

3M: Multi-style image caption generation using Multi-modality features under Multi-UPDOWN model

Chengxi Li and Brent Harrison

University of Kentucky
Lexington, KY 40506

Abstract

In this paper, we build a multi-style generative model for stylish image captioning which uses multi-modality image features, ResNeXt features, and text features generated by DenseCap. We propose the 3M model, a Multi-UPDOWN caption model that encodes multi-modality features and decodes them into captions. We demonstrate the effectiveness of our model on generating human-like captions by examining its performance on two datasets, the PERSONALITY-CAPTIONS dataset, and the FlickrStyle10K dataset. We compare against a variety of state-of-the-art baselines on various automatic NLP metrics such as BLEU, ROUGE-L, CIDEr, SPICE, etc ¹. A qualitative study has also been done to verify our 3M model can be used for generating different stylized captions.

Introduction

Factual image captioning is one of the fundamental tasks in deep learning. The issue with factual captions is that language generated is often stilted, and not necessarily representative of human communication. While classic image captioning approaches show deep understanding of image composition and language construction, it often lacks elements that make communication distinctly human. To address this issue, some researchers have tried to add personality to image captioning in order to generate stylish captions. In general, stylish captioning systems are divided into two categories based on how they are trained: single style and multi-style. Single-style training involves training one model for each personality, whereas multi-style techniques learn to generate captions in many different styles using one model.

Shuster *et al.* (Shuster et al. 2019) built a multi-style module by converting each personality to a multi-dimensional vector. Their generative model struggled to generate captions that accurately captured the given image context. This is likely because a multi-style captioners require greater knowledge about the input image when compared to single-style captioners. To address the inherent limitations of past multi-style captioning approaches, we propose the use of multi-modality image features to improve the quality

of multi-style image captioning. We believe that multi-modality features, specifically image features combined with features derived from text describing said image, will help the model better ground image features into text.

To effectively generate stylish captions, a model needs to incorporate elements of the local context of image regions and the global context of the image itself. To capture local context, our model will make use of region-based caption features generated by the DenseCap network (Johnson, Karpathy, and Fei-Fei 2016).

To complement dense caption features, we will also use ResNext features describing the global input image. To combine these features, we introduce a Multi-UPDOWN structure model where each UPDOWN structure is used to select the best feature from its own modality. These selections are then fused to generate the caption.

To evaluate the performance of our multi-style captioning model, we examine its performance on different stylish image captioning datasets. We evaluate its performance using various NLP metrics and compare against several state-of-the-art baselines. We perform an ablation study in which we examine how each part of our model contributes to the overall expressiveness and diversity of our generated captions. We also perform a qualitative evaluation in which we examine how well the captions generated by our model capture image and style context.

Related Work

Captions in FlickrStyle10K are created to have either a Humorous or Romantic linguistic style (Gan et al. 2017) while captions in PERSONALITY-CAPTIONS are created to be engaging and have a conversational style (Shuster et al. 2019). With FlickrStyle10K, researchers have built single-style captioners (Gan et al. 2017; Chen et al. 2018) where they make use of both factual captions and stylized captions for training. Later researchers explored training multi-style networks (Guo et al. 2019; Zhao, Wu, and Zhang 2020) that can generate multiple types of stylish outputs using a single model.

Shuster *et al.* released the PERSONALITY-CAPTIONS containing 215 personalities in 2019 for building engaging caption generations models. In their work, Shuster *et al.* built an image caption retrieval model and also explored the multi-style generative caption models along with various im-

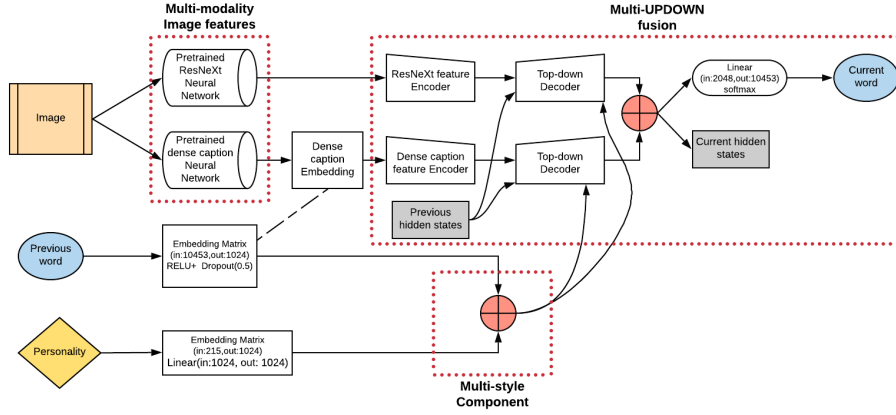


Figure 1: Architecture for Multi-style image caption generation using Multi-modality features under Multi-UPDOWN model

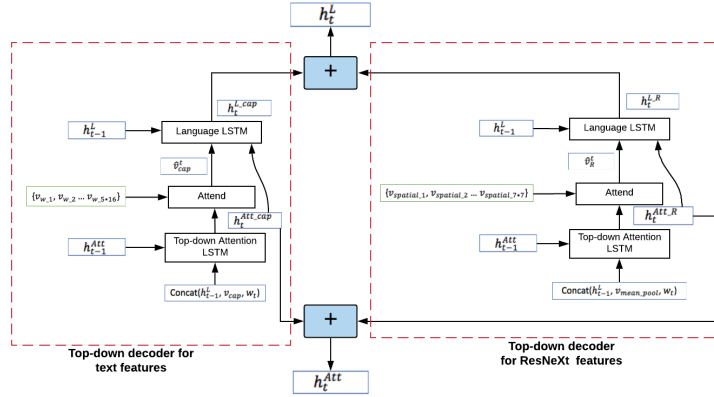


Figure 2: Two Decoders Fusion Details

age encoding strategies using several state-of-the-art image captioning models (Xu et al. 2015; Anderson et al. 2018). They applied a supervised learning model plus reinforcement learning training strategy using CIDEr score (2015) as reward. We extend the best performing supervised model presented in Shuster *et al.*'s work to build a multi-style model which incorporates multi-modality image features.

Method

The primary contribution of this paper is an architecture that utilizes multi-modality fusion for performing multi-style image captioning. This architecture specifically utilizes the soft fusion of two parallel encoder-decoder blocks, with each block containing an UPDOWN-like attention module. Our overall architecture for one step generation can be seen in Figure 1, where our multi-UPDOWN fusion blocks synthesize the information from multi-modality image features, multi-style components (previous word, personality) and previous hidden states to predict current word and hidden states at each time step.

Basically, we utilize two features from pre-trained net-

works: ResNeXt (Xie et al. 2017) visual features and text features describing the image itself (Johnson, Karpathy, and Fei-Fei 2016). These features allow the learner to better ground the image features into natural language.

Multi-style Component

As in the Figure 1, the desired style of the output caption is given as an input to our system using a one-hot vector. We then use an embedding matrix W_{p_embed} and a linear layer to encode each style into a fixed-size vector, we call it style vector p . For each word in our target stylized caption, we use another embedding matrix W_{embed} to embed each word. We will use W_{embed} to embed the dense captions too. This enables us to better connect image features to natural language. To better enable our network to generate words according to the given style, we concatenate each embedded word vector with the p to create a stylized word vector, w_t .

Multi-modality Image Features

Our architecture relies on two sets of bottom-up features extracted using pre-trained networks: ResNeXt features and

dense caption features. Specifically, we extract mean-pooled image features and spatial features from the ResNeXt network (2019) and 5 dense captions from each image with a dense caption network (2016). Each word in the dense captions is embedded using W_{embed} . By collecting both visual and text features, we provide our architecture with a more complete understanding of the full context of the image.

Multi-UPDOWN fusion Model

Our fusion model is composed of two individual encoders, the ResNext feature encoder and the dense caption encoder. Our model also uses a fused Top-down fashion decoder, which used to decode captions from encoded image features.

ResNeXt Feature Encoder and Dense Caption Encoder We encode the ResNeXt mean-pooled image features and spatial features using a linear layer, dropout layer and activation layer and get mean-pooled feature vector $\mathbf{v}_{mean.pool}$ and spatial feature vector $\mathbf{v}_{spatial.1}, \mathbf{v}_{spatial.2}, \dots, \mathbf{v}_{spatial.7*7}$. These are used as input features for the decoding process showed in the right branch of Figure 2. Then, we encode each embedded caption vector $Cap_i, i \in \{1, 2, 3, 4, 5\}$ using the Dense Caption Encoder, which is an LSTM network (1997) shown below where $\mathbf{w}_{t,i}^{dp}$ denotes a word vector in Cap_i at time t .

$$\mathbf{h}_{t,i}^{dp}, \mathbf{c}_{t,i}^{dp} = LSTM(\mathbf{w}_{t,i}^{dp}, (\mathbf{h}_{t-1,i}^{dp}, \mathbf{c}_{t-1,i}^{dp})) \quad (1)$$

We concatenate all 5 hidden states \mathbf{h}_i^{dp} into one vector \mathbf{v}_{cap} , which we call the *caption vector*. To apply attention on specific words during the decoding procedure, we keep all word states $\mathbf{c}_{t,i}^{dp}$ from the LSTM encoding process denoted as $\mathbf{v}_{w_1}, \mathbf{v}_{w_2} \dots \mathbf{v}_{w_L}$ where 5 captions contain total L words.

Top-down Decoder Fusion As we show in Figure 2, we apply the Top-down decoder model on encoded visual features and text features. The left branch is the Top-down decoder for our text features generated by the dense caption network and the right branch is the Top-down decoder for the ResNeXt features. At each time step, the Top-down decoder for text features generates a caption attention vector $\mathbf{h}_t^{Att.cap}$ by taking in the previous attention vector hidden states \mathbf{h}_{t-1}^{Att} as well as the concatenation of previous language model hidden states \mathbf{h}_{t-1}^L , the caption vector \mathbf{v}_{cap} and the previous stylized word vector \mathbf{w}_t as input.

$$\mathbf{h}_t^{Att.cap} = TopDownAttLSTM([\mathbf{h}_{t-1}^L, \mathbf{v}_{cap}, \mathbf{w}_t], \mathbf{h}_{t-1}^{Att}) \quad (2)$$

To calculate the *attended caption feature vector* We use a process inspired by (2018). We use vectors $\mathbf{v}_{w_1}, \mathbf{v}_{w_2} \dots \mathbf{v}_{w_L}$ and the caption attention vector $\mathbf{h}_t^{Att.cap}$ in the below equations:

$$\mathbf{a}_{i,t} = \mathbf{w}_a^T \tanh(W_{va} \mathbf{v}_{w_i} + W_{ha} \mathbf{h}_t^{Att.cap}) \quad (3)$$

$$\boldsymbol{\alpha}_t = \text{softmax}(\mathbf{a}_t) \quad (4)$$

$$\hat{\mathbf{v}}_{cap}^t = \sum_{i=1}^K \alpha_i^t \mathbf{v}_{w_i} \quad (5)$$

where $W_{va} \in \mathbb{R}^{H \times V}, W_{ha} \in \mathbb{R}^{H \times M}$ and $\mathbf{w}_a \in \mathbb{R}^H$ are learned parameters. This attention vector $\hat{\mathbf{v}}_{cap}^t$ is used

as the input to the language LSTM layer where the initial state is the previous hidden state from the language model, \mathbf{h}_{t-1}^L . This language LSTM then outputs the current language model hidden states $\mathbf{h}_t^{L.cap}$ for our text features as below:

$$\mathbf{h}_t^{L.cap} = LanguageLSTM([\hat{\mathbf{v}}_{cap}^t, \mathbf{h}_t^{Att.cap}], \mathbf{h}_{t-1}^L) \quad (6)$$

We calculate the ResNeXt attention vector $\mathbf{h}_t^{Att.R}$, and current language model hidden states from ResNeXt features $\mathbf{h}_t^{L.R}$, using a similar process with a separate network (shown in Figure 2 right branch). We generate the final language hidden states of the current step \mathbf{h}_t^L by fusing $\mathbf{h}_t^{L.cap}$, $\mathbf{h}_t^{L.R}$ as below:

$$\mathbf{h}_t^L = \mathbf{h}_t^{L.cap} + \mathbf{h}_t^{L.R} \quad (7)$$

We generate the final attention hidden states of the current step \mathbf{h}_t^{Att} by fusing $\mathbf{h}_t^{Att.cap}, \mathbf{h}_t^{Att.R}$ as below:

$$\mathbf{h}_t^{Att} = \mathbf{h}_t^{Att.cap} + \mathbf{h}_t^{Att.R} \quad (8)$$

We get the final language output as below:

$$\mathbf{h}_t^{output} = Dropout(\mathbf{h}_t^{L.cap}) + Dropout(\mathbf{h}_t^{L.R}) \quad (9)$$

Then we apply a linear layer to project the final language output \mathbf{h}_t^{output} to the vocabulary space and use a log softmax layer to convert it to a log probability distribution.

Experimental Methodology

To demonstrate the effectiveness of our model on stylish image captioning, we use the PERSONALITY-CAPTIONS dataset, which contains 215 distinct personalities. To prove our model is expandable to linguistic stylized captions, we train our model using the FlickrStyle10K dataset (Gan et al. 2017) which contains humorous and romantic personalities. We compare our results with the state-of-the-art work on the same datasets based on their automatic evaluation metrics. Ablation studies are also done to justify the contributions of each component of our method. We also perform a qualitative examination of the outputs of our model.

Dataset Details

The ground truth captions in PERSONALITY-CAPTIONS (Shuster et al. 2019; Thomee et al. 2016) are created to be engaging and have a human-like style. Each data entry in this dataset is represented as a triple containing an image, personality trait, and caption. In total, 241,858 captions are included in this dataset. In this work, we do not use the full PERSONALITY-CAPTIONS dataset due to accessibility of some examples. In total, our reduced dataset contains 186698 examples in the training set, 4993 examples in the validation set, and 9981 examples in the test set. The total vocabulary size of PERSONALITY-CAPTIONS after replacing infrequent tokens with 'UNK' is 10453.

The FlickrStyle10K dataset captions focus on linguistic style. Since only 7000 images are publicly available, we evaluate using a similar process to the one outlined in (2019; 2020). First we randomly select 6,000 images as the training data and use the remaining 1000 images as testing data. We further split 10% data from training data as validation data. Total vocabulary size of FlickrStyle10K is 8889.

Method	Caption Model	Training Method	Text Features	ResNeXt	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE
UPDOWN (2019)	UPDOWN	Supervised+Reinforcement	No	Yes	44.0	8.0	27.4	16.5	5.2
UPDOWN (2019)	UPDOWN	Supervised	No	Yes	40.5	6.9	26.2	16.2	4.0
3M	Multi-UPDOWN	Supervised	Yes	Yes	43.0	8.0	27.6	18.6	4.8

Table 1: Performance of Generative Models on PERSONALITY-CAPTIONS Dataset. Note: Results of (2019) under supervised learning are from re-training due to performance on supervised method not reported in (2019) and some data of original dataset not available. We also listed original result of (2019) which is under supervised and reinforcement learning for reference.

Caption Model	Personality	Text Features	ResNeXt	BLEU1	BLEU4	ROUGE-L	CIDEr	SPICE	Unique words(#)
Multi-UPDOWN	No	Yes	Yes	34.0	3.5	22.3	11.1	3.6	257
UPDOWN	Yes	No	Yes	42.4	7.5	26.7	17.9	4.4	1558
UPDOWN	Yes	Yes	No	43.2	8.1	27.6	18.0	4.6	1048
Multi-UPDOWN	Yes	Yes	Yes	43.0	8.0	27.6	18.6	4.8	1378

Table 2: Results of Ablation Studies on PERSONALITY-CAPTIONS Dataset

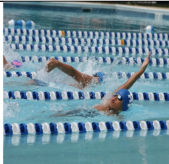
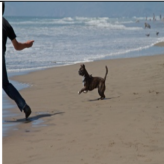



Index	R1	R2	R3	W1	W2
Dataset	PERSONALITY-CAPTIONS	PERSONALITY-CAPTIONS	FlickrStyle10K	PERSONALITY-CAPTIONS	FlickrStyle10K
Image					
Five Dense Captions	1. a small brown dog 2. blue and white shorts 3. a blue and white surfboard 4. man is wearing a wetsuit 5. man holding a surfboard	1. a dog in the air 2. the dog is on the beach 3. two horses in the sand 4. the man is jumping 5. the dog is black	1. dog playing with frisbee 2. a dog with a red collar 3. a dog wearing a collar 4. a dog with a green collar 5. a red collar on a dog	1. large gray rock wall 2. large rock in the background 3. large rock in the background 4. rocks on the ground 5. large rock in the background	1. a man riding a bicycle 2. a clear blue sky 3. a bike on a bike 4. a red and white hat 5. the bike is blue
Generated Captions/ Ground Truth	<i>what a great way to spend the day together. (Sophisticated)</i>	<i>what a cute dog! i want to play with him! (Exciting)</i>	<i>two dogs are running through a grassy field to search for bones. (Humorous)</i> <i>two dogs in love are running through a grassy field. (Romantic)</i>	<i>i don't know what i'm looking at, but i'm (Anxious)</i> ----Other style generations---- i wonder how many rocks have been there. (Mystical) i would love to climb this rock! (Whimsical (Playful, Fanciful)) i feel bad for the people who have to live in this area. (Empathetic) i can't wait to climb this rock! (Energetic)	<i>a man in a helmet rides a bike on a bike to win the. (Humorous)</i> <i>a man on a motorcycle rides a bike to impress his lover. (Romantic)</i>

Figure 3: R1-R3: Generated Captions samples using 3M trained on PERSONALITY-CAPTIONS and FlickrStyle10K (underscored). W1-W2: Imperfect Captions generations samples using 3M trained on PERSONALITY-CAPTIONS and FlickrStyle10K (underscored) along with generations from the same image and other personalities, personality are listed in parenthesis, ground truth has the same personality as the underscored generations

Training and Inference

In the training, we use entropy as loss function and Adam optimization with initial learning rate of $5e-4$. The learning rate decays every 5 epochs. In total, we train 30 epochs when using the PERSONALITY-CAPTIONS dataset (2019) with a batch size of 128 and evaluate the model every 3000 iterations. We train for 100 epochs when using the FlickrStyle10K dataset (2017) with batch size 128 and evaluate model every 100 iterations.

During inference, we generate captions using beam search with beam size 5. During this process, we impose a penalty to discourage the network from repeating words, from ending on words such as an, the, at, etc and from generating special tokens, like 'UNK'.

Quantitative Analysis

Our quantitative analysis is meant to show that our 3M model can outperform several state-of-the-art baselines in terms of a set of automated NLP metrics. In addition, we run an ablation study to validate the need for each part of

the 3M model.

Baselines and Evaluation Metrics To test if our 3M model can be used to generate human-like captions, we train it using the above settings on the PERSONALITY-CAPTIONS dataset. We compare against the model introduced previously by Shuster *et al.* (2019). Since we use a subset of the original PERSONALITY-CAPTIONS dataset, we retrain the method outlined by Shuster *et al.* using similar settings. We compare the performance of our 3M model against their model using BLEU (Papineni *et al.* 2002), ROUGE-L (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (Anderson *et al.* 2016). The comparison results are listed in Table 1.

To evaluate the extensibility of our model, we also applied our method on the FlickrStyle10K dataset. This is meant to evaluate how well our method can generate captions that capture linguistic style. We compare against the following state-of-the-art baselines:

- StyleNet (2017), a single style model trained with paired factual sentences and unpaired stylized captions.

Method	style	training method	BLEU1	BLEU3	Meteor	CIDEr	ppl	cls
SF-LSTM (2018)	romantic	single-style	27.8	8.2	11.2	37.5	-	-
SF-LSTM (2018)	humorous	single-style	27.4	8.5	11.0	39.5	-	-
StyleNet (2017)	romantic	single-style	13.3	1.5	4.5	7.2	52.9	37.8
StyleNet (2017)	humorous	single-style	13.4	0.9	4.3	11.3	48.1	41.9
MsCap (2019)	romantic	multi-style	17.0	2.0	5.4	10.1	20.4	88.7
MsCap (2019)	humorous	multi-style	16.3	1.9	5.3	15.2	22.7	91.3
MemCap (2020)	romantic	multi-style	19.7	4.0	7.7	19.7	19.7	91.7
MemCap (2020)	humorous	multi-style	19.8	4.0	7.2	18.5	17.0	97.1
3M	romantic	multi-style	25.6	6.7	10.1	29.3	8.33	92.8
3M	humorous	multi-style	25.5	6.7	10.0	28.4	7.29	95.3

Table 3: Performance of Generative Models on FlickrStyle10K Dataset. Note: Due to only 7K out of 10K dataset publicly available, all the result are reported based on 7k data. All results except 3M are referred from paper (2020).

- SF-LSTM (2018), a single style model trained with paired stylized caption and paired factual captions.
- MsCap (2019), a multi-style model trained with paired factual sentences and unpaired stylized captions.
- MemCap (2020), a multi-style model trained with paired factual sentences and unpaired stylized captions.

Following (Zhao, Wu, and Zhang 2020), on FlickrStyle10K, we trained a logistic regression classifier for style classification and a pretrained language model using SRILM toolkit (2002) to measure perplexity. We report BLEU, Meteor (Banerjee and Lavie 2005), CIDEr, the style classification accuracy (cls) and the average perplexity (ppl) for comparison and results are showed in Table 3.

Ablation Study Additionally, to evaluate the benefits of each component of our model, we perform an ablation study using the PERSONALITY-CAPTIONS dataset. We compare the full 3M model against the following variations: no personality features, no text features, and no ResNeXt features. BLEU, ROUGE-L, CIDEr, and SPICE are reported in Table 2 for evaluating the relevance between image and generations. we also report the number of unique words used across all generated captions per model in Table 2 to show the expressiveness of each generative model.

Qualitative Analysis

Specifically, we seek to explain that our model is capable of generating captions that match the given style as well as the image context. We first list the given image and five given dense captions, sample generations along with personality in the parenthesis, in Figure 3 as R1-R3. We discuss the whether caption generations matching the context in three aspects: 1. whether the multi-style component working for connecting caption generations with given personality; 2. whether valid text features could help for generations to match the image; 3. whether ResNext feature could help make reasonable generations when the given text features fails to connect with the image. To give a more complete view of the text that our model can generate, we also list the imperfect sample generations underlined in Figure 3 as W1-W2.

Results and Discussion

In this section, we will outline the results of our experiments and illustrate them in both quantitative and qualitative ways.

Quantitative Analysis

Comparison with baselines As seen in Table 1, our 3M model outperforms UPDOWN models under the same training method across all the NLP metrics we used for evaluation. We also achieve better results on ROUGE-L, CIDEr compared with Shuster’s model trained under reinforcement learning. This provides evidence that our approach is effective at multi-style caption generation.

We also show that our 3M model does well on linguistic style captioning even though it was not designed for that task. As Table 3 shows, our 3M model significantly outperforms two other multi-style models, MsCap (2019) and MemCap (2020) on BLEU, CIDEr, Meteor, and ppl on the FlickrStyle10K dataset. Note that our 3M model also achieved high cls values, which show how well our captions capture the given style.

We also achieve comparable performance to the SF-LSTM model across the automated metrics we examined. Given that the SF-LSTM model is designed for a single-style generation task, whereas our 3M model was designed for multi-style generation, we feel that this shows how robust our model is.

Ablation Study From Table 2, we can see if our model is trained without the multi-style component, the performance of all the nlp metrics drops, proving how critical this component is. Examining the results obtained from a model using only text features against a model that only had access to ResNeXt features shows that using only text features limits the overall expressiveness of generated captions as shown by the low number of unique words generated.

Our full model has achieved the highest ROUGE-L, CIDEr and SPICE score and improves expressiveness compared with model with only text features and improves the relevancy compared to a model with only Resnext features.

Qualitative Analysis

For our qualitative analysis, we will discuss the quality of the trained 3M models across two datasets assessing whether our model is capable of generating captions that match the

given style and image context, and assessing whether our model can assist in finding reasons for imperfect captions.

From all generations in Figure 3, we can see our 3M model is able to generate captions matching the given personality, which certify that our multi-style component is able to help direct the generations in the desired personality tone. From R2-R3 we can see that when there is a valid text feature available, the 3M model could make use of them. The generation in R1 is expressed in a more conservative and global way since text features cannot provide correct information, which necessitates the use of ResNext features.

One of the advantages of the 3M model is that it can easily generate multiple captions with different styles. This can enable us to better contextualize incomplete or erroneous captions. In W1 of Figure 3, for example, the generation appear incomplete for the “Anxious” personality. Looking at captions for other personalities, we see that our model can correctly identify image context. This leads us to believe that we simply set the caption length too low for the “anxious” example. In W2, our model generates the incorrect phrases “a bike on a bike.” By examining the text features used for generation, we can see that this was likely caused by our input text, and not the model itself.

Conclusion

In this paper we introduce the 3M model, which is a multi-style image captioner which integrates multi-modal features and a multi-UPDOWN encoder-decoder model. We demonstrate the effectiveness of our 3M model by comparing against state-of-the-art work using automatic evaluation methods. Ablation studies have also be done to evaluate the contributions of each component of our 3M model. And we certify that our 3M model could generate more expressive and diverse generations without losing the connection with context. The qualitative study helps understand how well our 3M performs and shows how our model can also explain the imperfectness of generations.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 382–398. Springer.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Banerjee, S., and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics.

Chen, T.; Zhang, Z.; You, Q.; Fang, C.; Wang, Z.; Jin, H.; and Luo, J. 2018. “factual”or“emotional”: Stylized image

captioning with adaptive learning and attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 519–535.

Gan, C.; Gan, Z.; He, X.; Gao, J.; and Deng, L. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3137–3146.

Guo, L.; Liu, J.; Yao, P.; Li, J.; and Lu, H. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4204–4213.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Johnson, J.; Karpathy, A.; and Fei-Fei, L. 2016. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Shuster, K.; Humeau, S.; Hu, H.; Bordes, A.; and Weston, J. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12516–12526.

Stolcke, A. 2002. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM* 59(2):64–73.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

Zhao, W.; Wu, X.; and Zhang, X. 2020. Memcap: Memorizing style knowledge for image captioning. In *AAAI*, 12984–12992.