Netizen-Style Commenting on Fashion Photos: Dataset and Diversity Measures

Wen Hua Lin, Kuan-Ting Chen, Hung Yueh Chiang and Winston Hsu
National Taiwan University, Taipei, Taiwan
q868686qq@gmail.com, ktchen@cmlab.csie.ntu.edu.tw, kenny5312012@gmail.com, whsu@ntu.edu.tw

ABSTRACT

Recently, deep neural network models have achieved promising results in image captioning task. Yet, "vanilla" sentences, only describing shallow appearances (e.g., types, colors), generated by current works are not satisfied netizen style resulting in lacking engagements, contexts, and user intentions. To tackle this problem, we propose *Netizen Style Commenting (NSC)*, to automatically generate characteristic comments to a user-contributed fashion photo. We are devoted to modulating the comments in a vivid "netizen" style which reflects the culture in a designated social community and hopes to facilitate more engagement with users. In this work, we design a novel framework that consists of three major components: (1) We construct a large-scale clothing dataset named NetiLook, which contains 300K posts (photos) with 5M comments to discover netizen-style comments. (2) We propose three unique measures to estimate the diversity of comments. (3) We bring diversity by marrying topic models with neural networks to make up the insufficiency of conventional image captioning works. Experimenting over Flickr30k and our NetiLook datasets, we demonstrate our proposed approaches benefit fashion photo commenting and improve image captioning tasks both in accuracy and diversity.

CCS CONCEPTS

• **Information systems → Document topic models**; • **Computing methodologies → Natural language generation**; **Image representations**;

KEYWORDS

Fashion; Image Captioning; Commenting; Diversity; Deep Learning; Topic Model

ACM Reference Format:
Wen Hua Lin, Kuan-Ting Chen, Hung Yueh Chiang and Winston Hsu. 2018. Netizen-Style Commenting on Fashion Photos: Dataset and Diversity Measures. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3184558.3186354

1 INTRODUCTION

In accordance with [30], fashion has a vital impact on our society because clothing typically reflects a person's social status. This is also expected in the growing online retail sales, reaching 529 billion dollars in the US, and 302 billion euros in Europe by 2018 [9]. Still today, people either wear up their new clothes or upload their new clothing photo on social media to receive comments of new clothes. However, dressing inappropriately sometimes causes embarrassing. Therefore, people tend to know whether they dress properly beforehand. As the promising results achieved by image captioning, the problem could be solved by fashion image captioning works, which automatically describe the outfit with netizen-like comments.

Whereas, image captioning [7][8][14][24][33][35] is still a challenging and under researching topic despite deep learning developing rapidly in recent years. To generate a human-like captioning, machines not only recognize objects in an image but express their relationships in natural language, such as English. Large corpora of paired images and descriptions, such as MS COCO [17] and Flickr30k [28] are proposed to address the problem. Several deep recurrent neural network models are devised to follow the datasets and reach promising results. However, modern methods only focus on optimizing metrics used in machine translation, which causes absence of diversity — producing conservative sentences. These sentences can achieve good scores in machine translation metrics

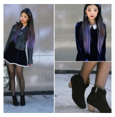
	(a) **Human**: wow so stunning!! love this look, and u r so pretty ♥
	(b) **CaptionBot**: A woman is laying on the grass.
	(c) **NC**: love the skirt
	(d) **Attention**: love the shoes
	(e) **NSC**: Love this button up bustier! It's fabulous! This look is amazing.
	(a) **Human**: Wow! Perfection!
	(b) **CaptionBot**: A group of people standing around each other.
	(c) **NC**: love the blazer!!!!!
	(d) **Attention**: love the shoes
	(e) **NSC**: I love your outfits. This one isn't an exception <3
	(a) **Human**: love ur ombre hair <3333 +1
	(b) **CaptionBot**: A group of women standing next to a woman.
	(c) **NC**: love the dress
	(d) **Attention**: love the shoes
	(e) **NSC**: I love the combinations :)) My heart for today goes to you! :)

Figure 1: Five sentences for each image from distinct commenting (captioning) methods. (a) One of the users' comments (i.e., ground truth) randomly picked from the post (photo) in our collected NetiLook dataset. (b) The sentences from Microsoft CaptionBot. (c) The results from neural image caption generation (NC) [33] (d) The results from neural image caption generation with visual attention (Attention) [35]. (e) Our proposed NSC. It marries style-weight to achieve vivid netizen style results.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186354

but are short of humanity. Compared with human comments as shown in Figure 1 (a), due to the limitation of training data, current methods (e.g., Figure 1 (b)) merely describe "vanilla" sentences with low utilities, which are merely describing the shallow and apparent appearances (e.g., color, types) in photos and generate meaningless bot tokens to users — lacking engagement, contexts, and feedbacks for user intentions, especially in the circumstances of online social media.

In order to generate human-like online comments (e.g, clothing style) for fashion photos, we collect a large corpus of paired user-contributed fashion photos and comments, called NetiLook, from an online clothing style community. To the best of our knowledge, our collected NetiLook is the largest fashion comment dataset. In our experiment on NetiLook, we found that these methods overfit to a general pattern, which makes captioning results insipid and banal (e.g., "love the ...") (cf., Figure 1 (c) and (d)). Therefore, to compensate for the deficiency, we propose integrating latent topic models with state-of-the-art methods and make the generated sentences vivacious (cf., Figure 1 (e)). Besides, for evaluating diversity, we propose three novel measures to quantize variety.

For richness and diversity in text content, we propose a novel method to automatically generate characteristic fashion photo comments for user-contributed fashion photos by marrying *style-weight* (cf., Section 4.2) from topic discovery models (i.e., latent Dirichlet allocation (LDA) [2]) with neural networks to achieve diverse comments with vivid "netizen" style. We look forward the breakthrough will foster further applications in social media, online customer services, e-commerce, chatbot developments, etc. It will be more exciting, in the very near future; for example, if the solution can work as an agent (or expert) in a living room and can comment for a user as testing the outfit in front of the mirror. To sum up, our main contributions are as follows:

- To our best knowledge, this is the first work to address the diverse measures of photo captioning in a large-scale fashion commenting dataset (cf., Section 1-2).
- We collect a brand new large-scale clothing dataset, NetiLook, which contains 300K posts (photos) with 5M comments (cf., Section 3).
- We investigate the diversity of clothing captioning and propose three measures to estimate the diversity (cf., Section 5).
- We leverage and investigate the merit of latent topic models, which is able to make up the insufficiency of conventional image captioning works (cf., Section 4).
- We demonstrate that our proposed approach significantly benefits fashion photo commenting and improves image captioning task both in accuracy and diversity over Flickr30k and NetiLook datasets (cf., Section 6).

2 RELATED WORK

Image captioning which automatically describes the content of an image with properly formed sentences enables many important applications such as helping visually impaired users and human-robot interaction. According to [1][31], a CNN-RNN framework, taking high-level features extracted from a deep convolution neural network (CNN) as an input for a recurrent neural network (RNN) to generate a complete sentence in natural language, has performed promisingly in image captioning tasks during the last few years. For example, [33] is an end-to-end CNN model followed by language generation of RNN. It was able to produce a grammatically correct sentence in natural language from an input image.

Following CNN-RNN frameworks, attention-based models ([35], [23]) were proposed. As human beings put different attentions at distinct objects while watching a photograph, attention-based models allow machines to put diverse weights on salient features. Compared with taking high-level representations of a whole image as input, attention-based models are able to dynamically weight various parts of images. Especially, when a lot of objects appear in an image, attention-based models can give more thorough captions [35].

Presently, state-of-the-art works are majorly attention-based models ([25], [16]) because they focus on correctness of descriptions. [5] assigned different weights to different words for fixing misrecognition. [18] focused on evaluating the correctness of attention in neural image captioning models.

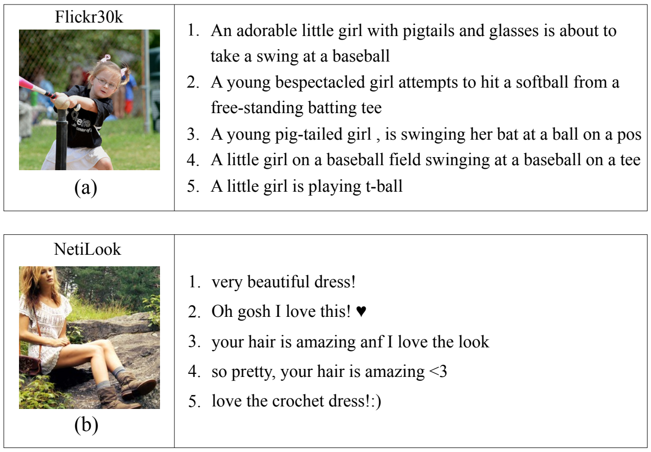
While applying current methods to generate comments, the demand for diversity is unveiled. Compared with depicting images, giving comments is more challenging because it needs to not only understand images but take care of engagement with users. To generate vivid comments, diversity is necessary. Besides commenting, diversity is also important in other areas (e.g., information retrieval [3]). In [13], to increase the utility of automatically generated response options of email, diversity is essential. Moreover, in building general-purpose conversation agents, which are required for intelligent agents' interaction with humans in natural language, diversity is also requisite. Therefore, we blend topic models with conventional methods to complement the diversity part of them.

Meanwhile, there has been increasing interest in clothing product analysis from the computer vision and multimedia communities. Most existing fashion analysis works focused on the investigation of the clothing attributes, such as clothing parsing ([21], [19], [22]), fashion trend ([11], [4]) and clothing retrieval ([10], [20]). In contrast to other works, we develop a novel framework that can leverage the learned Netizen-style embedding for commenting on fashion photos. Moreover, to our best knowledge, this is the first work to address the diverse measures of photo captioning in an all-time large-scale fashion commenting dataset. We detail the dataset and our method in the following sections.

3 DATASET — NETILOOK

[1] mentioned that current captioning datasets are relatively small compared with object recognition datasets, such as ImageNet [6]. Besides, the descriptions require costly manual annotation. With the growing of social media, such as Facebook and Instagram, people constantly share their life with the world. Consequently, these are all potentially valuable training data for image captioning (or commenting). Among social platforms, there are some specific websites just for clothing style. Lookbook[1], an example shown in Figure 3, is an online clothing style community where members share their personal style and draw fashion inspiration from each other. Such a rich and engaging social medium is potential to benefit

[1] lookbook.nu



Flickr30k

(a)

1. An adorable little girl with pigtails and glasses is about to take a swing at a baseball
2. A young bespectacled girl attempts to hit a softball from a free-standing batting tee
3. A young pig-tailed girl , is swinging her bat at a ball on a pos
4. A little girl on a baseball field swinging at a baseball on a tee
5. A little girl is playing t-ball

NetiLook

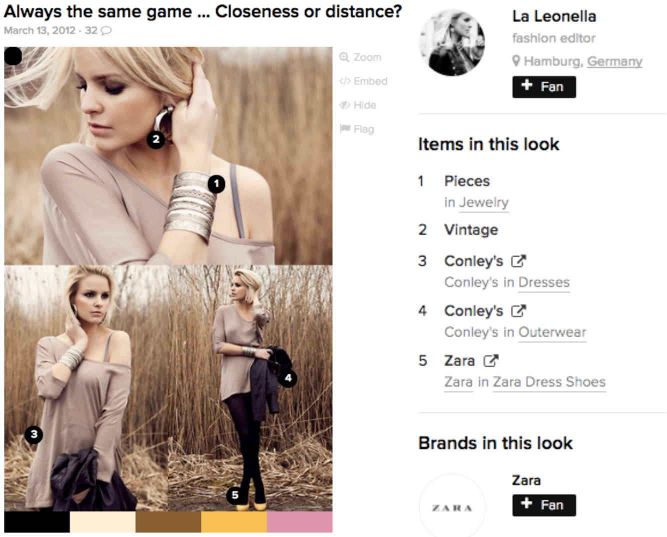
(b)

1. very beautiful dress!
2. Oh gosh I love this! ♥
3. your hair is amazing anf I love the look
4. so pretty, your hair is amazing <3
5. love the crochet dress!:)

Figure 2: Examples from Flickr30k and our NetiLook. (a) Most sentences are describing the shallow appearances (e.g., types, colors) and have similar sentence patterns (e.g., "A little girl ..."). (b) The sentences involve diverse user intentions with abundant styles. Furthermore, emojis and emoticons inside make it much more intimate with people.

Table 1: Comparison with other image captioning benchmarks (Flickr30k [28] and MS COCO [17]). Our collected dataset, Netilook, has the most diverse and realistic sentences in the social media (e.g., largest unique words)

Dataset	Images	Sentences	Average Length	Unique Words
Flickr30k	30K	150K	13.39	23,461
MS COCO	200K	1M	10.46	54,231
NetiLook	350K	5M	3.75	597,629



Always the same game ... Closeness or distance?

March 13, 2012 · 32

La Leonella

fashion editor

Hamburg, Germany

+ Fan

Items in this look

1 Pieces
in Jewelry

2 Vintage

3 Conley's
Conley's in Dresses

4 Conley's
Conley's in Outerwear

5 Zara
Zara in Zara Dress Shoes

Brands in this look

Zara

+ Fan

Figure 3: An example to show the attributes of a post in Lookbook. The post includes a title named by the author, country of the author, a publishing date, names, brands, and pantone of the clothes.

intelligent and human-like commenting applications. Hence, we collected a large corpus of paired user-contributed fashion photos and comments from Lookbook called *NetiLook*.

NetiLook[2]: To the best of our knowledge, this is the first and the largest netizen-style commenting dataset. It contains 355,205 images from 11,034 users and 5 million associated comments collected from Lookbook. As the examples shown in Figure 1, most of the images are fashion photos in various angles of views, distinct filters and different styles of collage. As Figure 2 (b) shows, each image is paired with (diverse) user comments. The maximum number of comments is 427 and the average number of comments is 14 per image in our dataset. Note that we observe that there are 7% of images with no comments and we remove these images in our training stage. Besides, each post has a title named by an author, a publishing date and the number of hearts given by other user. Moreover, some users add names, brands, pantone of the clothes, and stores where they bought the clothes. Furthermore, we collect the authors' public information. Some of them contain age, gender, country and the number of fans (cf., Figure 3). We believe all of these are valuable to boost the domain of fashion photo commenting. In this paper, we only use the comments and the photos from our dataset. Other attributes can be used to refine the system in future work. For comparing the results on Flickr30k, we also sampled 28,000 for training, 1,000 for validation and 1,000 for testing. Besides, we also sampled five comments for each image.

Compared to general image captioning datasets such as Flickr30k [28], the data from social media are quite noisy, full of emojis, emoticons, slang and much shorter (cf., Figure 2 (b) and Table 1), which makes generating a vivid "netizen" style comment much more challenging. Moreover, plenty of photos are in different styles of collage (cf., photos in Figure 1). Therefore, it makes the image features much more noisy than single view photos. To completely generate comments that entirely reflect the culture in social media, we demonstrate our method in the following section.

4 METHOD – NETIZEN STYLE COMMENTING

In NetiLook, we observed that user comments are much more diverse while comparing them with the sentences in general image captioning datasets. In addition, there are some frequently used sentences along with posts (e.g., "love this!", "nice") which cause current models inclined to generate similar sentences. The output comments become meaningless and insipid. To immerse the model in vivid netizen style, we fuse style-weight from topic models to image captioning models in order to keep long-range dependencies and take different comments from distinct points of view as topics.

4.1 Image Captioning

We follow [33] to extract image features from an image I by a CNN and feed it into the image captioning model at $t = -1$ to inform a LSTM (cf., CNN in Figure 4). We extract the FC7 (a fully-connected

[2] https://mashyu.github.io/NSC

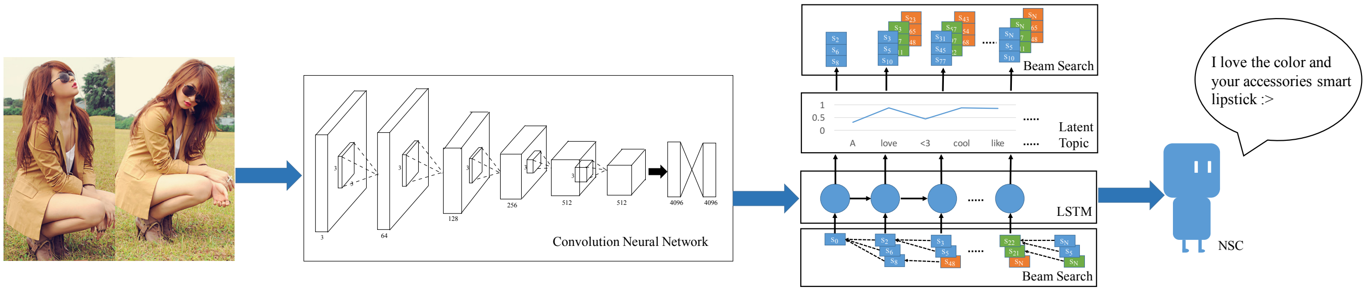


Figure 4: An illustration of our proposed framework. Our system consists of LSTM (cf., Section 4.1), topic models (cf., Section 4.2) and beam search to boost results (cf., Section 6.1). Our proposed approach leverages the outputs of image captioning model based on CNN-RNN frameworks and style-weight from LDA to generate a diverse comment with vivid "netizen" style.

layer) features as high-level meaning of the image from and feed it into the LSTM.

$$\mathbf{x}_{-1} = \mathrm{CNN}(\mathrm{I}). \tag{1}$$

We represent a word as a one-hot vector $\mathbf{s}$ of dimension equal to the size of dictionary. T is the maximum length of output sentences. We represent word embeddings as W_e.

$$\mathbf{x}_t = W_e \mathbf{s}, \ t \in 0 \ldots T-1. \tag{2}$$

With the CNN features, we can obtain probabilities of words in each generating step from the image captioning model. Sentences from general image captioning dataset basically depict common content of images. Therefore, conventional image captioning models are able to focus on accuracy. Nevertheless, to strike a balance between accuracy and diversity in current frameworks is arduous. To keep the merit of conventional models, we modify the generating processes of modern models with topic models and make outputs diverse while facing vivid netizen comments.

4.2 Style Captioning

To consider vivid netizen style comments, we introduce style-weight $\mathbf{w}_{style}$ element-wised multiplied ($\circ$) with outputs at each step of LSTM to season generated sentences.

$$\mathbf{p}_{t+1} = \mathrm{Softmax}(\mathrm{LSTM}(\mathbf{x}_t)) \circ \mathbf{w}_{style}, \ t \in 0 \ldots T-1. \tag{3}$$

Style-weight $\mathbf{w}_{style}$ represents the comment style, which teaches models to be acquainted with style in the corpus while generating captioning.

However, abstract concepts are hard for people to give a specific definition. To obtain the comment style in NetiLook, we apply LDA to discover latent topics and fuse with current captioning models.

Suppose, a corpus contains M comments. Comments are composed of a subset of N words. We specify K (K topics (z_1, z_2, ..., z_K)) for LDA. It gives N dimensional topic-word vectors and K dimensional comment-topic vectors.

Topic-word vectors: Each topic z has a probabilistic vector of N words in dictionary. The vector describes the word distribution of the topic. The topic-word vector ϕ_z of topic z is

$$\phi_{\mathbf{z}} = \{P(w_1|z), P(w_2|z), \ldots, P(w_N|z)\}. \tag{4}$$

where w_1, w_2, ..., w_N are N words in dictionary.

Comment-topic vectors: Each comment m is also associated with a probabilistic vector of topics, which means the topics probability of the comment. The comment-topic vectors θ_m of comment m is

$$\theta_{\mathbf{m}} = \{P(z_1|m), P(z_2|m), \ldots, P(z_k|m)\}. \tag{5}$$

where z_1, z_2, ..., z_K are different K topics. To find the topic distribution in corpus, each comment votes the topic with highest probability by $\arg\max(\theta_m)$. $\mathbf{t}_m^i$ is the i-th dimension of $\mathbf{t}_m$. In our finding, the voting gives the most characteristic style in the corpus. The mathematical notation can be represented as follow:

$$Let \ \mathbf{t}_m^i = \{ \begin{array}{l} 1 \text{ if } i = \arg\max(\theta_{\mathbf{m}}) \\ 0 \text{ otherswise} \end{array} . \tag{6}$$

The topic distribution of the corpus $\mathbf{y}$ now can be computed by normalizing the summation of the number of topics from $\mathbf{t}_m$ by the total number of comments. It means the proportion of various points of view of comments in the corpus:

$$\mathbf{y} = \sum_{m=1}^{M} \mathbf{t}_m / M. \tag{7}$$

With the topic distribution of corpus $\mathbf{y}$ and topic-word vectors ϕ, our style-weight $\mathbf{w}_{style}$ is now defined as:

$$\mathbf{w}_{style} = \sum_{k=1}^{K} \mathbf{y}^k \phi_k. \tag{8}$$

where $\mathbf{y}^k$ means the k-th dimension of $\mathbf{y}$.

As we embed style-weight in Equation (3), which could guide the generating process to select words that are much closer to the netizen style learned in the social media (e.g., we observe that one style-wight highlights emoji style), LSTM is capable to generate the sentences with the style in corpus. (cf., Latent Topic in Figure 4).

5 DIVERSITY MEASURES

Since BLEU and METEOR are not for diversity measure, diversity measures are being put importance on sentence generation models. Currently, [15] and [32] report the degree of diversity by calculating the number of distinct words in generated responses scaled by the total number of generated tokens. However, this is not enough for diverse comments from the Internet, since comments can be represented not only in natural language but in various sentence patterns, such as emojis, and emoticons. Therefore, to compensate the defects of BLEU and METEOR, we propose three novel measures to judge the diversity of comments generated from captioning models.

We observed that more diverse sentences are generated, more unique words are used. Thus we devise an intuitive and trivial unique words measure, called DicRate.

DicRate: The dictionary rate we proposed in this paper is measured through counting number of unique words among generated sentences divided by unique words among ground truth sentences. The number of unique words in ground truth sentences is N_t. The number of unique words in generated sentences is N_g. The DicRate is computed as follow:

$$\mathrm{DicRate}(N_t, N_g) = N_g / N_t. \tag{9}$$

DicRate reflects the abundance of vocabulary of a model, but it is still not incapable to measure sentence diversity. Inspired by the paper [29] for conversation response generation, two novel measures based on entropy are carried out to judge the diversity of comments on fashion photos. Descriptions of the measures are as follows:

WF-KL: The Kullback-Leibler divergence (KL divergence) of word frequency distribution between ground truth sentences and generated sentences. It shows how well a model learned the tendency of choosing words in a dataset. The number of unique words in the dataset is N. The occurrence times of each word in ground truth sentences are $\mathbf{w}_t$. The word frequency distribution of ground truth sentences is $\mathbf{w}_{ft}$. The occurrence of each word in generated sentences are $\mathbf{w}_g$. The word frequency distribution of generated sentences is $\mathbf{w}_{fg}$. By referring to the formula of term frequency-inverse document frequency (tf-idf), to avoid division by zero, we add one to $\mathbf{w}_t$ and $\mathbf{w}_g$. $\mathbf{w}^i$ is the i-th dimension of $\mathbf{w}$.

$$\mathbf{w}^i_{ft} = (\mathbf{w}^i_t + 1) / \sum_{i=1}^{N} (\mathbf{w}^i_t + 1). \tag{10}$$

$$\mathbf{w}^i_{fg} = (\mathbf{w}^i_g + 1) / \sum_{i=1}^{N} (\mathbf{w}^i_g + 1). \tag{11}$$

The WF-KL can be computed as follow:

$$\text{WF-KL}(\mathbf{w}_{ft}, \mathbf{w}_{fg}) = \sum_{i=1}^{N} \mathbf{w}^i_{ft} \log(\mathbf{w}^i_{ft} / \mathbf{w}^i_{fg}). \tag{12}$$

POS-KL: The KL divergence of part-of-speech (POS) tagging frequency distribution between ground truth sentences and generated sentences. POS is a classic natural language processing task. One of the applications is identifying which spans of text are products in user search queries [12]. Besides word distribution, POS also demonstrates the interaction between words in a sentence. The number of unique tags in the dataset is N. The occurrence times of each tag in ground truth sentences are $\mathbf{t}_t$. The tag frequency distribution of ground truth sentences is $\mathbf{t}_{ft}$. The occurrence times of each tag in generated sentences are $\mathbf{t}_g$. The tag frequency distribution of generated sentences is $\mathbf{t}_{fg}$. To avoid division by zero, we also add one to $\mathbf{t}_t$ and $\mathbf{t}_g$. $\mathbf{t}^i$ is the i-th dimension of $\mathbf{t}$.

$$\mathbf{t}^i_{ft} = (\mathbf{t}^i_t + 1) / \sum_{i=1}^{N} (\mathbf{t}^i_t + 1). \tag{13}$$

$$\mathbf{t}^i_{fg} = (\mathbf{t}^i_g + 1) / \sum_{i=1}^{N} (\mathbf{t}^i_g + 1). \tag{14}$$

The POS-KL can be computed as follow:

$$\text{POS-KL}(\mathbf{t}_{ft}, \mathbf{t}_{fg}) = \sum_{i=1}^{N} \mathbf{t}^i_{ft} \log(\mathbf{t}^i_{ft} / \mathbf{t}^i_{fg}). \tag{15}$$

6 EXPERIMENTS

6.1 Experiment Setting

To our best knowledge, this is the first captioning method that focuses on corpus style and sentence diversity. Generally, current methods are devoted to optimizing machine translation scores. Therefore, we only choose two famous captioning methods for comparison rather than other state-of-the-art methods (e.g., [1], [34]). To demonstrate the improvement of diversity, we apply our style-weight to our baselines.

Dataset: Note that we only adopt Flick30k for our experiments to compare with NetiLook because of the characteristic of Flick30k that mainly depicts humans, which is closer to NetiLook. Additionally, images in Flick30k and NetiLook are all collected from social media, which makes images in a similar domain.

Pre-processing: We argue that the learning process should be autonomous and leverage the freely and hugely available online social media. To avoid noise, we follow [33] to remove the sentences that contain a word frequency that is less than five times in training set. We also filter the sentences that are more than 20 words in the dataset to reduce advertisement and also make sentence more readable [26]. Noted that in order to thoroughly convey users' intention and comment style, we do not remove any punctuation in sentences.

Evaluation: BLEU and METEOR are conventional machine translation scores, which base on the matching of answers without considering diversity. The difference between BLEU and METEOR is that METEOR can handle stemming and synonymy matching. In BLEU scores, we report it in 4-grams because it has the highest correlation with humans [27]. For BLEU and METEOR, the higher scores mean that sentences are much correct according to the matching with ground truth. In our diversity measures, the higher DicRate shows that the more abundance of vocabulary of a model. Moreover, the lower WF-KL and POS-KL mean that the generated corpus is closer to the ground truth word distribution and sentence patterns.

Baseline: We duplicate two famous captioning methods (NC [33] and Attention [35]) in Table 2 and Table 3. NC is a CNN-RNN framework method that considers global features of images. Attention is an attention-based method, which puts distinct weights on salient features. By comparing NC with Attention in [35], BLEU and METEOR have similar relation like the result reported in Table 2. Our proposed method, NSC, fuses style-weight in the decoding stage. Following [33], we adopt beam search, an approximate inference algorithm, which is widely used in image captioning to boost the

Table 2: Performance on Flickr30k testing splits.

Method	BLUE-4	METEOR	WF-KL	POS-KL	DicRate
Human	0.108	0.235	1.090	0.013	0.664
NC	0.094	0.147	1.215	0.083	0.216
Attention	**0.121**	**0.148**	1.203	0.302	0.053
$\mathrm{NSC}_{\mathrm{NC}}$	0.089	0.146	1.217	**0.075**	**0.228**
$\mathrm{NSC}_{\mathrm{Attention}}$	0.119	**0.148**	**1.202**	0.319	0.055

Table 3: Performance on NetiLook testing splits.

Method	BLEU-4	METEOR	WF-KL	POS-KL	DicRate
Human	0.008	0.172	0.551	0.004	0.381
NC	0.013	0.151	0.665	1.126	0.036
Attention	0.020	0.133	**0.639**	1.629	0.011
$\mathrm{NSC}_{\mathrm{NC}}$	0.013	**0.172**	0.695	**0.376**	**0.072**
$\mathrm{NSC}_{\mathrm{Attention}}$	**0.030**	0.139	0.659	1.892	0.012

results. Because the number of possible sequences grows exponentially with the sentence length, beam search can explore generating process by spreading the most promising node in a limited set. We compare various beam sizes in our experiments and these methods get the best performance at the beam size of 3. Note that the optimal beam size might vary due to the properties of a dataset [13]. In our experiments, for LDA, analysis of the performance sensitivity is made by varying K from 1 to 15. For the first experiment on Flickr 30k, we set the number of topics to be 3 (K = 3); for the experiment on NetiLook, we have K = 5 in $\mathrm{NSC}_{\mathrm{NC}}$ and K = 3 in $\mathrm{NSC}_{\mathrm{Attention}}$. We observe that topic models can reflect some semantic "style" of comments (e.g. emoji style). Therefore, compared to Flickr 30k, more topic models are selected in NetiLook because user comments are much more diverse in this dataset. Interestingly, the proper number of topic models in $\mathrm{NSC}_{\mathrm{NC}}$ is higher than $\mathrm{NSC}_{\mathrm{Attention}}$. We observe that more topic models would not benefit the attention-based approach for the reason that attention-based models are greatly restricted the word selection.

6.2 Quantitative Analysis – Dataset

Traditional captioning dataset such as Flickr30k [28] and MS COCO [17] only focus on image description and do not emphasize on style and comment-like sentences. Therefore, we address the problem in the paper and contribute the dataset for brand-new problem definition. For comparing models with human and characteristics of datasets, we not only evaluate the generated sentences but also evaluate human comments. Also, as we can see differences from human evaluation between Table 2 and Table 3, the comparison does highlight the distinctions between Netilook and Flicr30k. For a comment given by a human or machine, it is difficult to be evaluated on conventional measures such as BLEU in NetiLook (e.g. 0.108 in Table 2 vs. 0.008 in Table 3 in BLEU-4). Thus, we propose our measures DicRate, WF-KL and POS-KL to evaluate comments.

In the scenario of online social media, punctuation, slang, emoticons and emojis are important for conveying emotion in a sentence. Thus, Netilook has much more diversity and unique words than other datasets (0.664 in Table 2 vs. 0.381 in Table 3 in DicRate). NetiLook specializes in describing clothing style as examples shown in Figure 5. Still, there are some common words and general patterns to describe and comment on the clothing style in comparison with Flickr30k, which mixes all types of images in the dataset. Thus NetiLook has lower score on WF-KL and POS-KL (e.g. 1.090 in Table 2 vs. 0.551 in Table 3 in WF-KL).

For such a diverse and characteristic dataset, machines are required considering overall corpus distribution and mimic comment style in order to get high performance in our evaluations. Nonetheless, for learning human commenting style, it is still challenging for general captioning models to generate diverse words while there are some general comments can achieve universally low loss (e.g., "nice", "I love this!"). However, our style-weight brings human style in machine generated sentences.

6.3 Quantitative Analysis – Model Evaluation

Table 2 summarizes performances for the Flickr30k dataset. Attention models put weights on salient features in images, thus the models easily describe objects inside pictures and reach a better BLEU and METEOR (e.g. 0.094 vs. 0.121 in BLEU-4). However, attention-based models are greatly restricted the word selection while decoding stage. In our experiments, POS-KL and DicRate are much worse (e.g. 0.053 vs. 0.216 in DicRate) comparing Attention with NC. With our style-weight, the model $\mathrm{NSC}_{\mathrm{NC}}$ expands the word diversity without sacrifice much on BLEU and METEOR. Style-weight encourages models choosing the words that are closer to the original distribution rather than the words that can generally get the lowest loss during the training phase. As we show in Table 2, DicRate and POS-KL are improved comparing $\mathrm{NSC}_{\mathrm{NC}}$ with NC (e.g. 0.216 vs. 0.228 in DicRate). The impact of style-weight is also shown in Attention model. However, we observed that embedding style-weight does not improve much in Flickr30k dataset in terms of diversity, because the sentences are objectively depicting humans performing various activities in Flickr30k.

In NetiLook, the experiment in Table 3 shows that our method can greatly improve the diversity. Comparing NC with Attention in Table 3, NC performs better than Attention (e.g. 0.036 vs. 0.011 in DicRate) except for BLEU-4 and WF-KL (e.g. 0.665 vs. 0.639 in WF-KL) because the selection of words is affected by salient features in an image, which makes the model miss the intention of the corpus while the whole dataset has similar objects. However, with style-weight, our $\mathrm{NSC}_{\mathrm{NC}}$ outperforms other baselines in POS-KL and DicRate (e.g. 0.376 of $\mathrm{NSC}_{\mathrm{NC}}$ in POS-KL). This proves that style-weight can guide the generating process to the comment that is much closer to the users' behaviour in the social media, making machine mimic online netizen comment style.

6.4 Image Commenting Results

We show some real examples of fashion commenting results on NetiLook with various methods. Though there are emojis generated from Microsoft CaptionBot, the comments are still lacking engagement and can not afford to process photos in collage. While training general captioning models (e.g., NC and Attention) on NetiLook, the comments are much shorter than Human's and fixed in some patterns, which lacks diversity.

Methods	(a)	(b)	(c)	(d)	(e)	(f)
Human	Gorgeous! Love pants, sunglasses and everything else! ;)	amazing look! Love everything in it!Hyped!	amazing picture :-)	great, man!!	cool	Cute top!
CaptionBot	a couple of women standing next to a woman and she seems 😁.	a group of people standing next to a woman in a black wet suit and they seem 😯😐😐.	a man is jumping in the air on a skateboard.	a group of people standing around each other.	a little girl walking down the street and they seem 😐😐.	a woman standing next to a brick wall and she seems 😐.
NC	I love your dress!	I love your shoes <3	I love your hair and look	nice look	nice	Love your dress!
Attention	I love your look	I love your shoes	I love the shoes	I love your shoes	I love your look	I love your shoes
Our proposed $\mathrm{NSC}_{\mathrm{NC}}$	I love the combinations :)) My heart for today goes to you! :)	I love your shoes!!! HYPED! <3	I love your hair and the coat, unbelievably gorgeous. Hyped!	I love your style!! :D	I love your hair	Love the dress!

Figure 5: Examples of comments generated by different methods. The examples show that our proposed approach $\mathrm{NSC}_{\mathrm{NC}}$ can help generate more diverse and vivid comments.

Similar intention like human: With the style-weight, $\mathrm{NSC}_{\mathrm{NC}}$ can generate the comment that is much closer to users' intention (cf., Figure 5 (a)).

More vivid comments: While conveying the same intention, $\mathrm{NSC}_{\mathrm{NC}}$ is able to use emoticons, punctuations and capitalizations to generate more netizen-style comments than other captioning models (cf., Figure 5 (b)).

Another point of view: By considering the topic distribution of data, $\mathrm{NSC}_{\mathrm{NC}}$ generates comments that are different from general captioning models' and much closer to human beings (cf., Figure 5 (c) - (e)).

Wrong objects: However, there are still some drawbacks in our $\mathrm{NSC}_{\mathrm{NC}}$, such as describing wrong objects in the images. Because the $\mathrm{NSC}_{\mathrm{NC}}$ is still based on image captioning models, it will also generate wrong comments as other captioning models due to the similarity of images (cf., Figure 5 (f)). It can be improved by jointly training the topic model with attention-based models.

6.5 User Study

Motivated by the paper [34] which conducts a human evaluation of image captioning by presenting images to three workers, we conducted a user study from 23 users to demonstrate the effect of diverse comments. The users are about 25 year-old and familiar with netizen style community and social media. The sex ratio in our user study is 2.83 males/female. They are asked to rank comments for 35 fashion photos. Each photo has 4 comments — from one randomly picked human comments, NC, Attention and our $\mathrm{NSC}_{\mathrm{NC}}$. Therefore, each of the users has to appraise 140 comments generated from different methods. Furthermore, we collect user feedback to understand user judgements on comments generated by different methods.

As Table 4 shows, 36.8% out of 805 votes (35 × 23) rank the sentences generated from $\mathrm{NSC}_{\mathrm{NC}}$ at the first place which outperforms NC and Attention. It means that our NSC defeats human comments in some images. Furthermore, The difference between humans and $\mathrm{NSC}_{\mathrm{NC}}$ in rank 1 is only 9.3%. In top two ranks, the performance of $\mathrm{NSC}_{\mathrm{NC}}$ reaches 76.6%. This also demonstrates that our $\mathrm{NSC}_{\mathrm{NC}}$ can generate sentences with human-like quality. In our user study, people generally regard our $\mathrm{NSC}_{\mathrm{NC}}$ sentences as human comments. According to our user study, the main concern of people's ranking is emoticons. Emoticons is an important component in the sentences to connect human emotions and also make sentences more vivid. For instance, Figure 5 (d) in the user study, the voting of $\mathrm{NSC}_{\mathrm{NC}}$ outperforms Human at the first rank (39.1% vs. 34.8%). Relevance between comments and images takes the second concern of people's ranking. Objects mentioned in sentences should not be trivial or mismatch in the photos. For example, Figure 5 (c), $\mathrm{NSC}_{\mathrm{NC}}$ captures the outfit (coat) and floating hair resulting in the same voting as human in rank 1 (39.1%) in the user study. To sum up, our style-weight makes captioning model mimic human style and generates human-like comments which most people agree with in our user study.

Table 4: Result of user study. $\mathrm{NSC}_{\mathrm{NC}}$'s comments are more likely to be regarded as human than other methods.

Ranking	Human	NC	Attention	$\mathrm{NSC}_{\mathrm{NC}}$
Rank 1	46.1%	10.8%	6.3%	36.8%
Rank 2	24.5%	21.4%	14.4%	39.8%
Rank 3	18.1%	31.9%	34.3%	15.7%
Rank 4	11.3%	35.9%	45.0%	7.8%

7 CONCLUSIONS

We present style-weight that greatly influences on current captioning models to immerse into human online society. Also, we contribute our dataset NetiLook, which is a brand new large-scale clothing dataset, to achieve netizen style commenting with style-weight. An image captioning model automatically generates characteristic comments for user-contributed fashion photos. NSC leverages the advantage of style-weight which can keep long-range dependencies to achieve vivid "netizen" style comments. Experiments on Flickr30k and NetiLook datasets, we demonstrate our proposed approaches benefit fashion photo commenting and improve image captioning task both in accuracy (quantified by conventional measures of image captioning) and diversity (quantified by our proposed measures). A user study is carried showing that our proposed idea can generate sentences with human-like quality. It is worth noting that our proposed approach can be applied on other fields (e.g., conversation response generation or question-answering system) to help generate sentences with various styles by the idea of style-weight. Moreover, NetiLook contains abundant attributes, researchers are able to use those attributes to build a more comprehensive system. For example, comments from different genders in future work. We believe that the integration of image captioning models, style-weight and the dataset proposed in this paper will have a great impact on related research domains.

8 ACKNOWLEDGEMENT

This work was supported in part by Microsoft Research Asia and the Ministry of Science and Technology, Taiwan, under Grant MOST 105-2218-E-002-032. We also benefit from the grants from NVIDIA and the NVIDIA DGX-1 AI Supercomputer and the discussions with Dr. Ruihua Song, Microsoft Research Asia.

REFERENCES

[1] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–10.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[3] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. 2012. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 6.

[4] KuanTing Chen, Kezhen Chen, Peizhong Cong, Winston H Hsu, and Jiebo Luo. 2015. Who are the devils wearing prada in new york city?. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 177–180.

[5] Minghai Chen, Guiguang Ding, Sicheng Zhao, Hui Chen, Qiang Liu, and Jungong Han. 2017. Reference Based LSTM for Image Captioning.. In *AAAI*. 3981–3987.

[6] Jia Deng, Alexander Berg, Kai Li, and Li Fei-Fei. 2010. What does classifying more than 10,000 image categories tell us? *ECCV 2010* (2010), 71–84.

[7] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[8] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1473–1482.

[9] Centre for Retail Research. 2017. Online Retailing: Britain, Europe, US and Canada 2017. www.retailresearch.org/onlineretailing.php. (2017). Accessed: 2018-01-25.

[10] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE International Conference on Computer Vision*. 3343–3351.

[11] Shintami C Hidayati, Kai-Lung Hua, Wen-Huang Cheng, and Shih-Wei Sun. 2014. What are the fashion trends in new york?. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 197–200.

[12] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[13] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, László Lukács, Marina Ganea, Peter Young, et al. 2016. Smart reply: Automated response suggestion for email. *arXiv preprint arXiv:1606.04870* (2016).

[14] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[15] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055* (2015).

[16] Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. 2017. Image Caption with Global-Local Attention.. In *AAAI*. 4133–4139.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755.

[18] Chenxi Liu, Junhua Mao, Fei Sha, and Alan L Yuille. 2017. Attention Correctness in Neural Image Captioning.. In *AAAI*. 4176–4182.

[19] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. 2014. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia* 16, 1 (2014), 253–265.

[20] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. 2012. Hi, magic closet, tell me what to wear!. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 619–628.

[21] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 3330–3337.

[22] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1096–1104.

[23] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. *arXiv preprint arXiv:1612.01887* (2016).

[24] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632* (2014).

[25] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2017. Text-Guided Attention Model for Image Captioning.. In *AAAI*. 4233–4239.

[26] Nirmaldasan. 2008. The Average Sentence Length. https://strainindex.wordpress.com/2008/07/28/the-average-sentence-length/. (July 2008). Accessed: 2017-04-06.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[28] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 139–147.

[29] Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating Long and Diverse Responses with Neural Conversation Models. *arXiv preprint arXiv:1701.03185* (2017).

[30] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. 2015. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 869–877.

[31] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[32] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *arXiv preprint arXiv:1610.02424* (2016).

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[34] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 988–997.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.. In *ICML*, Vol. 14. 77–81.