO UPDOWN model described in Section 4.2.

different possible personality traits. We show that such captions are far more engaging to humans than traditional ones.

We then develop model architectures that can simultaneously understand image content and provide engaging captions for humans. To build strong models, we consider both retrieval and generative variants, and leverage state-of-the-art modules from both the vision and language domains. For image representations, we employ the work of Mahajan et al. (2018) that uses a ResNeXt architecture trained on 3.5 billion social media images which we apply to both. For text, we use a Transformer sentence representation following (Mazaré et al., 2018) trained on 1.7 billion dialogue examples. Our generative model gives a new state-of-the-art on caption generation on COCO, and our retrieval architecture, TransResNet, yields the highest known hits@1 score on the Flickr30k dataset. To make the models more engaging to humans, we then adapt those same architectures to the PERSONALITY-CAPTIONS task by conditioning the input image on the given personality traits, giving strong performance on our new task. In particular, when compared to human captions, annotators preferred our retrieval model’s captions over human ones 49.5% of the time, where the difference is not statistically significant.

2 RELATED WORK

A large body of work has focused on developing image captioning datasets and models that work on them. In this paper we also perform experiments on the COCO (Chen et al., 2015) and Flickr30k (Young et al., 2014) datasets, comparing to a range of models, including both generative models such as in (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018) and retrieval based such as in (Gu et al., 2017; Faghri et al., 2017; Nam et al., 2016). These setups measure the ability of models to understand the content of an image, but do not address more natural human communication.

A number of works have tried to induce more engaging captions for human readers. One area of study is to make the caption personalized to the reader, e.g. by using user level features such as location and age (Denton et al., 2015) or knowledge of the reader’s active vocabulary (Park et al., 2017). Our work does not address this issue. Another research direction is to attempt to produce amusing captions either through wordplay (puns) (Chandrasekaran et al., 2017) or training on data from humour websites (Yoshida et al., 2018). Our work focuses on a general set of personality traits, not on humour. Finally, closer to our work are approaches that attempt to model the style of the caption. Some methods have tried to learn style in an unsupervised fashion, as a supervised dataset like we have built in this work was not available. As a result, evaluation was more challenging in those works, see e.g. Mathews et al. (2018). Others such as You et al. (2018) have used small datasets like SentiCap (Mathews et al., 2016) with ~ 800 images to inject sentiment into captions. Gan et al. (2017) collect a somewhat bigger dataset with 10,000 examples, FlickrStyle10K, but only covers two types of style (romantic and humorous). In contrast, our models are trained on the PERSONALITY-CAPTIONS dataset that has 215 traits and $\sim 200,000$ images.

Our work can also be linked to the more general area of human communication, separate from just factual captioning, in particular image grounded conversations between humans (Mostafazadeh

Table 1: PERSONALITY-CAPTIONS dataset statistics.

Split	train	valid	test
Number of Examples	186,858	5,000	10,000
Number of Personality Types	215	215	215
Vocabulary Size	35559	5557	8137
Average Tokens per Caption	11.6	11.2	11.4

et al., 2017) or dialogue in general where displaying personality is important (Zhang et al., 2018). In those tasks, simple word overlap based automatic metrics are shown to perform weakly (Liu et al., 2016) due to the intrinsically more diverse outputs in the tasks. As in those domains, we thus also perform human evaluations in this work to measure the engagingness of our setup and models.

In terms of modeling, image captioning performance is clearly boosted with any advancements in image or text encoders, particularly the former. In this work we make use of the latest advancements in image encoding by using the work of Mahajan et al. (2018) which provides state-of-the-art performance on Imagenet image classification, but has so far not been applied to captioning. For text encoding we use the latest advances in attention-based representations using Transformers (Vaswani et al., 2017); in particular, their use in retrieval models for dialogue by large-scale pretraining (?) is adapted here for our captioning tasks.

3 PERSONALITY-CAPTIONS

The PERSONALITY-CAPTIONS dataset is a large collection of (image, personality trait, caption) triples that we collected using crowd-workers, and will be made publicly available upon acceptance.

We considered 215 possible personality traits which were constructed by selecting a subset from a curated list of 638 traits³ that we deemed suitable for our captioning task. The traits are categorized into three classes: positive (e.g., sweet, happy, eloquent, humble, perceptive, witty), neutral (e.g., old-fashioned, skeptical, solemn, questioning) and negative (e.g., anxious, childish, critical, fickle, frivolous). Examples of traits that we did not use are allocentric, insouciant, flexible, earthy and invisible, due to the difficulty of their interpretation with respect to captioning an image.

We use a randomly selected set of the images from the YFFC100M Dataset⁴ to build our training, validation and test sets, selecting for each chosen image a random personality trait from our list.

In each annotation round, an annotator is shown an image along with a trait. The annotators are then asked to write an engaging caption for the image in the context of the personality trait. It was emphasized that the personality trait describes a trait of the author of the caption, not properties of the content of the image. See Section D in the appendix for the exact instructions given to annotators.

4 MODELS

We consider two classes of models for caption prediction: retrieval models and generative models. Retrieval models produce a caption by considering any caption in the training set as a possible candidate response. Generative models generate word-by-word novel sentences conditioned on the image and personality trait (using a beam). Both approaches require an image encoder.

4.1 IMAGE ENCODERS

We build both types of model on top of pretrained image features, and compare the performance of two types of image encoders. The first is a residual network with 152 layers described in (He et al., 2015) trained on Imagenet (Russakovsky et al., 2014) to classify images among 1000 classes, which we refer to in the rest of the paper as *ResNet152* features. We used the implementation provided in the torchvision project (Marcel & Rodriguez, 2010). The second is a ResNeXt $32 \times 48d$ (Xie

³<http://ideonomy.mit.edu/essays/traits.html>

⁴<https://multimediacommons.wordpress.com/yfcc100m-core-dataset/>

et al., 2016) trained on 3.5 billion Instagram pictures following the procedure described by Mahajan et al. (2018), which we refer to in the rest of the paper as *ResNeXt-IG-3.5B*. The authors provided the weights of their trained model to us. Both networks embed images in a 2048-dimensional vector which is the input for most of our models. In some of the caption generation models that make use of attention, we keep the spatial extent of the features by adapting activation before the last average pooling layer, and thus extract features with $7 \times 7 \times 2048$ dimensions.

4.2 CAPTION GENERATION MODELS

We re-implemented three widely used previous/current state-of-the-art methods (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018) for image captioning as representatives of caption generation models. We refer them as SHOWTELL, SHOWATTTELL and UPDOWN respectively.

Image and Personality Encoders We extract the image representation r_I using the aforementioned image encoders. The SHOWTELL model uses image features with 2048 dimensions and the other models use image features with $7 \times 7 \times 2048$ dimensions. In the case where we augment our models with personality traits, we learn an embedding for each trait, which is concatenated with each input of the decoder.

Caption Decoders The SHOWTELL model first applies a linear projection to reduce image features into a feature vector with 512 dimensions. Similar to Vinyals et al. (2015), this embedding is the input for a LSTM model that generates the output sequence. In SHOWATTTELL, while the overall architecture is similar to Xu et al. (2015), we adopt the modification suggested by Rennie et al. (2017) and input the attention-derived image features to the cell node of the LSTM. Finally, we use the UPDOWN model exactly as described in Anderson et al. (2018).

Training and Inference We perform a two-stage training strategy to train such caption generation models as proposed by Rennie et al. (2017). In the first stage, we train the model to optimize the standard cross-entropy loss. In the second stage, we perform policy gradient with REINFORCE to optimize the non-differentiable reward function (CIDEr score in our case). During inference, we apply beam search (beam size=2) to decode the caption.

4.3 CAPTION RETRIEVAL MODELS

We define a simple yet powerful retrieval architecture, named TransResNet. It works by projecting the image, personality, and caption in the same space S using image, personality, and text encoders.

Image and Personality Encoders The representation r_I of an image I is obtained by using the 2048-dimensional output of the image encoder described in Sec. 4.1 as input to a multi-layer perceptron with ReLU activation units and a final layer of 500 dimensions. To take advantage of personality traits in the PERSONALITY-CAPTIONS task, we embed each trait to a 500-dimensional vector to obtain its representation r_P . Image and personality representations are then summed.

Caption Encoders Each caption is encoded into a vector r_C of the same size using a Transformer architecture (Vaswani et al., 2017), followed by a two layer perceptron. We try two sizes of Transformer: a larger architecture (4 layers, 300 hidden units, 6 attention heads) and a smaller one (2 layers, 300 hidden units, 4 attention heads). We consider either training from scratch or pretraining our models. We either pretrain only the word embeddings, i.e. where we initialize word vectors trained using fastText (Bojanowski et al., 2016) trained on Wikipedia, or pretrain the entire encoder. For the latter, we follow the setup described in Mazaré et al. (2018): we train two encoders on a next-utterance retrieval task on a dataset of dialogs containing 1.7 billion pairs of utterances, where one encodes the context and another the candidates for the next utterance, their dot product indicates the degree of match, and they are trained with negative log-likelihood and k -negative sampling. We then initialize our system using the weights of the candidate encoder only, and then train on our task.

For comparison, we also consider a simple bag-of-words encoder (pretrained or not). In this case, r_C is the sum of the 300-dimensional word embeddings of the caption.

In each case, given an input image and personality trait (I, P) and a candidate caption C , the score of the final combination is then computed as $s(I, P, C) = (r_I + r_P) \cdot r_C$.

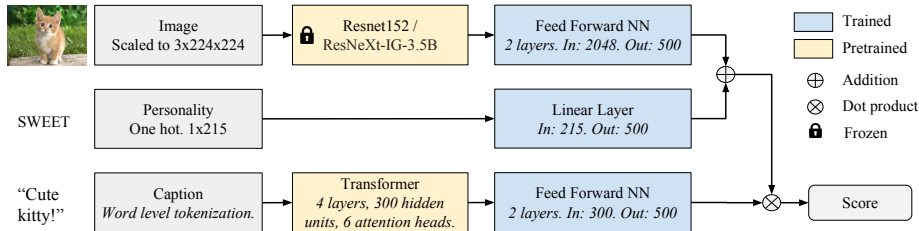


Figure 2: Our architecture TransResNet, used for our retrieval models.

Training and Inference Given a pair I, P , and a set of candidates (c_1, \dots, c_N) , at inference time the predicted caption is the candidate c_i that maximizes the score $s(I, P, c_i)$. At training time we pass a set of scores through a softmax and train to maximize the log-likelihood of the correct responses. We use mini-batches of 500 training examples; for each example, we use the captions of the other elements of the batch as negatives. Our overall TransResNet architecture is detailed in Figure 2.

5 EXPERIMENTS

We first test our architectures on traditional caption datasets to assess their ability to factually describe the contents of images in a neutral tone. We then apply the same architectures to PERSONALITY-CAPTIONS to assess their ability to produce engaging captions conditioned on personality. The latter is tested with both automatic metrics and human evaluation of engagingness.

5.1 AUTOMATIC EVALUATION ON TRADITIONAL CAPTION DATASETS

Generative Models For our generative models, we test the quality of our implementations of existing models (SHOWTELL, SHOWATTTELL and UPDOWN) as well as the quality of our image encoders, where we compare ResNet152 and ResNeXt-IG-3.5B. We report performance on the COCO caption dataset (Lin et al., 2014). We evaluate BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016) and compare model’s performances to state-of-the-art models under Karpathy & Fei-Fei (2015)’s setting.

The results are shown in Table 3. Models trained with ResNeXt-IG-3.5B features consistently outperform their counterparts with ResNet152 features, demonstrating the effectiveness of ResNeXt-IG-3.5B beyond the original image classification and detection results in Mahajan et al. (2018). More importantly, our best model (UPDOWN) either outperforms or is competitive with state-of-the-art single model performance (Anderson et al., 2018) across most metrics (especially CIDEr).

Retrieval Models We compare our retrieval architecture, TransResNet, to existing models reported in the literature on the COCO caption and Flickr30k tasks. We evaluate retrieval metrics R@1, R@5, R@10, and compare our model performance to state-of-the-art models under the setting of (Karpathy & Fei-Fei (2015)). The results are given in Table 4 (for more details, see Tables 7 and 10 in the appendix for COCO and Flickr30k, respectively). For our model, we see large improvements using ResNeXt-IG-3.5B compared to Resnet152, and stronger performance with a Transformer-based text encoding compared to a bag-of-words encoding. Pretraining the text encoder also helps substantially (see Appendix A for more analysis of pretraining of our systems). Our best models are competitive on COCO and are state-of-the-art on Flickr30k by a large margin (68.4 R@1 for our model vs. 56.8 R@1 for the previous state-of-the-art).

5.2 AUTOMATIC EVALUATIONS ON PERSONALITY-CAPTIONS

Generative models We first train the aforementioned caption generation models without using the personality traits. This setting is similar to standard image captioning, and Table 5 shows that the three caption generation models that we considered are ranked in the same order, with the UPDOWN model being the most effective. The best results are again obtained using the ResNeXt-IG-3.5B features. Adding the embedding of the personality trait allows our best model to reach a CIDEr score of 22.0, showing the importance of modeling personality in our new task.

Table 2: Predictions from our best TransResNet model on the PERSONALITY-CAPTIONS valid set.


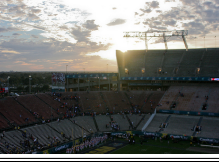


Image	Personality	Generated comment
	Anxious	I love cats but i always get so scared that they will scratch me.
	Happy	That cat looks SO happy to be outside.
	Vague	That’s a nice cat. Or is it a lion?
	Dramatic	That cat looks so angry; it might claw your eyes out!
	Charming	Awww, sweet kitty. You are so handsome!
	Sentimental	The arena reminded me of my childhood.
	Argumentative	I dislike the way the arena has been arranged
	Cultured	The length of this stadium coincides rather lovely with the width.
	Sweet	It was such a nice day at the game. These fans are the best.
	Romantic	Basking at the game with my love
	Skeptical	So many fireworks, there is no way they set them all off at one
	High-spirited	Those are the most beautiful fireworks I have ever seen!
	Cultured	Fireworks have been used in our celebrations for centuries.
	Arrogant	fireworks are overrated and loud
	Humble	I’m so grateful for whoever invented fireworks!
	Romantic	A charming home that will call you back to days gone by.
	Anxious	This house and this street just makes me feel uneasy.
	Creative	I could write a novel about this beautiful old home!
	Sweet	What a cute little neighborhood!
	Money-minded	Call APR now to get your house renovated!

Table 3: Generative model performance on COCO caption using the test split of (Karpathy & Fei-Fei, 2015)

Method	Image Encoder	BLEU1	BLEU4	ROUGE-L	CIDE _F	SPICE
Adaptive (Lu et al., 2017)	ResNet	74.2	32.5	-	108.5	19.5
Att2in (Rennie et al., 2017)	ResNet	-	33.3	55.3	111.4	-
NBT (Lu et al., 2018)	ResNet	75.5	34.7	-	107.2	20.1
UPDOWN (Anderson et al., 2018)	ResNet FRCNN	79.8	36.3	56.9	120.1	21.4
SHOWTELL (Our)	ResNet152	75.2	31.5	54.2	103.9	18.4
SHOWATTELL (Our)	ResNet152	76.5	32.4	55.1	109.7	19.2
UPDOWN (Our)	ResNet152	77.0	33.9	55.6	112.7	19.6
SHOWTELL (Our)	ResNeXt-IG-3.5B	78.2	35.0	56.6	119.9	20.8
SHOWATTELL (Our)	ResNeXt-IG-3.5B	78.8	35.6	57.1	121.8	20.6
UPDOWN (Our)	ResNeXt-IG-3.5B	79.3	36.4	57.5	124.0	21.2

Note that all scores are lower than for the COCO captioning task. Indeed standard image captioning tries to produce text descriptions that are semantically equivalent to the image, whereas PERSONALITY-CAPTIONS captures how a human responds to a given image when speaking to another human when both can see the image – which is rarely to simply state its contents. Hence, PERSONALITY-CAPTIONS has intrinsically more diverse outputs, similar to results found in other human communication tasks (Liu et al., 2016). For that reason we perform human evaluation in Section 5.3 in addition to automatic evaluations.

Retrieval models Similarly we compare the effect of various configurations of our retrieval model, TransResNet. The models are evaluated in terms of R@1, where for each sample there are 100 candidates to rank: 99 randomly chosen candidates from the test set plus the true label.

Table 6 shows the scores obtained on the test set of PERSONALITY-CAPTIONS. Again, the impact of using the image encoder trained on billions of images is considerable, we obtain 53.5% for our best ResNeXt-IG-3.5B model, and 34.4% for our best Resnet152 model. Conditioning on the personality traits is also very important (53.5% vs. 38.5% R@1 for the best variants with and without conditioning). Transformer text encoders also outperform bag-of-word embeddings encoders,

Table 4: Retrieval model performance on Flickr30k and COCO caption using the splits of (Karpathy & Fei-Fei, 2015). COCO caption performance is measured on the 1k image test split.

Model	Text Pre-training	Flickr30k			COCO		
		R@1	R@5	R@10	R@1	R@5	R@10
UVS (Kiros et al., 2014)	-	23.0	50.7	62.9	43.4	75.7	85.8
Embedding Net (Wang et al., 2018)	-	40.7	69.7	79.2	50.4	79.3	69.4
sm-LSTM (Huang et al., 2016)	-	42.5	71.9	81.5	53.2	83.1	91.5
VSE++ (ResNet, FT) (Faghri et al., 2017)	-	52.9	80.5	87.2	64.6	90.0	95.7
GXN (i2t+t2i) (Gu et al., 2017)	-	56.8	-	89.6	68.5	-	97.9
<i>TransResNet model variants:</i>							
Transformer, ResNet152	Full	10.3	27.3	38.8	21.7	45.6	58.9
Bag of words ResNeXt-IG-3.5B	None	50.0	81.1	90.0	51.6	85.3	93.4
Transformer ResNeXt-IG-3.5B	None	55.6	83.2	90.5	64.0	90.6	96.3
Bag of words ResNeXt-IG-3.5B	Word	58.6	87.2	92.9	54.7	87.1	94.5
Transformer ResNeXt-IG-3.5B	Word	68.4	90.6	95.3	67.3	91.7	96.5

Table 5: Generative model caption performance on the PERSONALITY-CAPTIONS test set.

Method	Image Encoder	Personality Encoder	Personality				
			BLEU1	BLEU4	ROUGE-L	CIDE _r	SPICE
SHOWTELL	ResNet152	Yes	12.4	1.4	13.2	14.5	1.6
SHOWATTTELL	ResNet152	Yes	15.3	1.3	13.1	15.2	3.4
UPDOWN	ResNet152	Yes	15.4	1.4	14.6	16.9	4.9
SHOWTELL	ResNeXt-IG-3.5B	No	15.2	0.9	13.3	14.4	4.6
SHOWATTTELL	ResNeXt-IG-3.5B	No	13.8	0.9	13.1	17.6	5.4
UPDOWN	ResNeXt-IG-3.5B	No	14.3	1.0	13.5	18.0	7.0
SHOWTELL	ResNeXt-IG-3.5B	Yes	14.2	1.2	14.5	15.4	2.2
SHOWATTTELL	ResNeXt-IG-3.5B	Yes	15.0	1.4	14.6	18.8	5.9
UPDOWN	ResNeXt-IG-3.5B	Yes	15.6	1.6	15.0	22.0	7.3

where pretraining for either type of encoder helps. For Transformers pretraining the whole network performed better than just pretraining the word embeddings, see Appendix A.

Example predictions of our best model, TransResNet (ResNeXt-IG-3.5B), are given in Table 2.

5.3 HUMAN EVALUATION ON PERSONALITY-CAPTIONS

The goal of PERSONALITY-CAPTIONS is to be engaging to human readers by emulating human personality traits. We thus test our task and models in a set of human evaluation studies.

Evaluation Setup Using 500 random images from the YFCC-100M dataset that are not present in PERSONALITY-CAPTIONS, we obtain captions for them using a variety of methods, as outlined in the sections below, including both human authored captions and model predicted captions. Using a separate set of human annotators, comparisons are then done pairwise: we show each image, with two captions to compare, to five separate annotators and ask them to choose the “more engaging” caption. For experiments where both captions are conditioned on a personality, we show the annotator the personality; otherwise, the personality is hidden. We then report the percentage of the time one method is chosen over the other. The results are summarized in Figure 3.

Traditional Human Captions We compare human authored PERSONALITY-CAPTIONS captions to human authored traditional neutral (COCO-like) captions. Captions conditioned on a personality were found to be significantly more engaging than those that were neutral captions of the image, with a win rate of 64.5%, which is statistically significant using a binomial two-tailed test.

Human vs. Model Engagingness We compare the best-performing models from Section 5.2 to human authored PERSONALITY-CAPTIONS captions. For each test image we condition both human and model on the same (randomly-chosen) personality trait. Our best TransResNet model from Sec.

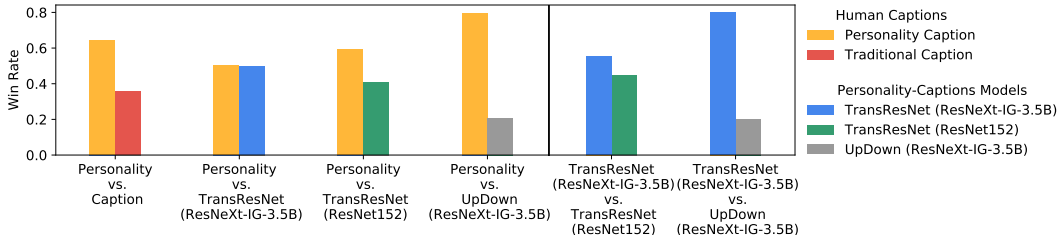


Figure 3: Human evaluations on PERSONALITY-CAPTIONS. Engagingness win rates of various pairwise comparisons: human annotations of PERSONALITY-CAPTIONS vs. traditional captions, vs. PERSONALITY-CAPTIONS model variants, and models compared against each other.

Table 6: Results for TransResNet retrieval variants on the PERSONALITY-CAPTIONS test set.

Text Encoder	Pre-training	Image Encoder	Personality Encoder	R@1
Transformer	Full	ResNet152	No	16.6
Bag of Words	None	ResNet152	Yes	24.2
Transformer	None	ResNet152	Yes	26.8
Bag of Words	Word	ResNet152	Yes	28.5
Transformer	Full	ResNet152	Yes	34.4
Transformer	Full	ResNeXt-IG-3.5B	No	38.5
Bag of Words	None	ResNeXt-IG-3.5B	Yes	38.6
Transformer	None	ResNeXt-IG-3.5B	Yes	42.9
Bag of Words	Word	ResNeXt-IG-3.5B	Yes	45.7
Transformer	Full	ResNeXt-IG-3.5B	Yes	53.5

5.2, using the ResNext-IG-3.5B image features, almost matched human authors, with a win rate of 49.5% (difference not significant, $p > 0.6$). The same model using ResNet152 has a win rate of 40.9%, showing the importance of strongly performing image features. The best generative model we tried, the UPDOWN model using ResNext-IG-3.5B image features, performed worse with a win rate of 20.7%, showing the impact of retrieval for engagement.

Model vs. Model engagingness We also compare our models in a pairwise fashion directly, as measured by human annotators. The results given in Figure 3 (all statistically significant) show the same trends as we observed before: TransResNet with ResNext-IG-3.5B outperforms the same model with ResNet152 features with a win rate of 55.2%, showing the importance of image features. Additionally, TransResNet with ResNext-IG-3.5B image features (with no pretraining) also substantially outperforms the UPDOWN model using ResNext-IG-3.5B with a winrate of 80.1%.

6 CONCLUSION

In this work we consider models that can simultaneously understand image content and provide engaging captions for humans. To build strong models, we first leverage the latest advances in image and sentence encoding to create generative and retrieval models that perform well on standard image captioning tasks. In particular, we attain a new state-of-the-art on caption generation on COCO, and introduce a new retrieval architecture, TransResNet, that yields the highest known hits@1 score on the Flickr30k dataset.

To make the models more engaging to humans, we then condition them on a set of controllable personality traits. To that end, we collect a large dataset, PERSONALITY-CAPTIONS to train such models. Using automatic metrics and human evaluations, we show that our best system is able to produce captions that are close to matching human performance in terms of engagement. Our benchmark will be made publicly available to encourage further model development, leaving the possibility of superhuman performance coming soon in this domain.

REFERENCES

- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pp. 382–398. Springer, 2016.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. *CVPR*, 2018.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.
- Arjun Chandrasekaran, Devi Parikh, and Mohit Bansal. Punny captions: Witty wordplay in image descriptions. *arXiv preprint arXiv:1704.08224*, 2017.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017.
- Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. User conditional hashtag prediction for images. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1731–1740. ACM, 2015.
- Aviv Eisenschat and Lior Wolf. Capturing deep correlations with 2-way nets. *CoRR*, abs/1608.07973, 2016. URL <http://arxiv.org/abs/1608.07973>.
- Martin Engilberge, Louis Chevallier, Patrick Prez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- Fartash Faghri, David J. Fleet, Ryan Kiros, and Sanja Fidler. VSE++: improved visual-semantic embeddings. *CoRR*, abs/1707.05612, 2017. URL <http://arxiv.org/abs/1707.05612>.
- Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proc IEEE Conf on Computer Vision and Pattern Recognition*, pp. 3137–3146, 2017.
- Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. *CoRR*, abs/1711.06420, 2017. URL <http://arxiv.org/abs/1711.06420>.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *arXiv preprint arXiv:1802.08218*, 2018.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL <http://arxiv.org/abs/1512.03385>.
- Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. What we instagram: A first analysis of instagram photo content and user types. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.
- Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal LSTM. *CoRR*, abs/1611.05588, 2016. URL <http://arxiv.org/abs/1611.05588>.

- Timothy Jay and Kristin Janschewitz. Filling the emotion gap in linguistic theory: Commentary on potts' expressive dimension. *Theoretical Linguistics*, 33(2):215–221, 2007.
- Jonczyk and Rafał Jończyk. *Affect-language interactions in native and non-native English speakers*. Springer, 2016.
- Onno Kampman, Farhad Bin Siddique, Yang Yang, and Pascale Fung. Adapting a virtual agent to user personality. In *Advanced Social Interaction with Agents*, pp. 111–118. Springer, 2019.
- Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.
- Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. URL <http://arxiv.org/abs/1411.2539>.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.
- Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6, pp. 2, 2017.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7219–7228, 2018.
- L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal convolutional neural networks for matching image and sentence. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2623–2631, Dec 2015. doi: 10.1109/ICCV.2015.301.
- Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. *CoRR*, abs/1805.00932, 2018. URL <http://arxiv.org/abs/1805.00932>.
- Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1485–1488, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1874254. URL <http://doi.acm.org/10.1145/1873951.1874254>.
- Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8591–8600, 2018.
- Alexander Patrick Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *AAAI*, pp. 3574–3580, 2016.
- P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes. Training Millions of Personalized Dialogue Agents. *ArXiv e-prints*, September 2018.
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spathourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *CoRR*, abs/1701.08251, 2017. URL <http://arxiv.org/abs/1701.08251>.

- Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. *CoRR*, abs/1611.00471, 2016. URL <http://arxiv.org/abs/1611.00471>.
- Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1899–1907, Oct 2017. doi: 10.1109/ICCV.2017.208.
- Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, and Christos Faloutsos. Automatic image captioning. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 3, pp. 1987–1990. IEEE, 2004.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.
- Chunseong Cesc Park, Byeongchan Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. 2017.
- Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, volume 1, pp. 3, 2017.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL <http://arxiv.org/abs/1409.0575>.
- Matthias Scheutz, Paul Schermerhorn, and James Kramer. The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 226–233. ACM, 2006.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.
- Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *CoRR*, abs/1511.06361, 2015. URL <http://arxiv.org/abs/1511.06361>.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.
- L. Wang, Y. Li, J. Huang, and S. Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2797921.
- Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5005–5013, 2016.
- Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016. URL <http://arxiv.org/abs/1611.05431>.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.
- Kota Yoshida, Munetaka Minoguchi, Kenichiro Wani, Akio Nakamura, and Hirokatsu Kataoka. Neural joking machine: Humorous image captioning. *arXiv preprint arXiv:1805.11850*, 2018.

Quanzeng You, Hailin Jin, and Jiebo Luo. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv preprint arXiv:1801.10121*, 2018.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

A IMPACT OF PRETRAINED WORD EMBEDDINGS AND TEXT ENCODERS

Table 7: More detailed results for retrieval model performance on COCO Captions using the splits of (Karpathy & Fei-Fei, 2015). For our TransResNet models, we compare two types of pretraining: Full indicates a model with a pretrained text encoder, while Word indicates a model with pretrained word embeddings only.

Model	Text Encoder Pretraining	Caption retrieval			
		R@1	R@5	R@10	Med Rank
1k Images					
m-CNN (Ma et al., 2015)		42.8	-	84.1	2.0
UVS (Kiros et al., 2014)		43.4	75.7	85.8	2.0
HM-LSTM (Niu et al., 2017)		43.9	-	87.8	2.0
Order Embeddings (Vendrov et al., 2015)		46.7	-	88.9	2.0
Embedding Net (Wang et al., 2018)		50.4	79.3	69.4	-
DSPE+Fisher Vector (Wang et al., 2016)		50.1	-	89.2	-
sm-LSTM (Huang et al., 2016)		53.2	83.1	91.5	1.0
VSE++ (ResNet, FT) (Faghri et al., 2017)		64.6	90.0	95.7	1.0
GXN (i2t+t2i) (Gu et al., 2017)		68.5	-	97.9	1.0
Engilberge et al. (2018)		69.8	91.9	96.6	1.0
Transformer [†] , Resnet152	Word	21.7	45.6	58.9	7.0
Bag of words, ResNeXt-IG-3.5B	None	51.6	85.3	93.4	1.4
Bag of words [†] , ResNeXt-IG-3.5B	Word	54.7	87.1	94.5	1.0
Transformer, ResNeXt-IG-3.5B	None	63.4	90.6	96.3	1.0
Transformer [†] , ResNeXt-IG-3.5B	Word	66.6	90.6	96.3	1.0
Transformer*, ResNeXt-IG-3.5B	Full	67.3	91.7	96.5	1.0
5k Images					
Order Embeddings (Vendrov et al., 2015)		23.3	-	65.0	5.0
VSE++ (ResNet, FT) (Faghri et al., 2017)		41.3	71.1	81.2	2.0
GXN (i2t+t2i) (Gu et al., 2017)		42.0	-	84.7	2.0
Transformer, Resnet152	Word	7.8	21.9	31.2	30.0
Bag of words, ResNeXt-IG-3.5B	None	26.6	58.6	73.0	4.0
Bag of words, ResNeXt-IG-3.5B	Word	29.7	62.9	75.7	3.0
Transformer, ResNeXt-IG-3.5B	None	38.8	71.6	82.7	2.0
Transformer, ResNeXt-IG-3.5B	Word	44	73.7	84	2.0
Transformer, ResNeXt-IG-3.5B	Full	44.3	74.5	83.9	2.0

Table 8: Retrieval model performance on Flickr30k using the splits of (Karpathy & Fei-Fei, 2015). For our models, we compare two types of pretraining: Full indicates a model with a pretrained text encoder, while Word indicates a model with pretrained word embeddings only.

Model	Text Encoder Pretraining	Caption retrieval			
		R@1	R@5	R@10	Med Rank
UVS (Kiros et al., 2014)		23.0	50.7	62.9	5.0
UVS (Github)		29.8	58.4	70.5	4.0
Embedding Net (Wang et al., 2018)		40.7	69.7	79.2	-
DAN (Nam et al., 2016)		41.4	73.5	82.5	2.0
sm-LSTM (Huang et al., 2016)		42.5	71.9	81.5	2.0
2WayNet (Eisenschat & Wolf, 2016)		49.8	67.5	-	-
VSE++ (ResNet, FT) (Faghri et al., 2017)		52.9	80.5	87.2	1.0
DAN (ResNet) (Nam et al., 2016)		55.0	81.8	89.0	1.0
GXN (i2t+t2i) (Gu et al., 2017)		56.8	-	89.6	1.0
Transformer, Resnet152	Word	10.3	27.3	38.8	19
Bag of words, ResNeXt-IG-3.5B	None	50.0	81.1	90.0	1.5
Transformer, ResNeXt-IG-3.5B	None	55.6	83.2	90.5	1.0
Bag of words, ResNeXt-IG-3.5B	Word	58.6	87.2	92.9	1.0
Transformer, ResNeXt-IG-3.5B	Full	62.3	88.5	94.4	1.0
Transformer, ResNeXt-IG-3.5B	Word	68.4	90.6	95.3	1.0

Table 9: Comparing Generative model caption performance on the PERSONALITY-CAPTIONS test set: pretrained word embeddings vs. no pretraining. Pretraining makes a very small impact in this case, unlike in our retrieval models.

Method	Image Encoder	Personality Encoder	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
<i>no pretraining:</i>							
SHOWTELL	ResNeXt-IG-3.5B	Yes	14.2	1.2	14.5	15.4	2.2
SHOWATTTELL	ResNeXt-IG-3.5B	Yes	15.0	1.4	14.6	18.8	5.9
UPDOWN	ResNeXt-IG-3.5B	Yes	15.6	1.6	15.0	22.0	7.3
<i>with word embedding pretraining:</i>							
SHOWTELL [†]	ResNeXt-IG-3.5B	Yes	15.6	1.4	14.7	17.0	3.0
SHOWATTTELL [†]	ResNeXt-IG-3.5B	Yes	15.0	1.5	14.9	18.5	4.8
UPDOWN [†]	ResNeXt-IG-3.5B	Yes	16.4	1.6	15.5	21.5	7.5

Table 10: Retrieval model performance on PERSONALITY-CAPTIONS. We compare two types of pretraining: Full indicates a model with a pretrained text encoder, while Word indicates a model with pretrained word embeddings only.

Text Encoder		Image Encoder	Personality Encoder	R@1
Encoder Type	Pretraining			
Transformer	Full	ResNeXt-IG-3.5B	Yes	53.5
Transformer	Word	ResNeXt-IG-3.5B	Yes	48.6
Bag of Words	Word	ResNeXt-IG-3.5B	Yes	45.7
Transformer	None	ResNeXt-IG-3.5B	Yes	42.9
Bag of Words	None	ResNeXt-IG-3.5B	Yes	38.6
Transformer	Full	ResNeXt-IG-3.5B	No	38.5
Transformer	Full	Resnet152	Yes	34.4
Transformer	Word	Resnet152	Yes	30.2
Bag of Words	Word	Resnet152	Yes	28.5
Transformer	None	Resnet152	Yes	26.8
Bag of Words	None	Resnet152	Yes	24.2
Transformer	Full	Resnet152	No	16.6

B ENGAGING CAPTIONS, WITH NO PERSONALITY CONDITIONING

Engaging-only Captions Instead of asking to author a caption based on a personality trait, we can ask humans to simply write an “engaging” caption instead, providing them with no personality cue. We found that human annotators overall preferred captions written by those unconditioned on a personality by a slight margin ($\sim 54\%$). To further understand this difference, we split the images into three subsets based on the personality on which the PERSONALITY-CAPTIONS annotator conditioned their caption, i.e. whether the personality was positive, negative, or neutral. We then examined the engagingness rates of images for each of these subsets. In the set where PERSONALITY-CAPTIONS annotators were provided with positive personalities, which totaled 185 out of the 500 images, we found that human annotators preferred the captions conditioned on the personality to those that were not. However, in the other two sets, we found that the unconditioned captions were preferred to the negative or neutral ones. For these two subsets, we believe that, without the context of any personality, annotators may have preferred the inherently more positive caption provided by someone who was asked to be engaging but was not conditioned on a personality.

Table 11: Pairwise win rates of various approaches, evaluated in terms of engagingness

Type of caption A	WIN PERCENTAGE		Type of caption B
Human (all) personality captions	45.5	54.5	Human engaging captions
Human (positive) personality captions	51.2	48.8	Human engaging captions

Diversity of captions We found that the captions written via our method were not only more engaging for positive personality traits, but also resulted in more diversity in terms of personality traits. To measure this diversity, we constructed a model that predicted the personality of a given comment. The classifier consists in the same Transformer as described in 4.3, pre-trained on the same large dialog corpus, followed by a softmax over 215 units. We then compare the total number of personality types as predicted by the classifier among each type of human-labeled data: “engaging” captions conditioned on personalities, “engaging” captions not conditioned on personalities, and traditional image captions. That is, we look at each caption given by the human annotators, assign it a personality via the classifier, and then look at the total set of personalities we have at the end for each set of human-labeled data. For example, out of the 500 human-generated traditional captions, the classifier found 63% of all possible positive personalities in this set of captions. As indicated in Table 12, the human annotators who were assigned a personality produce more diverse captions, particularly negatively and neutrally conditioned ones, as compared to human annotators who are just told to be “engaging” or those who are told to write an image caption.

Table 12: Caption diversity in human annotation tasks. PERSONALITY-CAPTIONS provides more diverse personality traits than traditional captions or collecting engaging captions without specifying a personality trait to the annotator, as measured by a personality trait classifier.

Annotation Task	Personality Trait Coverage		
	Positive	Neutral	Negative
Given Personalities	100%	100%	99.0%
Traditional Caption	63.0%	83.3%	47.0%
Engaging, No Conditioning	81.5%	91.7%	71.4%
PERSONALITY-CAPTIONS	82.7%	94.4%	87.8%

C COMPARING GENERATIVE AND RETRIEVAL MODELS ON COCO

The ultimate test of our generative and retrieval models on PERSONALITY-CAPTIONS is performed using human evaluations. Comparing them using automatic metrics is typically difficult because retrieval methods perform well with ranking metrics they are optimized for and generative models perform well with word overlap metrics they are optimized for, but neither of these necessarily correlate with human judgements, see e.g. Zhang et al. (2018).

Nevertheless, here we compare our generative and retrieval models directly with automatic metrics on COCO. We computed the BLEU, CIDEr, SPICE, and ROUGE-L scores for our best TransResNet model. The comparison is given in Table 13.

Table 13: Generative and retrieval model performance on COCO caption using the test split of (Karpathy & Fei-Fei, 2015). All models use ResNeXt-IG-3.5B image features.

Model	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
TransResNet	50.6	10.9	38.0	49.1	13.9
SHOWTELL	78.2	35.0	56.6	119.9	20.8
SHOWATTTELL	78.8	35.6	57.1	121.8	20.6
UPDOWN	79.3	36.4	57.5	124.0	21.2

D HUMAN ANNOTATION SETUP

Comment on an Image

Description

In this task, you will be shown 5 images, and will write a comment about each image. The goal of this task is to write something about an image that someone else would find engaging.

STEP 1


With each new photo, you will be given a **personality trait** that you will try to emulate in your comment. For example, you might be given "snarky" or "glamorous". The personality describes **YOU**, not the picture. It is *you* who is snarky or glamorous, not the contents of the image.

STEP 2

You will then be shown an image, for which you will write a comment *in the context of your given personality trait*. Please make sure your comment has at least **three words**. Note that these are *comments*, not captions.

E.g., you may be shown an image of a tree. If you are "snarky", you might write "What a boring tree, I bet it has bad wood;" or, if you were "glamorous", you might write "What an absolutely beautiful tree! I would put this in my living room it's so extravagant!"

Image



Your assigned personality is:

Adventurous

Reminder - please do not write anything that involves any level of discrimination, racism, sexism and offensive religious/politics comments, otherwise the submission will be rejected.

Instructions for the annotation task collecting the data for PERSONALITY-CAPTIONS.

E SAMPLES FROM PERSONALITY-CAPTIONS

Table 14: Some samples from PERSONALITY-CAPTIONS. For each sample we asked a person to write a caption that fits both the image and the personality.



Sarcastic
Yes please sit by me



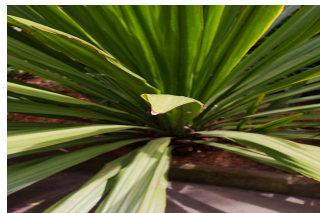
Mellow
Look at that smooth easy catch of the ball. like ballet.



Zany
I wish I could just run down this shore!



Contradictory
Love what you did with the place!



Contemptible
I can't believe no one has been taking care of this plant. Terrible



Energetic
About to play the best tune you've ever heard in your life. Get ready!



Kind
they left me a parking spot



Spirited
That is one motor cycle enthusiast!!!



Creative
Falck alarm, everyone. Just a Falck alarm.



Crazy
I drove down this road backwards at 90 miles per hour three times



Morbid
I hope this car doesn't get into a wreck.



Questioning
Why do people think its cool to smoke cigarettes?

F EXAMPLES FROM HUMAN EVALUATION SET

Image and Pers.	Use pers.	Captioning	Caption
 Spirited	No	Standard	A city on the background, a lake on the front, during a sunset.
	No	Engaging	Talk about summer fun! Can I join? :)
	Yes	Human	i feel moved by the sunset
	Yes	TransResNet	The water at night is a beautiful sight.
	Yes	UPDOWN	This is a beautiful sunset!
 Ridiculous	No	Standard	Rose colored soft yarn.
	No	Engaging	I really want to untangle that yarn.
	Yes	Human	I cannot believe how yummy that looks.
	Yes	TransResNet	What is up with all the knitting on my feed
	Yes	UPDOWN	I would love to be a of that fruit!
 Maternal	No	Standard	A beautiful mesa town built into the cliffs.
	No	Engaging	That is a strange cave
	Yes	Human	It must be very dangerous if children play there
	Yes	TransResNet	I hope my kids don't climb on this.
	Yes	UPDOWN	I hope this is a beautiful place.
 Sophisticated	No	Standard	Hockey players competing for control of the hockey puck.
	No	Engaging	Great save, goalie!!
	Yes	Human	Hockey is a little too barbaric for my taste.
	Yes	TransResNet	Hockey players gracefully skate across the ice.
	Yes	UPDOWN	This hockey is like they are a great of the game.
 Curious	No	Standard	two people walking through a snowy forest.
	No	Engaging	Too cold for me.
	Yes	Human	I wonder what's at the finish line for these guys?
	Yes	TransResNet	I wonder why they are running.
	Yes	UPDOWN	I wonder what they are a?
 Happy	No	Standard	Hollywood Tower at Night
	No	Engaging	I went to that theme park, but was too scared to get on that ride!
	Yes	Human	I am so excited to be here!
	Yes	TransResNet	I remember going to disney world, it was one of the best trips I've ever done.
	Yes	UPDOWN	This looks like a beautiful view!

Table 15: Example variants of the captions shown to human annotators in the human evaluation tasks in Section 5.3. The first two captions are human annotations not conditioned on a personality; the next three are captions conditioned on the listed personality, and are generated via a human annotator, TransResNet, and UPDOWN respectively.

G MORE EXAMPLES FROM TRANSRESNET

Image	Personality	Generated comment
	Sweet Vague Cultured Paranoid Overimaginative	I love, love, love these chairs! I want the big one in my house! This chair is either covered in snow or the snow is covered in the chair. These chairs remind me of the Swedish interior design revolution of the 70's. What if someone fell off those chairs. Those chairs look like they could be in a doll house.
	Arrogant Overimaginative Vague Optimistic Charming	I've seen better sunsets elsewhere. that sunset is so orange it could be a fruit It's the sunset. The sunset makes look forward to a happy tomorrow. The way the sun is hitting the water makes for a romantic evening.
	Sweet Skeptical Sympathetic Vague Wishful	What a cute puppy, reminds me of my friends. I don't think this dog will bite me. poor dog! It looks so hungry :c it's a dog I wish that I had a dog as cute as him.
	Cultured Skeptical Sweet Overimaginative Sympathetic	I love a cultural celebration. I'm not sure if these are guys in costumes or time travelers. I love that they are celebrating their traditions and culture. They look like they could be dancers in a fantasy movie with dragons! I feel sorry for him having to wear that
	Romantic Humble Paranoid Creative Money-minded	If I was an insect, I would definitely make this my mate. I am grateful that spiders eat these disgusting bugs. What is going on? Are these insects dangerous? I made something like this from colored toothpicks once how much are those? those looks expensive
	Happy Optimistic Critical Charming Adventurous	That is so cool! I I love street art! The future is bright for people who can dream in artistic ways. I do believe this taggers verbage is a tad juvenile What a charming wall. I think I could create art like that, I will go learn and take action.
	Dramatic Wishful Sweet Romantic Happy	The color of this flower is absolutely astounding. I can't believe it. I always wish I could grow these types of flowers. Beautiful flowers! I would give them to you. The pink flowers would make a beautiful bouquet for my wife. Oh my, what a lovely purple color of nature's new sprouts!

Table 16: More example predictions from our best TRANSRESNET model on the PERSONALITY-CAPTIONS validation set.

Image	Personality	Generated comment
	Adventurous Vague Charming Optimistic Paranoid	This biking event looks like something that I would try! Those people are riding a bike. I bet a wonderful couple uses this bike to tour the countryside together. A hopeful cyclist trying to catch up to the pack What if all those bikes just tipped over!
	Adventurous Cultured Vague Dramatic Sympathetic	I am so ready for the conference. This conference is one of the most important ones in the country. The organization on that table is uncertain. OMG!! This ceremony is frightening! I feel bad for these people being so cramped in this room.
	Old-fashioned Charming Argumentative Anxious Dramatic	Such old fashioned script, a true lost art. I could use these to write to my loved ones. Can you even read this through all the jpeg artifacts? I hope this paper doesnt tear, history will be destroyed. Some of the most profound things ever written have been on linen.
	Happy Wishful Boyish Romantic Cultured	It finally snowed, it makes me feel awesome I wish there was enough for snow angels. Can I go sledding now? What a beautiful frost! Looks like the perfect place to fall in love! The white of the snow provides a glistening contrast to the dead trees.
	Wishful Money-minded Critical Humble Paranoid	I wish I could have a life as easy as a plant. This plant is probably worth a lot of money the leaf is ruining the picture This plant is a symbol of life in humble opinion. Just gorgeous! If you eat this leaf it definetly will not poison you. Or will it...
	Romantic Boyish Creative Sweet Money-minded	This valentine concert is for lovers. It's always fun to get down and jam with the boys! musician performing a song of theirs oh what lovely young musicians I wonder how much the musicians have in student loan debt.
	Skeptical Paranoid Happy Arrogant Humble	I wonder why the ships are all parked further down the deck. I hope those ships don't sink Look how beautiful the port is at this time of day! :) Those boats don't need to be docked at this time of night We are so lucky to have these boats available locally

Table 17: More example predictions from our best TRANSRESNET model on the PERSONALITY-CAPTIONS validation set.