Conditional GANs for Image Captioning with Sentiments

Tushar Karayil[1,2](✉), Asif Irfan[1], Federico Raue[2], Jörn Hees[2],
and Andreas Dengel[1,2]

[1] TU Kaiserslautern, Kaiserslautern, Germany
[2] DFKI, Kaiserslautern, Germany
{tushar.karayi,asif.irfanl,federico.raue,jorn.hees,
andreas.dengel}@dfki.de

Abstract. The area of automatic image captioning has witnessed much progress recently. However, generating captions with sentiment, which is a common dimension in human generated captions, still remains a challenge. This work presents a generative approach that combines sentiment (positive/negative) and variation for caption generation. The presented approach consists of a Generative Adversarial Network which takes as input, an image and a binary vector indicating the sentiment of the caption to be generated. We evaluate our model quantitatively on the state-of-the-art image caption dataset and qualitatively using a crowd-sourcing platform. Our results, along with human evaluation prove that we competitively succeed in the task of creating variations and sentiment in image captions.

Keywords: GAN · Sentiment · Caption · Policy Gradient

1 Introduction

A caption of an image is a short piece of text, provided by the user and describes the user's interpretation of the image. Automatic image captioning, where machines generate a short piece of text given an image, lies at the intersection of computer vision and natural language processing. In the last decade, there has been significant progress in generating descriptive image captions [16,17,19]. However, these image captioning approaches often only focus on describing the content of the image without any emotional or sentimental dimension. An analysis of the 3000 captions prevalent across the social media platforms (e.g. Flickr) reveal two characteristic dimensions of these textual descriptions [1]. First, sentiments are often prominently present in captions. These captions are more than mere *factual* descriptions of the image. Second, humans often use a wide variety of captions while describing images [3]. These aspects are often neglected in state-of-the-art image captioning models where the intention is to generate a caption which is often as close to the ground-truth as possible. Therefore,

T. Karayil and A. Irfan—Equal contribution from authors.

I. V. Tetko et al. (Eds.): ICANN 2019, LNCS 11730, pp. 300–312, 2019.
https://doi.org/10.1007/978-3-030-30490-4_25

in order for machines to generate effective captions for images it must at least include the above mentioned dimensions.

Our motivating question then becomes: *"Can a model generate captions with intended sentiment and variations?"*. The authors of [2] show that adjectives can be used to add sentiments and an adjective-noun pair can express the visual sentiment of an image. Therefore, we assume that incorporating adjectives into a caption enhances the sentiment component of the same. To combine sentiment and variability in a single model we use a generative approach.

Generative models have shown to be effective at approximating unknown distributions. The most successful among generative models, called Generative Adversarial Network, GAN [5] has proved to be highly efficient at tasks like image generation, image completion [10, 21] etc. A typical GAN includes a generator network which, given a noise vector z, generates data items and a discriminator network which evaluates these items (if generated or real). Together, they perform a *min-max* game, where the generators objective is to generate data which can fool the discriminator and the discriminators objective is to accurately distinguish the generated data from real. A variant of GANs, called Conditional-GAN, CGAN [10] follows an architecture where generator and discriminators are conditioned on an external input. Our method takes inspiration from the CGAN architecture where sentiment acts as the external condition.

Training GANs for text generation is a challenging task, mainly due to two reasons: First, the process of generating language is a sequential-sampling procedure which is non-differentiable, making the direct application of backpropagation difficult; Second, the generator receives the feedback from the discriminator only after the entire sequence has been produced. This leads to several problems for training sequences like vanishing gradients. The authors of [3] have shown that reinforcement learning algorithms like Policy Gradients and Monte-Carlo rollouts can be used to mitigate these effects in order to train a GAN for caption generation. Our final model takes as input, an image and a binary variable (indicating the desired positive or negative sentiment of the caption) to generate captions accordingly. Figure 1 and Table 2 show the basic architecture of the model and few examples of generated captions respectively.

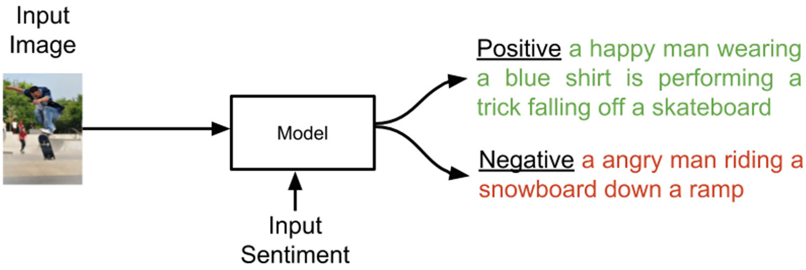

Input Image
Model
Input Sentiment
Positive a happy man wearing a blue shirt is performing a trick falling off a skateboard
Negative a angry man riding a snowboard down a ramp

Fig. 1. Basic overview of our model. The input to the model is the image and a binary vector indicating the required sentiment (positive/negative) of the output captions. The model generates the caption which has the input sentiment and multiple variations.

In this regard, the contributions of this paper can be stated as follows:

- We design a generative sentiment-captioning model which is different from the regular encoder-decoder captioning models.
- The generative model can create captions with an input sentiment and variations. The required sentiment can be provided as an external input along with the image.
- We show that sentiment information can be embedded into a GAN with a two phase training approach. Our model outperforms the state-of-the-art for objective and sentiment captions.

The rest of the paper is organized as follows. Section 2 lists related work relevant to this paper. Section 3 gives a detailed description of the model architecture and training. Section 4 describes our experiments and discusses our findings, including the human evaluation. Section 5 concludes the paper and charts out the future direction.

2 Related Work

Image captioning frameworks generally follow an encoder-decoder architecture [16,17]. The input image is encoded into a n-dimensional space using a Convolutional Neural Network (CNN). The encoded image acts as the initial state for the decoder which is a Long Short Term Memory Network (LSTM) to generate a text sequence. The network is trained using a *maximum likelihood* loss (e.g. Cross Entropy Loss). Deviating from this convention, we can have a GAN architecture wherein the generator is an LSTM network and the discriminator, as second neural network, evaluates the caption. The generator given an image, generates a sequence which is evaluated by the discriminator.

To the best of our knowledge, there has been no previous work which combines sentiments and variability together in captioning. Therefore, this section lists out the related research in two directions: (a) captioning approaches for sentiment captions, (b) captioning approaches which use generative models (for variability) to generate captions.

2.1 Sentiment Captioning

The goal of these methods is to generate captions which have some sentiment information inside them. As mentioned previously, the addition of adjectives is one method to increase the sentiment of a caption. Therefore, the methods mentioned below have an extra module to inject sentiment into the caption.

In [14], authors use an additional CNN, along with the regular image captioning encoder-decoder model which learns sentiment features. The model keeps track of the noun with the highest probability in the generated caption to add a sentiment adjective to it. The additional CNN along with the encoder-decoder model means that the number parameters required are almost doubled. A parallel Recursive Neural Network (RNN) was used in [9] with the standard encode-decoder model which emphasizes on the descriptions with sentiments. The model is trained with a switching mechanism using a binary sentiment variable associated with each word. The parallel RNN architecture here is not truly end to end and needs to be trained first with the positive set of captions and again with a negative set of captions. The authors of [11] used a method which takes into account the current focus point inside the image (attention) to get the sentiments into the generated captions. A high-level as well as word level sentiment information with two different sentiment vectors are used here to capture the general sentiment of the image. Modified LSTM modules have also been proposed by [18] to inject sentiments.

The above mentioned approaches differ from our model in two ways: First, they are based on a variant of encoder decoder model. Second, they are trained on the *Maximum Likelihood* principle. This enforces the model to generate very rigid captions lacking the extra dimension of variability.

2.2 Generative Methods

The approaches listed here are based on generative models, specifically GANs. GAN training is often aided by the use of reinforcement learning algorithms. Here the discriminator acts as a reward agent and the objective of the generator is to create sequences which maximize the reward. The authors of [20] propose a framework to use GANs to train a captioning framework with the policy gradient algorithm. Policy gradient approaches try to find an optimal policy/rule through feedback from the discriminator. The work in [3] applies the aforementioned policy gradient based training to introduce variations in caption generation.

The above mentioned generative methods differ from our model in two ways: First, their focus is on generating neutral or objective captions and the sentiment dimension is found missing. Second, the z vector of the model is a random input whereas our model uses the z dimensions to encode input sentiment.

3 Proposed Method

We use an architecture similar to a CGAN [3] but with one generator and two discriminators. Our training also differs from the adversarial approach [5]. Briefly put, our training contains two phases. In the first phase, we train both the generator and discriminator. After the first phase, the discriminator weights are frozen and they now act as reward agents. In the second phase, the rewards produced by the discriminators (for the generated captions) act as a feedback to further train the generator. Training using this reinforcement technique is called policy gradients.

3.1 Architecture

Our model consists of a generator, G and two discriminator networks, D_r, D_s (Fig. 2 shows the detailed architecture of the model). G is a single layer LSTM network (hidden dimension h_g) which takes an image along with a noise vector

$z \in \mathbb{R}^m$ as input and generates a caption by sampling discretely from the output. The input image is first converted into a feature vector, $f \in \mathbb{R}^n$ using the last fully connected layer of a pretrained CNN. The objective of the generator is to generate captions which are relevant to the image and have a positive/negative sentiment based on the input noise vector z. The objective of D_r and D_s is to accurately judge the relevance and the sentiment of the generated caption respectively.

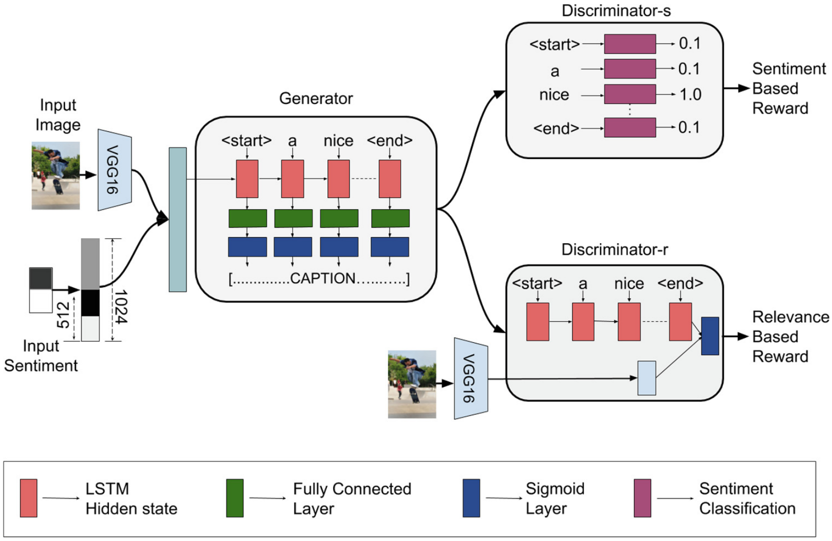


Fig. 2. Detailed architecture of our model. The Generator takes the image and a binary sentiment vector as input. Discriminator-r uses the same image to evaluate the quality of the generated caption. Discriminator-s uses the input sentiment vector and the generated caption while evaluating its reward.

The first discriminator D_r, is a LSTM network (hidden dimension h_d), which given an image and a caption, distinguishes between the captions generated by G from the ones present in the training set. D_r also takes into account the semantic relevance of the generated caption given the input image and the true caption of the input image. The objective function to train the discriminator is an extended version used by [3]. For D_r with parameters η, given an image I, the objective function (1) and reward (2) can be formulated as:

$$\begin{aligned} L_{D_r}(I;\eta) = \mathbb{E}_{S_r \sim S_T} \log R_{D_r}(I, S_r) + \alpha \cdot \mathbb{E}_{S_g \sim S_G} \log\left(1 - R_{D_r}(I, S_g)\right) \\ + \beta \cdot \mathbb{E}_{S_n \in S_N} \log\left(1 - R_{D_r}(I, S_n)\right) \end{aligned} \tag{1}$$

$$R_{D_r} = \sigma(f_\theta(I) \cdot h_\eta(S)) \tag{2}$$

where η represents parameters of D_r, θ represents the parameters of the CNN, f and h are embedding functions of image and caption respectively, $<\cdot>$ is the dot product, S_T is the true caption for I from the training set, S_G is a generated caption from G for I and S_N is a "irrelevant-caption" from the training set that does not belong to I. α and β are balancing coefficients.

The second discriminator, D_s takes the generated caption from G, the input sentiment vector and assigns a reward for each of the tokens generated by the generator. Our experiments showed that a pre-trained sentiment classifier can also be used with our modified objective function[1]. D_s provides a high reward if the computed sentiment is the same as the expected sentiment and punishes G for deviations. Thus, the reward from D_s can be defined as follows:

$$R_{D_s}(S, \omega) = \mathbb{E}_{S \sim S_G}[\delta_{wp} \log f_p(S) + \delta_{wn} \log f_n(S)] \tag{3}$$

$$f_p(t) = \begin{cases} 1, & s(t) > 0.5, \\ 0.8, & 0 \leq s(t) \leq 0.5, \\ 0.1, & s(t) < 0 \end{cases} \qquad f_n(t) = \begin{cases} 1, & s(t) < -0.5, \\ 0.8, & -0.5 \leq s(t) \leq 0, \\ 0.1, & s(t) > 0 \end{cases}$$

where $\omega \in \{p, n\}$ is the input sentiment, $s(t)$ is sentiment value of token t assigned by D_s and δ is the kronecker delta. It should be noted that, after the discriminators are trained, their role is to provide a reward to each token in the caption generated by G.

3.2 Training

We divide the model training into two phases. In the first phase, the generator and discriminators are trained. After the first phase, the generator is able to generate words which are relevant to the image (without any specific language structure). For the second phase, the discriminators are frozen and the generator is further trained (via policy gradients) to incorporate sentiment and language variations in the caption. We found that this method of training increased the stability of the model and prevented the model from the "helvetica scenario" or mode collapse [5].

Phase 1. The generator G in this setup was pre-trained with maximum likelihood estimation technique for e_g epochs. This pre-training was done in order to stabilize the gradients. We reach a stage where the generator starts generating some relevant words related to the image. The discriminator D_r was then trained using this generator for e_r epochs (with the loss function in Eq. 1). Although, it can be argued that this is not truly an alternating adversarial training, we did not find any significant difference in variations of captions with alternating adversarial training. Moreover, with our approach it reduced the training time (in terms of number of required epochs) of Generator. The noise variable z was set to a 1024 sampling from $\mathcal{N}(0, 1)$.

[1] We used sentiment classifier provided by TextBlob (https://textblob.readthedocs.io/en/dev), which provides a sentiment value in $[-1, 1]$.

Phase 2. We used a policy gradient approach to further train G wherein the discriminators D_r and D_s act as reward agents. This means that G needs to generate captions through which it can maximize the rewards given by D_r and D_s. We also want to provide an input to G for the intended sentiment of the caption. To achieve this, we used the noise variable z. We split z into two parts: a 512-dimensional vector sampled from $\mathcal{N}(0,\,1)$ and a 512-dimensional latent code vector which is assigned values based on the sentiment in the ground truth caption. If the ground truth caption was positive, the latent code is assigned values such that the first 256 dimensions are set to 1 and the rest to 0 and vice-versa (using smaller dimensions than 256 made the generator to ignore these values). We then used D_s in a way which forces the generator to use this information. This ensures that there is high correlation between G's distribution and the dimensions of z vector. Therefore, the Phase-2 loss of G can be formulated as follows:

$$L_G(I) = \mathbb{E}_{S_g \in S_G}[-\gamma_1 \cdot R_{D_r} - \gamma_2 \cdot R_{D_s} + \gamma_3 \cdot \varOmega(S_g, S_t)] \tag{4}$$

where, $\gamma_1, \gamma_2, \gamma_3 \in (0, 1]$ are the balancing coefficients learned from the validation set and $\varOmega$ is a regularizing term used to prevent the discriminator from collapsing to trivial patterns. We found that setting $\varOmega$ to cross-entropy function (between generated caption S_g and true caption S_t of the image) gave the best results. During inference, to generate variations, the first 512-dimensions are sampled from $\mathcal{N}(0,\,1)$ (the last 512 are set according to the required sentiment as mentioned above). Table 2 shows these variations where the first 512 dimension of z are changed to create three different captions for a given sentiment.

4 Experiments

In this section we describe the dataset, hyperparameters and the results from our experiments. We evaluate the results both quantitatively and qualitatively to get a better understanding of our performance. For these experiments we have chosen only the positive and negative emotions. This is because the state-of-the-art models in sentiment captioning have used only positive/negative emotion [9,11]. Therefore, choosing the same number of emotions for our experiments gave us a clear way to compare our results against the state-of-the-art.

4.1 Sentiment Enhanced MSCOCO

Microsoft-COCO (MSCOCO) [8] is an image-caption dataset containing 150,000 image-caption pairs in total (train, validation and test) and is also the preferred dataset for state-of-the-art image-captioning research [3,9,11,16,18,19]. Therefore, we chose to use MSCOCO as this gave us a clear way to compare our results against the state-of-the-art. Although MSCOCO is the benchmark dataset for captioning models, the captions provided are quite objective and clearly lack the sentiment dimension. A sentiment classification showed us that there were only 29,521 and 26,851 captions with a positive and negative sentiment respectively. The rest, 61,915 were neutral captions. The sentiment captions dataset from [9] (with 998 images) was found to be too small for our training task. To overcome these challenges, we decided to modify each of the nouns present in the MSCOCO dataset with a suitable positive or negative adjective. The intention was to enhance the sentiment value of the training set. Rather than randomly adding positive and negative adjectives, we used the work [6], to find the list of suitable adjectives for each noun. We used the 2017 train/val split of MSCOCO which consists of 118,287 training images and 5,000 validation images[2]. For each of these images, there are 5 captions in the dataset. Following the sentiment enhancement, we processed each of these captions similar to [3]: (a) remove all the non alphabetic characters apart from comma, (b) convert all the words to lower-case, (c) add a START (<*start*>) and END (<*end*>) token at the beginning and end of each caption, (d) remove all the words with the frequency of less than 5 in training and validation set combined. This gave us the vocabulary size of 10,496 words. All the words that were not in the vocabulary were replaced with a token <*unk*>. We used the maximum sequence length of 16 and thus truncated all captions up until this length and padded the shortened sequences with token <*pad*>. After the changes, the sentiment-enhanced MSCOCO contained 50,303 positive and 67,981 negative image-caption pairs.

4.2 Hyperparameters

In this section we describe the set of parameters which were empirically determined based on the validation set. The hidden dimensions of the generator and discriminator lstm networks, h_g, h_d were both set to 512. The VGG16 network was used as the feature extractor for images with the feature vector $f \in \mathbb{R}^{4096}$. The noise vector z was from R^{1024}. The coefficients for Eq. 1, α, β were set to 1. The coefficients for Eq. 4, $\gamma_1, \gamma_2, \gamma_3$ were set to $1, 1, 0.5$ respectively. For the first phase training, epochs e_g and e_r were $50, 30$ respectively. For the second phase, e_g was 100.

4.3 Results

Since our work addresses the dimensions of sentiment and variability, the results were evaluated both quantitatively as well as qualitatively. Quantitative evaluation usually involves reporting conventional scores of BLEU [12], METEOR [4], ROUGE [7] and CIDEr [15] against the ground truth. Qualitative evaluation uses human subjects to evaluate the generated captions for sentiment and grammar.

Quantitative Results. Classical score like BLEU, METEOR, ROGUE and CIDEr are generally evaluated by matching n-grams between target and the

[2] MSCOCO does not have ground-truth captions for the test set.

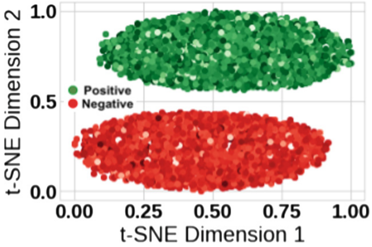
Table 1. Conventional metrics for Show n Tell (SnT), SentiCap (SCap) (for both positive and negative captions) and our model (with 1, 5 and 10 generated captions). Even though our objective is not to maximize conventional scores, we still outperform both objective and sentiment models in most of these scores as we increase the variations. SnT scores are the same for Positive and Negative captions because they generate a neutral caption.

	Metric	SnT [16]	SCap [9]	Ours		
				c = 1	c = 5	c = 10
Positive captions	BLEU-1	0.620	0.567	0.547	**0.621**	**0.656**
	BLEU-2	0.437	0.365	0.346	0.406	**0.439**
	BLEU-3	**0.306**	0.240	0.220	0.267	0.295
	BLEU-4	**0.218**	0.164	0.144	0.181	0.202
	METEOR	0.219	0.199	0.185	0.209	**0.221**
	ROUGE_L	0.473	0.443	0.418	0.469	**0.488**
	CIDEr	**0.752**	0.545	0.461	0.591	0.631
Negative captions	BLEU-1	0.620	0.572	0.570	**0.645**	**0.676**
	BLEU-2	0.437	0.367	0.362	0.428	**0.463**
	BLEU-3	0.306	0.246	0.234	0.287	**0.319**
	BLEU-4	0.218	0.164	0.151	0.191	**0.219**
	METEOR	0.219	0.200	0.199	**0.222**	**0.235**
	ROUGE_L	0.473	0.447	0.445	**0.483**	**0.504**
	CIDEr	**0.752**	0.516	0.509	0.627	0.688

generated captions. Therefore, a higher score would suggest that the generated caption is closer to the target sentence. Even though our models are not trained to emulate the ground truth (in turn maximize the benchmark scores), we would like to report these scores to show that we can still outperform the state-of-the-art, simply by increasing the variations for our captions. To compare against the state-of-the-art for objective captioning, we use the "Show n Tell" [16] model. For comparison against the state-of-the-art for sentiment captioning, we use "SentiCap" [9] model. We use the test set published by [9] which contains 433 positive and 433 negative image-caption pairs. The results show that even though we did not train the model according to the conventional criteria, we competitively outperform the state of art as shown in Table 1. As we increase our generated captions ($c = 1, 5, 10$), we also get some variations which are similar to the ground truth, thereby achieving high value for these scores. Furthermore, we have to use the same underlying vocabulary to generate variations. The nouns present in the ground truth caption, like park, kitchen, man etc. are present in the generated captions/variations as well (although their positions are different) providing a boost to these values.

In order to determine whether the z vector truly encodes the intended sentiment, we created 30,000 pairs (15,000 positive/negative each) of encoded z

vectors and calculated the sentiment of the generated caption. We then used t-SNE to visualize these vectors. Figure 3 (left) shows the distribution of these vectors. As can be seen, there are two clusters that represent two different sentiment encoded z vectors. The colors indicate the sentiment of the generated caption. Each of the two clusters are dominated by a single output sentiment (positive or negative) as indicated by their color coding. Figure 3 (right) shows the confusion matrix w.r.t the sentiment. As seen from the confusion matrix, the overall accuracy of the intended sentiment is 93.19%. The results (visual and confusion matrix) indicate that the encoding scheme is effective and achieves the intended sentiment in the generated caption.



		Generated	
		Pos	Neg
Target	Pos	13,498	1,502
	Neg	541	14,459

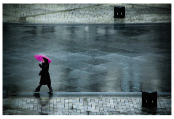
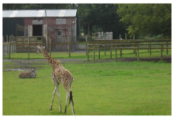
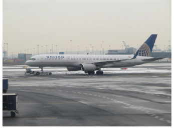
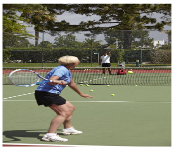
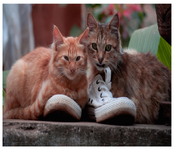
Fig. 3. The plot (left) shows the t-SNE projection of the z vectors onto a 2-D space. Each of the two clusters formed by z vectors are dominated by a single sentiment (of generated caption). The right side shows the confusion matrix w.r.t the sentiment (expected vs generated sentiment).

Human Evaluation. As our task involves generating variable and sentiment captions, a fair evaluation is only possible through humans. The evaluation should include judgments about: (a) the validity of the caption (given the image), (b) the sentiment of the caption. In this regard, we randomly sampled 200 *held-out* images from the validation set of MSCOCO. We generated 3 positive and 3 negative captions per image. Each image-caption pair was evaluated by 3 subjects and a majority vote decided the final answer. Through the entire experiment, we collected 3600 responses from human subjects. Each subject was shown an image and 6 captions (3 positive, 3 negative) but in a random order. For each caption, given the image, subjects were then asked to answer the following questions:

1. Is this a valid caption for the given image?
2. What is the general sentiment of the caption: positive, neutral, negative?

We used the crowd sourcing platform, Amazon Mechanical Turk to conduct our human evaluation. To ensure quality, in addition to a validation set, only contributors with a minimum rating of 75% were allowed to participate. From the 3600 responses collected, 77.7% of the generated captions were voted valid and having the intended sentiment. This clearly indicates that the captions (and

Table 2. Positive and negative captions generated by our model. Positive/Negative captions are generated by providing the respective external input to the GAN along with the image. For each sentiment (+/−), there are three variations shown.

Image	**Captions**
	+ a proud woman walking down the street holding a colorful umbrella. + a attractive person walking across a street holding a umbrella. + a great person walking with a umbrella on top of a street. − a dangerous person walking down the street in the rain. − a evil person walking across with a umbrella. − a dangerous person walking holding a pink umbrella.
	+ a beautiful giraffe standing on top of a lush green field. + a beautiful giraffe standing near a tree in a field. + a wonderful giraffe in a field with a bird in the background. − a sad giraffe standing in a field next to a bush. − a sick giraffe standing in a lush green field. − a sick giraffe standing in a field next to a tree.
	+ a white and blue great plane is on a runway. + a popular passenger jet is parked on the runway. + a large white great airplane sitting on a runway. − a white and blue jet sitting on a wrong runway. − a expensive passenger jet is parked on the runway at an airport. − a fake airplane that is sitting on a runway.
	+ a professional tennis player returns a real shot. + a thoughtful woman plays a forehand. + a realistic tennis player returns a shot. − a crazy person plays a shot. − a unpredictable tennis player. − a angry woman on the ground.
	+ a adorable kitten is sitting on white couch. + a cute cat sleeps on a white sheet. + a proud group of cats resting on a couch. − a dramatic cat rests on a table. − lazy group of cats by the table. − a ridiculous cat laying on top of a piece of luggage.

the variations) from our model were of high quality (semantically relevant) and had the intended sentiment. Table 2 shows few examples that we used for this task. In 10.3% of the cases, the subjects voted for a "neutral" sentiment because of the generated adjective not being strong enough to convey the sentiment.

5 Conclusion and Future Work

The paper presented a generative approach to combine sentiments and variations for captions in a single model. To achieve this goal, we used an architecture similar to a GAN, training it with policy gradients. We trained the generator with two different discriminators to generate sentiment captions with a variability dimension. We showed that our model competitively outperforms the two state-of-the-art models (for objective and sentiment captions) for image captioning. To further evaluate the results, we also performed a human evaluation and showed that 77.7% of the generated captions are valid with intended sentiments. Our results imply that it is possible to generate variable-sentiment captions with good degree of accuracy. We plan to explore the control that one can have w.r.t the variations and sentiment. An interesting direction in this regard is to understand how each dimension of the $z - vector$ contribute to sentiments/variations. We also plan to extend this model with the newly introduced Google Conceptual Captions dataset [13] with 3.2 million images.

Acknowledgements. This work was supported by the BMBF project DeFuseNN (Grant 01IW17002) and the NVIDIA AI Lab (NVAIL) program.

References

1. Blandfort, P., Karayil, T., Borth, D., Dengel, A.: Image captioning in the wild: how people caption images on Flickr. In: Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes, pp. 21–29. ACM (2017), https://doi.org/10.1145/3132515.3132522
2. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S.F.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 223–232. ACM (2013). https://doi.org/10.1145/2502081.2502282
3. Dai, B., Fidler, S., Urtasun, R., Lin, D.: Towards diverse and natural image descriptions via a conditional GAN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2970–2979 (2017). https://doi.org/10.1109/ICCV.2017.323
4. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380 (2014)
5. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
6. Karayil, T., Blandfort, P., Hees, J., Dengel, A.: The focus-aspect-value model for explainable prediction of subjective visual interpretation. In: International Conference of Multimedia Retrieval (2019). https://doi.org/10.1145/3323873.3325026

7. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
8. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
9. Mathews, A.P., Xie, L., He, X.: SentiCap: generating image descriptions with sentiments. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
10. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint (2014)
11. Nezami, O.M., Dras, M., Wan, S., Paris, C.: Senti-attend: image captioning using sentiment and attention. arXiv preprint arXiv:1811.09789 (2018)
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: 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). https://doi.org/10.3115/1073083.1073135
13. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 2556–2565 (2018)
14. Shin, A., Ushiku, Y., Harada, T.: Image captioning with sentiment terms via weakly-supervised sentiment dataset. In: BMVC (2016)
15. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015). https://doi.org/10.1109/CVPR.2015.729908
16. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015). https://doi.org/10.1109/CVPR.2015.7298935
17. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
18. You, Q., Jin, H., Luo, J.: Image captioning at will: a versatile scheme for effectively injecting sentiments into image descriptions. arXiv preprint arXiv:1801.10121 (2018)
19. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651–4659 (2016). https://doi.org/10.1109/CVPR.2016.503
20. Yu, L., Zhang, W., Wang, J., Yu, Y.: SeqGAN: sequence generative adversarial nets with policy gradient. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
21. Zhang, H., et al: StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5907–5915 (2017). https://doi.org/10.1109/ICCV.2017.629