

Engaging Image Captioning via Personality

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, Jason Weston

Facebook AI Research

{kshuster, samueulhumeau, hexianghu, abordes, jase}@fb.com

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 241,858 of such captions conditioned over 215 possible traits. We build models that combine existing work from (i) sentence representations [36] with Transformers trained on 1.7 billion dialogue examples; and (ii) image representations [32] 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 by spanning both the ability to understand and to be engaging. For agents to communicate the way people do, they must display personality as well as perform conversational function [21, 22, 45, 23]. Consider for example an online conversational agent or robot that can both perceive images and speak – the aforementioned capabilities would be expected from a good conversationalist.

Communication grounded in images is naturally engaging to humans [18], for example billions are shared and discussed daily online. In order to develop engaging conversational agents, it thus seems promising to allow them to comment on images naturally as humans do. Yet the majority of studies in the research community have so far focused on function only: standard image captioning [40] requires the machine to generate a sentence which factually describes

the elements of the scene in a neutral tone. Similarly, visual question answering [4] and visual dialogue [9] 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 [16].

Standard image captioning tasks simply state the obvious, and are not considered engaging captions by humans. For example, in the COCO [8] and Flickr30k [57] 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 vision research.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 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 241,858 captions, each conditioned on one of 215 different possible personality traits. We show that such captions are far more engaging 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 generative³ variants, and leverage state-of-the-

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

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

³“Generative” here refers to a model that generates a caption word-by-word as opposed to a retrieval model.

art modules from both the vision and language domains. For image representations, we employ the work of [32] 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 [36] trained on 1.7 billion dialogue examples. Our generative model gives a new state-of-the-art on COCO caption generation, and our retrieval architecture, TransResNet, yields the highest known R@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, see Figure 1. In particular, when compared to human captions, annotators preferred our retrieval model’s captions over human ones 49.5% of the time – very close to human performance. Our task is however a challenge for generative models which succeed on COCO, but fail on our task. We believe future work should address this important open problem.

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 [8] and Flickr30k [57] datasets, comparing to a range of models, including both generative models such as in [50, 54, 3] and retrieval based such as in [15, 13, 38]. 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 [10] or knowledge of the reader’s active vocabulary [42]. Our work does not address this issue. Another research direction is to attempt to produce amusing captions either through wordplay (puns) [7] or training on data from humour websites [55]. 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. [34]. Others such as [56] have used small datasets like SentiCap [35] with ~800 images to inject sentiment into captions. [14] collect a somewhat bigger dataset with 10,000 images, 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 ~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 [37] or dialogue in general where displaying personality is important [58]. In those tasks, simple word overlap based automatic metrics are shown to perform weakly [28] 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 [32] which provides state-of-the-art performance on ImagenNet 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 [47]; in particular, their use in retrieval models for dialogue by large-scale pretraining [36] 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, publicly available at http://parl.ai/projects/personality_captions.

Personality traits A large number of studies are dedicated to producing a model of the personality of an individual [20], such as the Big-Five [1], the Big-Two [1] and 16PF among others [6]. Those models usually project personality in a low dimension space, for instance the Big-Five describes a personality by weighting openness to experience, conscientiousness, extraversion, agreeableness and neuroticism. However such a description is not well adapted to a crowdsourced data collection task, where labelers are not familiar with those models. We found it clearer to use a single descriptor as a “personality trait” (e.g. “sweet”, “skeptical”, “solemn”, etc.). We considered 215 possible personality traits which were constructed by selecting a subset from a curated list of 638 traits⁴ 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). 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.

Data collection We use a randomly selected set of the images from the YFCC100M Dataset⁵ to build our train-

⁴<http://ideonomy.mit.edu/essays/traits.html>

⁵<https://multimediacommons.wordpress.com/yfcc100m-core-dataset/>; [46]

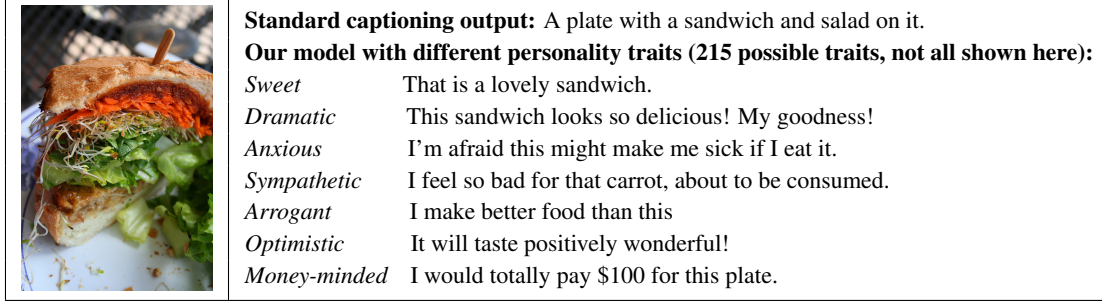


Figure 1: Our TransResNet model compared to a standard image captioning model on the same image conditioned on various personality traits. Our model is trained on the new PERSONALITY-CAPTIONS dataset which covers 215 different personality traits. The standard captioning system used for comparison is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

Type	Datasets With Personality				Datasets Without Personality			
Dataset	Personality-Captions			FlickrStyle10K	COCO		Flickr30k	
Split	train	valid	test	train	train	valid	train	valid
Number of Images	186,858	5,000	10,000	7000	82783	40504	29000	1014
Number of Captions	186,858	5,000	50,000	14000	414113	202654	145000	5070
Number of Personality Types	215	215	215	2	None	None	None	None
Vocabulary Size	33641	5460	16655	8889	23776	17724	17920	4283
Average Tokens per Caption	11.2	10.9	11.1	14.51	11.3	11.3	13.53	13.74

Table 1: PERSONALITY-CAPTIONS dataset statistics compared to other captioning datasets.

ing, validation and test sets, selecting for each chosen image a random personality trait, drawn uniformly from our list. The captions are written by a large number of crowdworkers, with the annotation task distributed among them. Test examples have 5 captions per image in order to compute multi-reference automatic evaluations such as BLEU.

In each annotation round, an annotator is shown an image along with a trait. The annotators are then asked to write an engaging utterance for the image in the context of the personality trait. Specifically, they are told to “write a comment *in the context of your given personality trait*... about an image that someone else would find engaging”. Note we do not use the word “caption” in these instructions because we felt it would be clearer to crowdworkers of our intent: not many humans have experience writing captions and they may misinterpret the word to mean a factual natural statement, whereas they have experience writing personality-based engaging comments. We thus aim to illicit more natural utterances that humans are used to writing. In this paper we refer to these labels as PERSONALITY-CAPTIONS.

The captions are constrained to include at least three words. It was emphasized that the personality trait describes a trait of the author of the caption, not properties of the content of the image. They were also instructed not to use the personality trait word itself in their caption. For quality control, crowdworkers were manually monitored and removed for poor performance. See Figure 3 in the appendix for more details of the exact instructions given to annotators.

The final dataset statistics are given in Table 1 and compared to the largest dataset we are aware of that also has personality based captions, FlickrStyle10k, which is significantly smaller in terms of images, examples and number of personalities. We also show standard captioning datasets COCO and Flickr30k for reference.

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 [17] trained on Imagenet [44] 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 [33]. The second is a ResNeXt $32 \times 48d$ [53] trained on 3.5 billion Instagram pictures following the procedure described by [32], 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 image captioning methods: SHOWTELL [50], SHOWATTTELL [54] and UPDOWN [3].

Image and Personality Encoders The image representation r_I is extracted using the aforementioned image encoder. For the SHOWTELL model, the 2048-dimensional outputs of image encoder is used. For the SHOWATTTELL and UPDOWN models, we keep the spatial extent and use the $7 \times 7 \times 2048$ dimensional outputs of image encoder. In all cases, the image features are ultimately reduced to a vector of dimension 512. In the SHOWTELL model, a linear projection is applied to do so. In both the SHOWATTTELL and UPDOWN models, the image features are first linearly reduced to a tensor of $7 \times 7 \times 512$ dimensions with a 1×1 convolution layer. Then the attention mechanism is used to weighted combine image features along its 7×7 spatial extent, into a vector of dimension 512. In the cases where personality traits are used, each personality trait is embedded by a vector of dimension 512, akin to a word embedding, giving a 215×512 matrix of weights to learn for PERSONALITY-CAPTIONS. The personality embedding is then input to the LSTM caption decoders, through concatenating with the input word vectors at each decoding step.

Caption Decoders In SHOWTELL, similar to [50], the dimensionality reduced image features are used as the first input word to a LSTM model to generate the output caption sequence. In SHOWATTTELL, while the overall architecture is similar to [54], we adopt the modification suggested by [43] and input the attention-derived image features to the cell node of the LSTM. Finally, we use the UPDOWN model exactly as described in [3]. The key difference to SHOWATTTELL is that two LSTM instead of one are used, of which one is responsible for generating the attention weight and the other is responsible of generating the caption. In all above models, the word vector of the previously predicted word (concatenated with personality embedding when applicable) is input to the LSTM caption decoder to predict the current word, at each caption decoding step.

Training and Inference We perform a two-stage training strategy to train such caption generation models as proposed

by [43]. 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 obtain its representation $r_P \in \mathbb{R}^{500}$. 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 [47], followed by a two layer perceptron. We consider a Transformer architecture with 4 layers, 300 hidden units and 6 attention heads. We either train from scratch, pretrain only the word embeddings, i.e. where we initialize word vectors trained using fastText [5] trained on Wikipedia, or pretrain the entire encoder. For the latter, we follow the setup described in [36]: 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, $r_C \in \mathbb{R}^{300}$ is the sum of the 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 the following dot product: $s(I, P, C) = (r_I + r_P) \cdot r_C$.

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

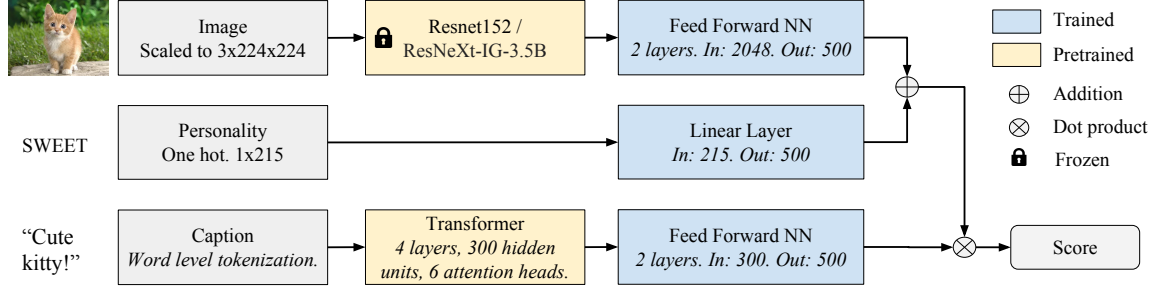


Figure 2: Our architecture TransResNet, used for our retrieval models.

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 both engagingness and fit.

5.1. Automatic evaluation on Traditional Captions

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, ResNet152 and ResNeXt-IG-3.5B. We report performance on the COCO caption dataset [27]. We evaluate BLEU [41], ROUGE-L [26], CIDEr [48] and SPICE [2] and compare models’ performances to state-of-the-art models under the setting of [24]. We provide additional ablations in Appendix C.

The results are shown in Table 2. 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 [32]. More importantly, our best model (UPDOWN) either outperforms or is competitive with state-of-the-art single model performance [3] 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 ([24]). The results are given in Table 3 (for more details, see Tables 9 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 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 4 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 16.5, showing the importance of modeling personality in our new task.

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. PERSONALITY-CAPTIONS has intrinsically more diverse outputs, similar to results found in other human communication tasks [28]. Besides, as in COCO [8], measures like BLEU do not correlate well with human judgements (see top row in Tables 2 and 4) hence we perform human evaluation of our models in Section 5.3.

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 500 candidates to rank: 495 randomly chosen candidates from the test set plus the true labels.

Table 5 shows the scores obtained on the test set of PERSONALITY-CAPTIONS. Again, the impact of using the

Method	Image Encoder	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
Human	-	66.3	21.7	48.4	85.4	19.8
Adaptive [29]	ResNet	74.2	32.5	-	108.5	19.5
Att2in [43]	ResNet	-	33.3	55.3	111.4	-
NBT [30]	ResNet	75.5	34.7	-	107.2	20.1
UPDOWN [3]	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
SHOWATTTELL (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
SHOWATTTELL (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

Table 2: Generative model performance on COCO caption using the test split of [24]

Model	Text Pre-training	Flickr30k			COCO		
		R@1	R@5	R@10	R@1	R@5	R@10
UVS [25]	-	23.0	50.7	62.9	43.4	75.7	85.8
Embedding Net [51]	-	40.7	69.7	79.2	50.4	79.3	69.4
sm-LSTM [19]	-	42.5	71.9	81.5	53.2	83.1	91.5
VSE++ (ResNet, FT) [13]	-	52.9	80.5	87.2	64.6	90.0	95.7
GXN (i2t+t2i) [15]	-	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 3: Retrieval model performance on Flickr30k and COCO caption using the splits of [24]. COCO caption performance is measured on the 1k image test split.

image encoder trained on billions of images is considerable, we obtain 77.5% for our best ResNeXt-IG-3.5B model, and 51.7% for our best Resnet152 model. Conditioning on the personality traits is also very important (77.5% vs. 53.9% R@1 for the best variants with and without conditioning). Transformer text encoders also outperform bag-of-word embeddings encoders, 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 6.

5.3. Human Evaluation on Personality-Captions

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

Engagingness 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 below, including both

human authored captions and model predicted captions. Using a large separate set of human crowdworkers, comparisons are then done pairwise: we show each image, with two captions to compare, to five separate annotators and give them the instruction: “The goal of this task is to pick which comment is the most engaging (interesting, captivating, attention-grabbing)”. This results in 2500 trials in total for each pairwise comparison test. 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 given in Table 7.

Traditional Human Captions We also collected traditional neutral (COCO-like) captions for our 500 test images. Specifically, the instructions were “You will be shown an image, for which you will provide a caption” with the example “E.g, if you are shown an image of a snow-covered tree in a park, you could write *A tree in a park, covered with snow*”. We then compared human authored PERSONALITY-CAPTIONS captions to these neutral captions. Captions conditioned on a personality were found to be significantly more engaging than the neutral captions, with a win rate

Method	Image Encoder	Personality	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
Human Baseline	-	Yes	30.1	2.8	20.1	10.8	5.1
SHOWTELL	ResNet152	No	35.6	3.6	21.5	6.0	2.2
SHOWATTTELL	ResNet152	No	37.8	4.5	23.2	9.3	3.3
UPDOWN	ResNet152	No	36.8	4.1	22.8	8.8	3.2
SHOWTELL	ResNet152	Yes	39.7	7.2	25.0	9.6	1.8
SHOWATTTELL	ResNet152	Yes	42.7	7.2	26.8	12.4	3.8
UPDOWN	ResNet152	Yes	43.9	8.0	27.3	13.6	3.9
SHOWTELL	ResNeXt-IG-3.5B	No	36.5	4.5	22.2	7.8	2.4
SHOWATTTELL	ResNeXt-IG-3.5B	No	38.5	4.9	23.5	11.4	4.0
UPDOWN	ResNeXt-IG-3.5B	No	38.9	4.8	23.5	12.0	4.1
SHOWTELL	ResNeXt-IG-3.5B	Yes	38.4	7.3	24.3	9.6	1.6
SHOWATTTELL	ResNeXt-IG-3.5B	Yes	43.3	7.1	27.0	12.6	3.6
UPDOWN	ResNeXt-IG-3.5B	Yes	44.0	8.0	27.4	16.5	5.2

Table 4: Generative model caption performance on the PERSONALITY-CAPTIONS test set.

Text Encoder	Pre-training	Image Encoder	Personality Encoder	R@1
Transformer	None	None	Yes	20.0
Transformer	Full	None	Yes	25.8
Transformer	Full	ResNet152	No	18.7
Bag of Words	None	ResNet152	Yes	35.4
Bag of Words	Word	ResNet152	Yes	40.5
Transformer	None	ResNet152	Yes	40.6
Transformer	Full	ResNet152	Yes	51.7
Transformer	Full	ResNeXt-IG-3.5B	No	53.9
Bag of Words	None	ResNeXt-IG-3.5B	Yes	58.6
Transformer	None	ResNeXt-IG-3.5B	Yes	65.9
Bag of Words	Word	ResNeXt-IG-3.5B	Yes	66.2
Transformer	Full	ResNeXt-IG-3.5B	Yes	77.5

Table 5: Results for TransResNet retrieval variants on the PERSONALITY-CAPTIONS test set.

of 64.5%, which is statistically significant using a binomial two-tailed test ($p < .001$).

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. 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 lower 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 Table 7 (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 text encoder pretraining, for a fairer comparison, denoted * in the table) also substantially outperforms UPDOWN ResNext-IG-3.5B with a winrate of 80.1%.

Human Evaluation of Caption Relevance In addition to our evaluation of engagingness it is important to also check that the produced captions are relevant to the corresponding image and the personality trait. In order to evaluate this we again performed crowd-sourced human evaluation for the same 500 evaluation images, where we asked annotators if captions “fit” the image and the personality trait. Results are presented in Table 8. Although human captioners are better at fitting the image (92.8% vs 90.2%), TransResNet actually outperforms them at choosing a caption that fits the personality (87.7% vs 83.1%). Note that human captioners were not told specifically that their captions should unambiguously fit the personality trait. Still, our main conclusion is that our model can indeed provide relevant captions.





Image	Personality	Generated comment
	Anxious Happy Vague Dramatic Charming	I love cats but i always get so scared that they will scratch me. That cat looks SO happy to be outside. That's a nice cat. Or is it a lion? That cat looks so angry; it might claw your eyes out! Awww, sweet kitty. You are so handsome!
	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.
	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
	Romantic Anxious Creative Sweet Money-minded	A charming home that will call you back to days gone by. This house and this street just makes me feel uneasy. I could write a novel about this beautiful old home! What a cute little neighborhood! Call APR now to get your house renovated!

Table 6: Predictions from our best TransResNet model on the PERSONALITY-CAPTIONS valid set.

Type of caption A	WIN PERCENTAGE		Type of caption B
Human personality captions	64.5	35.5	Human traditional captions
Human personality captions	50.5	49.5	TransResNet (ResNeXt-IG-3.5B)
Human personality captions	59.1	40.9	TransResNet (ResNet-152)
Human personality captions	79.3	20.7	UpDown (ResNeXt-IG-3.5B)
TransResNet (ResNeXt-IG-3.5B)	55.2	44.8	TransResNet (ResNet-152)
TransResNet (ResNeXt-IG-3.5B)*	80.1	19.9	UpDown (ResNeXt-IG-3.5B)

Table 7: 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. Our best model TransResNet (ResNeXt-IG-3.5B) is close to human performance.

Set of Captions	Fits Personality	Fits Image	Fits both
Human	83.1%	92.8%	80.5%
TransResNet	87.7%	90.2%	81.8%

Table 8: Human evaluation of caption fit.

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 intro-

duce a new retrieval architecture, TransResNet, that yields the highest known R@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. We show that our best system is able to produce captions that are close to matching human performance in terms of engagement and relevance. An important open problem that remains is to improve generative models on this task, which failed to do as well.

Capturing many types of human sentiment is crucial for moving towards agents that communicate the way people do, and in the future this research may help drive applications including safer chatbots, better text generation, and many others.

References

- [1] A. E. Abele and B. Wojciszke. Agency and communion from the perspective of self versus others. *Journal of personality and social psychology*, 93(5):751, 2007.
- [2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.
- [3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and vqa. *CVPR*, 2018.
- [4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.
- [6] H. E. Cattell and A. D. Mead. The sixteen personality factor questionnaire (16pf). *The SAGE handbook of personality theory and assessment*, 2:135–178, 2008.
- [7] A. Chandrasekaran, D. Parikh, and M. Bansal. Punny captions: Witty wordplay in image descriptions. *arXiv preprint arXiv:1704.08224*, 2017.
- [8] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [9] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017.
- [10] E. Denton, J. Weston, M. Paluri, L. Bourdev, and R. Fergus. User conditional hashtag prediction for images. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1731–1740. ACM, 2015.
- [11] A. Eisenschlat and L. Wolf. Capturing deep correlations with 2-way nets. *CoRR*, abs/1608.07973, 2016.
- [12] M. Engilberge, L. Chevallier, P. Prez, and M. Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [13] F. Faghri, D. J. Fleet, R. Kiros, and S. Fidler. VSE++: improved visual-semantic embeddings. *CoRR*, abs/1707.05612, 2017.
- [14] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. Stylenet: Generating attractive visual captions with styles. In *Proc IEEE Conf on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017.
- [15] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. *CoRR*, abs/1711.06420, 2017.
- [16] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *arXiv preprint arXiv:1802.08218*, 2018.
- [17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [18] Y. Hu, L. Manikonda, and S. 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.
- [19] Y. Huang, W. Wang, and L. Wang. Instance-aware image and sentence matching with selective multimodal LSTM. *CoRR*, abs/1611.05588, 2016.
- [20] J. C. S. Jacques, Y. Gülütürk, M. Pérez, U. Güllü, C. Andújar, X. Baró, H. J. Escalante, I. Guyon, M. van Gerven, R. van Lier, and S. Escalera. First impressions: A survey on computer vision-based apparent personality trait analysis. *CoRR*, abs/1804.08046, 2018.
- [21] T. Jay and K. Janschewitz. Filling the emotion gap in linguistic theory: Commentary on potts’ expressive dimension. *Theoretical Linguistics*, 33(2):215–221, 2007.
- [22] Jonczyk and R. Jończyk. *Affect-language interactions in native and non-native English speakers*. Springer, 2016.
- [23] O. Kampman, F. B. Siddique, Y. Yang, and P. Fung. Adapting a virtual agent to user personality. In *Advanced Social Interaction with Agents*, pages 111–118. Springer, 2019.
- [24] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [25] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.
- [26] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

- [27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [28] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. 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.
- [29] J. Lu, C. Xiong, D. Parikh, and R. 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, page 2, 2017.
- [30] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018.
- [31] 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)*, pages 2623–2631, Dec 2015.
- [32] D. Mahajan, R. B. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. *CoRR*, abs/1805.00932, 2018.
- [33] S. Marcel and Y. Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM ’10, pages 1485–1488, New York, NY, USA, 2010. ACM.
- [34] A. Mathews, L. Xie, and X. He. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8591–8600, 2018.
- [35] A. P. Mathews, L. Xie, and X. He. Senticap: Generating image descriptions with sentiments. In *AAAI*, pages 3574–3580, 2016.
- [36] P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes. Training Millions of Personalized Dialogue Agents. *ArXiv e-prints*, Sept. 2018.
- [37] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. P. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *CoRR*, abs/1701.08251, 2017.
- [38] H. Nam, J. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. *CoRR*, abs/1611.00471, 2016.
- [39] 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)*, pages 1899–1907, Oct 2017.
- [40] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos. Automatic image captioning. In *Multimedia and Expo, 2004. ICME’04. 2004 IEEE International Conference on*, volume 3, pages 1987–1990. IEEE, 2004.
- [41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.
- [42] C. C. Park, B. Kim, and G. Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6432–6440, July 2017.
- [43] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *CVPR*, volume 1, page 3, 2017.
- [44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. ImageNet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.
- [45] M. Scheutz, P. Schermerhorn, and J. 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*, pages 226–233. ACM, 2006.
- [46] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, Jan. 2016.
- [47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [48] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [49] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. *CoRR*, abs/1511.06361, 2015.

- [50] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [51] 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*, pages 1–1, 2018.
- [52] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5005–5013, 2016.
- [53] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016.
- [54] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [55] K. Yoshida, M. Minoguchi, K. Wani, A. Nakamura, and H. Kataoka. Neural joking machine: Humorous image captioning. *arXiv preprint arXiv:1805.11850*, 2018.
- [56] Q. You, H. Jin, and J. Luo. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv preprint arXiv:1801.10121*, 2018.
- [57] P. Young, A. Lai, M. Hodosh, and J. 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.
- [58] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.