DrunaliaCap: Image Captioning for Drug-Related Paraphernalia With Deep Learning

BEIGENG ZHAO
Cyber Crime Investigation Department, Criminal Investigation Police University of China, Shenyang 110035, China
e-mail: zhaobeigeng@qq.com
This work was supported by Fundamental Research Funds for the Central Universities under Grant D2019023.

ABSTRACT Image captioning is a process of generating textual descriptions of images. In recent years, research on publicly available large-scale datasets and deep learning-based algorithms has promoted the development of this field. However, little research has been conducted on captioning images of drug-related paraphernalia that, despite being an important topic for both drug prevention and police enforcement, is not covered by existing image captioning studies. In this paper, we propose DrunaliaCap—a deep learning-based system for autogenerating both “factual” (what is in the image) and “functional” (the usage of each paraphernalia during drug-taking) descriptions of images of drug-related paraphernalia. We constructed a new dataset containing 20 categories of drug-related items and trained deep learning-based models for the proposed system. We further proposed a method to evaluate and optimize the generation of captions to prevent them from missing important knowledge. Experiments were conducted to validate the performance of the newly proposed dataset and method. We analyzed the experimental results and discussed the significance, limitations, and potential applications of our work.

INDEX TERMS Image captioning, drug prevention, dataset construction, deep learning.

I. INTRODUCTION

Image captioning is a challenging and cross-disciplinary process involving computer vision, natural language processing, and machine learning. An image captioning model takes an image as the input and generates an output of a paragraph of human-readable text describing the image. For humans, describing images is easy; however, this process is not easy for machines and involves two main challenges—accurate recognition of objects in images and generation of meaningful descriptions in natural language.

In recent years, many deep learning-based approaches have been used to solve the two challenges mentioned above. To effectively recognize the objects in images, computer vision technologies such as convolutional neural networks (CNNs) are often used as encoders to extract image features. To generate descriptions, techniques in the field of natural language processing, such as recurrent neural networks (RNNs), are used as decoders. The image features extracted by the encoders are “translated” into human-readable text by the decoders. To train and evaluate deep learning models, datasets containing images and text annotations are also needed.

Most image captioning studies focus on generating “factual” captions, including what objects are in the images, as well as the attributes and relationships of the objects. However, there is an interesting phenomenon in which people often use “non-factual” words when describing images. Some recent studies have been devoted to the generation of these non-factual captions for images, such as sentiment [26], writing style [21], and personality [27].

Increasingly more research has been devoted to constructing characteristic image caption datasets with quality annotations and designing efficient deep learning-based methods. However, to the best of our knowledge, little effort has been devoted to captioning photos of drug-related paraphernalia using deep learning methods, despite this being meaningful work with great application value.

The term “drug paraphernalia” refers to any equipment or accessory that is intended or modified for making, using, or concealing drugs. In many scenarios, photos of drug paraphernalia and corresponding text describing the photos exist. For example, in drug prevention publicity and education, not only are photos of drug paraphernalia often displayed to the audience, but the corresponding names, structures, and usage of the paraphernalia are also explained. Another example is that when drug-related paraphernalia are seized during police raids, photos of the seizures should be taken and...
attached with the corresponding reports along with descriptions. In addition, images containing drug paraphernalia and the accompanying text descriptions are often seen on social media platforms and e-commerce websites.

The contributions of this paper are as follows. First, we propose a new process with potential application value to autogenerate factual and functional captions for images of drug-related paraphernalia with deep learning. Second, we construct a new dataset, train deep learning models for the proposed process, and further propose a method to evaluate and optimize the generation of captions. Third, we conduct experiments to validate the performance of our proposed system and discuss the significance, limitations, future exploration and potential applications of our work.

II. RELATED WORK

A. IMAGE CAPTIONING DATASETS

A number of different datasets have been constructed for the training and evaluation of image captioning methods. Microsoft Common Objects in Context (MS-COCO) [1] is a large-scale dataset containing more than 300,000 images, over 2 million instances, and 80 object categories. Each image is annotated with five captions. Flickr30k [2] is a popular benchmark dataset for sentence-based image description that contains over 30,000 images collected from Flickr with more than 158,000 human-annotated captions. Flickr8K [3] is another widely used dataset consisting of 8,000 images chosen from 6 different Flickr groups. The images in this dataset are manually selected to depict a variety of situations and scenes. Because these 3 well-annotated datasets contain a wide range of everyday scenes and objects, they are popularly used as benchmarks for the evaluation of image captioning models.

Datasets built using other perspectives also promote image captioning research. Different from the abovementioned three datasets, the Visual Genome [6] dataset focuses on captioning multiple regions in a picture separately. Tran et al. [7] and Park et al. [8] created datasets using images and text from Instagram, which is a photo sharing social platform. IAPR TC-12 [9] contains captions in multiple languages. AIChallenger [10], COCO-CN [11], and STAIR [12] focus on captioning in Chinese and Japanese languages.

In addition to factual captions, some research studies have concentrated on creating non-factual image captioning datasets. Mathews et al. [26] created SentiCap for autogenerating captions containing emotions. This dataset contains more than 1000 images selected from MS-COCO. At least three positive and three negative sentences containing emotions were rewritten for each image via embedding adjective-noun pairs into the original MS-COCO annotations. FlickrStyle10K [21] is constructed using 10,000 images from Flickr30K. Each picture in this dataset is additionally annotated with romantic and humorous stylized sentences. Shuster et al. [27] annotated a large-scale dataset called Personality-Captions containing more than 200,000 images. Each image in this dataset is annotated with one caption conditioned on one of 215 different possible personality traits.

 Inspired by the success of these datasets in different aspects of image captioning, we constructed a new dataset that
B. Zhao: DrunaliaCap: Image Captioning for Drug-Related Paraphernalia With Deep Learning

focuses on the autogeneration of factual and functional captions for images of drug paraphernalia.

B. IMAGE CAPTIONING METHODS

Many image captioning models are based on encoder-decoder architectures. Vinyals et al. [13] first combined a CNN encoder and a long short-term memory (LSTM) decoder for image captioning. In the proposed CNN-LSTM model, the CNN encoder is utilized to extract image features. Then, the LSTM decoder “translates” the extracted features into human-readable sentences. Xu et al. [5] first introduced an attention mechanism into an encoder-decoder image captioning architecture. This proposed attention mechanism is different from Vinyals et al.’s method in that it can focus on a salient part of a picture and generate a corresponding word at each time step. Wang et al. [22] proposed a CNN+CNN-based image captioning method that uses a hierarchical attention module to connect a vision CNN with a language CNN. More recent studies [28]–[30] have focused on accurately aligning different levels of image features and text fragments via innovative model architectures and attention mechanisms.

To generate non-factual image captions, Mathews et al. [26] proposed a switching RNN model. In this model, two RNNs are trained and run in parallel to generate factual and sentimental words. Gan et al. [21] designed a model called StyleNet, in which the weight matrices in LSTM networks are decomposed into several factors that are used to generate factual and stylized captions. Chen et al. [31] proposed a variant version of LSTM called Style-Factual LSTM. In this model, two groups of matrices are trained to capture factual and stylized information, respectively. A set of weights are designed to control the proportion of each matrix.

Several recent works have focused on learning non-factual knowledge from an unpaired corpus. Chen et al. [32] proposed generating non-factual image captions with the Domain Layer Norm, which enables the generation of various stylized sentences. MSCap [33] was designed to generate multiple stylized descriptions by training a single captioning model on an unpaired non-factual corpus with the help of several auxiliary modules. Zhao et al. [34] proposed a new model named MemCap, which resorts to explicitly encoding non-factual knowledge by building a memory module.

In our work, we trained CNN-LSTM models to autogenerate factual and functional captions on the basis of Xu et al.’s method [5], which is effective and has been used as a baseline in many other studies. We further proposed a method to evaluate and optimize the generation process to prevent the generated captions from missing important knowledge of drug paraphernalia.

III. METHOD

A. DATASET CONSTRUCTION

Our dataset contains 820 images, of which 450 are in the public domain and were collected from the Internet, while the other 370 were obtained by taking photos of teaching props. Each image in the dataset contains a close-up of 1–3 drug items on a light gray or white background. If two or more items appear in one image, these items should be able to be used together during drug-taking. For example, a piece of tinfoil and a lighter are often used together to heat and inhale heroin or methamphetamine fumes, while a spoon and a syringe are usually used together to liquefy and inject drugs.

Figure 2 shows some representative examples of images in the dataset.

Altogether, the dataset contains 20 different categories of drug paraphernalia. The distribution of instances in each category is shown in Figure 3. The average number of instances in each category is 44.9. Bongs and pipes have the largest numbers of instances, with 93 and 82, respectively, owing to their various shapes and colors. Electronic cigarettes have the smallest number of instances, with 23.

Each picture is annotated with one factual sentence and one functional sentence. A factual sentence describes the item(s) in an image, including their visible colors, shapes, and materials. In addition to describing the items in an image, a functional sentence describes the possible usage of the item(s) during drug-taking.

Before annotation, the following instructions were given to each annotator to understand the possible usage of each item in the dataset:

![Figure 2](image_url)

**Figure 2.** Examples of images in the constructed dataset. (a) A piece of tin foil and a lighter. (b) A homemade bong out of a Sprite bottle and straws. (c) A syringe and a spoon. (d) A small mirror and a short straw. (e) A red “sexy lady” glass bong. (f) A hookah. (g) A glass pipe with a bowl structure at one end. (h) A cigar. (i) A metal roach clip.
• Cigars, rolling papers and pre-rolled cones: These items are usually used with marijuana, which is often obtained in a loose form and needs to be packed into something for smoking. Marijuana rolled into a cigar is called a blunt, and marijuana smoked in rolling paper is called a joint.

• Roach clips: Roach clips are usually used to hold a joint or blunt to prevent fingers from getting burned. Some roach clips are adorned with feathers or other decorations at the tail.

• E-cigarettes and vape pens: Nicotine and marijuana cartridges can also be smoked in an e-cigarette or a vape pen. Some evidence suggests that young people who use e-cigarettes are more likely to smoke traditional cigarettes in the future.

• Bongs, hookahs and homemade bongs: Bongs and hookahs are water pipes used as filtration devices for smoking cannabis, tobacco, or other herbal substances. They can also be used to smoke PCP (phenylcyclidine), crack cocaine, crystal methamphetamines, opium, or other powerful drugs such as psychedelic DMT (N,N-Dimethyltryptamine). In addition to commercially available bongs and hookahs, drug users usually make bongs at home using inexpensive supplies such as drink bottles, straws, and different kinds of pipes.

• Tin or aluminum foil and lighters: Tin or aluminum foil is sometimes used to heat up drugs such as heroin, crack cocaine, or methamphetamines to inhale the fumes. Drug users often place chopped-up drugs onto a piece of foil and hold it over a lighter until it starts emanating smoke.

• Pipes: Various kinds of pipes can suggest multiple types of drug abuse, including the use of marijuana, crack cocaine, heroin, and crystal methamphetamines. Pipes with a bulb at the end are usually used to smoke methamphetamines or crack cocaine, while marijuana pipes often have a bowl structure at the end. Straight, long glass tubes are often used to smoke crack cocaine.

• Straws, paper tubes and rolled-up currency: These tools are used by people who snort drugs in powdered forms, such as cocaine or heroin, directly through their nose.

• Mirrors, razor blades and playing cards: Mirrors are often used as chopping boards and smooth surfaces for cutting and snorting powdered drugs. Razor blades, playing cards, and other kinds of cards are used to cut cocaine or methamphetamine into lines to be snorted.

• Syringes: Many drugs can be dissolved in a liquid and injected directly into the body by using a syringe, including cocaine, heroin, prescription painkillers, and methamphetamines.

• Spoons: People who inject drugs often use spoons to help liquefy or dissolve drugs in crystallized forms. Spoons may also be used to bring a small amount of cocaine up to the nose to snort.

Three volunteers with professional knowledge of drug prevention participated in the annotation work. Each annotator was first asked to understand the common usage of each item during drug-taking according to the instructions given above. Then, images were randomly presented to each annotator for them to annotate factual and functional captions.

To construct an annotated corpus with various sentence structures and large word variability, the annotators were encouraged to use various sentence structures and content to describe a specific item. For example, for describing the function of a bong in drug-taking, the following sentences with different structures were presented to the annotators:

- The bubble base glass bong shown in the image can be used as a filtration device for smoking crack cocaine or methamphetamines through water.
- Drug addicts often use such a bubble base glass bong to smoke powerful drugs such as methamphetamines.
- The smoke of drugs can be filtered through a bubble base glass bong.

Thus, different sentences, including but not limited to any of the above, can be used to describe the function of a bong.

| Table 1. Number of words per sentence in factual and functional captions. |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                                | Factual         | Functional      |                  |                  |                  |                  |
|                                | Min  | Max  | Average | Min  | Max  | Average |                  |                  |
|                                | 2    | 13   | 8.3     | 8    | 28   | 15.4    |                  |                  |

Table 1 shows the statistics corresponding to the number of words per sentence in factual and functional captions.

To ensure the quality of the annotations, when the process of the annotation work reached 40% and 70% completion and before finalizing the annotations, we randomly checked 30% of the annotated sentences for each annotator. Both good and bad samples were then returned to the corresponding annotator for reannotation and an overall review. Representative

![instances per category](image.png)
B. CNN-LSTM CAPTIONING MODEL

A CNN-LSTM architecture is a commonly used method for deep learning-based image caption generation. CNNs, which can efficiently extract different levels of image features, were originally designed for image classification and object recognition tasks. LSTM is a variant of an RNN, is explicitly designed to avoid the long-term dependency problem, and is widely used in sequence-to-sequence machine translation. In a CNN-LSTM architecture, one or several layers of a CNN are used as an encoder to extract image features. An LSTM network is utilized as a decoder to transform the extracted image features into human-readable text. When generating each word, an attention mechanism can help the LSTM decoder focus on a specific region rather than the whole image.

As shown in Figure 5, our captioning model includes a CNN encoder, an LSTM decoder, and an attention mechanism. During training, the model takes a single image I and a sequence S of 1-of-V encoded caption words:

\[ S = \{w_1, \ldots, w_T\}, \quad w_t \in \mathbb{R}^V \]  

where \( w_t \) is a vector corresponding to the \( t \)-th encoded word in the annotated caption text, \( V \) is the size of the vocabulary, and \( T \) is the sentence length. Image feature \( F \) extracted by the CNN is processed by the attention mechanism and is then taken by the LSTM decoder as the input. The LSTM network generates a sequence of words \( \tilde{w}_t \). Loss values are then calculated according to the generated sequence and \( S \) for back-propagation. Once the model is trained, a sequence of predicted words can be generated using a single image.

To extract image features, we borrowed layers (from Conv1 to Res5c) from the ResNet-101 model [16] pretrained on the ImageNet classification task [17]. The extracted image features were flattened into a tensor \( F \) of size \( L \times C \), where \( L \) and \( C \) represent the spatial and channel dimensions, respectively.

A standard LSTM network [18] was employed to generate the caption words. Following the model proposed by Xu et al. [5], our RNN model contains a soft spatial attention mechanism. At each time step \( t \), a vector \( \alpha_t \) with the same
dimension as that of $F$ is calculated as follows:

$$e_{hi} = g(F_i, h_{i-1})$$

$$\alpha_{hi} = \frac{\exp(e_{hi})}{\sum_{k=1}^{L} \exp(e_{ki})}$$

where $F_i, i = 1, \ldots, L$ represents the features extracted from different positions of the image, and $\alpha_{hi}$ is the weight of $F_i$ used to generate the word $w_i$. The attention model $g$ is a multilayer perceptron that calculates the weight. $F$ is multiplied element-wise with $\alpha_{hi}$, and the product is input into the LSTM cell. The transition to the next hidden state can be denoted as

$$h_t = LSTM(\alpha_t, F, h_{t-1})$$

The LSTM network generates one word $\tilde{w}_t$ at each time step based on the soft image attention $\alpha_t = F$ and the previous generated hidden state $h_{t-1}$. Once the model is trained, sentences captioning an image can be generated word by word according to Equation 4.

### C. IMPORTANT KNOWLEDGE GENERATION

When annotating functional captions, different annotators tend to describe an image from different perspectives and with different vocabularies. In our proposed captioning task, some words and phrases contain important knowledge related to specific drug paraphernalia and are distributed in different sentences. For example, when describing the usage of cigars, words such as “blunt”, “marijuana”, “smoking”, and “cannabis” used by different annotators are distributed in different ground truth sentences. During training, the relationships between the “cigar” objects in the images and these important words distributed in different sentences are learned by the CNN-LSTM network described in Section III.B. However, if only one sentence is generated for an image, the generated sentence may contain only a portion of the important words and miss the others. To prevent missing important knowledge in the generated captions, we use a beam search algorithm to generate multiple functional captions for each image and a Dictionary of Important Knowledge (DIK) to help evaluate and choose the appropriate beam size.

Beam search [35] is a heuristic search algorithm widely used to decode output sequences from RNNs. At each time step, beam search stores the top-$K$ highest scoring partial solutions, where $K$ is known as the beam size. When generating the next word, all possible single-word extensions of existing solutions are considered, and the $K$ highest scoring extensions are retained.

Denote the set of top-$K$ solutions at the end of time step $t - 1$ as $S_{t-1} = \{s_{1,t-1}, \ldots, s_{K,t-1}\}$, where $s_{k,t-1}$ is one of the top-$K$ scoring word sequences at time step $t - 1$. At time step $t$, let $\tilde{S}_t = \{s| (s_{t-1} \in S_{t-1}) \land (w_t \in V) \}$ denote the set of all possible single word extensions of each solution, where $w_{t}$ represents a word in the vocabulary $V$. Then, the top-$K$ solutions are updated as

$$S_t = \arg\max_{s_1, \ldots, s_K} \sum_{k \in [1,K]} \Theta_t(s_{k,t}; x)$$

$$s.t. \ s_t \neq s_j (\forall i, j \in [1,K] \text{ and } i \neq j)$$

where $x$ represents the encoded image features, and $\Theta_t(s; x) = \sum_{i \in [1,T]} \log p(r_{i} | s_{i-1}, x)$ is the log probability of the generated sequence at time step $T$ calculated by the CNN-LSTM network described in Section III.B.

To evaluate the captions generated by the beam search algorithm with different beam sizes, we constructed a DIK for each category of drug-related paraphernalia in the DrunaliaCap dataset. We first programmed counting all the words appearing in the factual annotations and manually screened out the nouns representing drug-related paraphernalia, such as “cigar” and “bong”, to form a nonduplicate vocabulary denoted as $V_d$. Then, we used the same method (this time for the functional annotations) to construct a nonduplicate vocabulary of words containing important knowledge about drug paraphernalia, such as “marijuana”, “smoking”, and “blunt”, denoted as $V_u$.

The finally constructed DIK is a dictionary denoted as $D$, each key of which corresponds to a noun in $V_u$. We iterate the factual sentences $y_i^f$ and functional sentences $y_i^u$ of each image $x_i$ in the dataset. For each noun $w$ in $V_u$ that also exists in $y_i^f$, words containing important knowledge (those in $V_u$ and also appearing in $y_i^f$) are added to $D[w]$, which represents a nonduplicate set.

We propose an algorithm to evaluate the performance of the captioning model with a beam size of $k$ for generating important knowledge of drug paraphernalia. When generating functional captions for an image $x_i$, the algorithm first looks up $D[w]$ to obtain a nonduplicate set $l$ containing words of important knowledge, where $w$ represents each of the drug paraphernalia nouns in the factual sentences $y_i^f$. The number of nonduplicate words in $l$ that appear in the generated $k$ functional sentences is counted and denoted as $n$. The average number $\bar{n}$ of $n$ for all $|D|$ images is used to evaluate the performance of the captioning model in generating important knowledge of drug paraphernalia. A larger $\bar{n}$ represents better performance. The related pseudo code is shown in Algorithm 1.

### IV. EXPERIMENTS

#### A. SETUP

Our experiments include two parts. In the first part, we validate the necessity of the newly constructed dataset and evaluate the performance of different models for auto-generating factual and functional captions. For comparison, CNN-LSTM models were separately trained on MS-COCO [1], Flickr30k [2], Flickr8k [3], and the newly constructed DrunaliaCap dataset. In the second part, we validate the algorithm described in Section III.C by analyzing the generation of functional captions under different beam sizes.
Algorithm 1: Our Proposed Algorithm to Construct the DIK and to Evaluate Beam Search With a Size of k. See Section III.C for Details

Input: \( \mathcal{D} = \{x_i, y_i^f, y_i^p\} \): image captioning dataset, \( \mathcal{V}_a \): vocabulary of drug paraphernalia nouns, \( \mathcal{V}_d \): vocabulary of knowledge words, \( k \): beam size

Output: \( \mathcal{D} \): the constructed DIK, \( \bar{n} \): average words in \( \mathcal{D} \) in the generated captions

1. **Function** BUILD-DICTIONARY(\( \mathcal{D}, \mathcal{V}_a, \mathcal{V}_d \)):
   1. \( n \leftarrow 0 \)
   2. **foreach** \( w \in \mathcal{V}_a \) **do**
      1. \( \mathcal{D}[w] \leftarrow \emptyset \)
   3. **end**
   4. **foreach** \( w \in y_i^d \) **do**
      1. \( \mathcal{D}[w] \leftarrow \mathcal{D}[w] \cup (y_i^d \cap \mathcal{V}_d) \)
   5. **end**
   6. return \( \mathcal{D} \);

2. **Function** EVALUATE-BEAM-SIZE(\( k, \mathcal{D}, \mathcal{D} \)):
   1. \( n \leftarrow 0 \)
   2. **foreach** \( (x_i, y_i^f, y_i^p) \in \mathcal{D} \) **do**
      1. \( l \leftarrow \emptyset \)
      2. **foreach** \( w \in (y_i^f \cap \mathcal{V}_a) \) **do**
         1. \( l \leftarrow l \cup \mathcal{D}[w] \)
      3. **end**
      4. \( S \leftarrow \emptyset \)
      5. **while** not all \( s \in S \) meet \(<\text{End}>\) **do**
         1. update \( S \) with Equation 5
      6. **end**
   3. \( h \leftarrow \emptyset \)
   4. **foreach** word \( w \in (S \cap l) \) **do**
      1. \( h \leftarrow h \cup w \)
   5. **end**
   6. \( n \leftarrow n + |h| \)
   7. **end**
   8. return \( n \div |\mathcal{D}| \);

9. **End Function**

10. \( \mathcal{D} \leftarrow \text{BUILD-DICTIONARY}(\mathcal{D}, \mathcal{V}_a, \mathcal{V}_d) \)
11. \( \bar{n} \leftarrow \text{EVALUATE-BEAM-SIZE}(k, \mathcal{D}, \mathcal{D}) \)

The experimental results of the two parts are analyzed in detail in Sections IV.C and IV.D, respectively.

1) DATA SPLIT
We randomly selected 15% of the data from the DrunaliaCap dataset as a held-out test set and another 15% of the data as a validation set for hyperparameter optimization. The remaining 70% of the data were used for training. For MS-COCO, Flickr30k, and Flickr8k, we followed the publicly available splits from a previous study by Karpathy et al. [19].

2) TRAINING STRATEGY
During training, we set a dropout rate of 0.5 for the multilayer perceptron, and early stopping using the Bilingual Evaluation Understudy (BLEU) score was applied according to Xu et al.’s previous work [5]. Further, we used the Adam optimizer with a mini-batch size of 32. The initial learning rate was set to 0.0004, which decayed every 5 epochs with a decay factor of 0.8.

3) IMPLEMENTATION
Our trainable CNN-LSTM model was implemented on the basis of PyTorch [20]. The ImageNet-pretrained ResNet-101 model was obtained from PyTorch’s torchvision module. The weights of the proposed image feature extractor were taken from the pretrained model and fixed during training. The other layers in the proposed network were trained from scratch. The models were trained using a desktop computer with the 64-bit Ubuntu18.04 operating system and an NVIDIA GTX-2080ti GPU.

B. EVALUATION CRITERIA
To evaluate and compare the captions generated by different methods, we used three metrics commonly used in image captioning, namely, the BLEU [23], the Metric for Evaluation of Translation with Explicit Ordering (METEOR) [24], and the Consensus-based Image Description Evaluation (CIDEr) [25]. The characteristics of each metric are described below. In addition, in the second part of the experiment, Algorithm 1 presented in Section III.C was used to evaluate the performance of the models for generating important knowledge of drug-related paraphernalia.

BLEU was originally designed to be used in the field of machine translation. It is now also widely used to evaluate the performance of image captioning models. This algorithm evaluates the quality of machine-generated text by the proportion of the same n-grams that occur in the ground truth sentences. The notation BLEU-n indicates the algorithm using n-grams (n can be an integer from 1-4).

METEOR is based on the harmonic mean of the unigram precision and recall, with the recall weighted higher than the precision. Compared with BLEU, this metric produces a good correlation with human judgment at the sentence level rather than the corpus level.

CIDEr was designed to evaluate the quality of an image caption. It encodes sentences in the form of term frequency–inverse document frequency vectors. The cosine similarity between a machine-generated caption and ground truth caption is calculated as a score.

For all three metrics, a larger score means better performance.

C. GENERATING FACTUAL AND FUNCTIONAL DESCRIPTIONS
Table 2 shows the performance of the models (without beam search) trained on different datasets in the proposed factual and functional captioning tasks. In both the factual and functional tasks, the captions generated by the methods other than DrunaliaCap obtained BLUE-3/4 scores of 0 and extremely low BLEU-1/2, METEOR and CIDEr scores. These results
The BLEU-4/CIDEr/METEOR scores obtained by the models trained using the DrunaliaCap dataset are 0.226/2.973/0.559 and 0.209/1.818/0.473 in factual and functional tasks, respectively. Compared with the scores obtained by the other methods, these scores indicate that the models trained using the newly constructed dataset generate factual and functional captions with a certain quality.

We further investigated the output of all models. Some typical examples are shown in Figure 6. As shown in the figure, the models trained on the DrunaliaCap dataset generate expected factual and functional captions. Meanwhile, drug-related items are incorrectly recognized by Caption Bot and the models trained on MS-COCO, Flickr30k, and Flickr8k. Although these state-of-the-art large-scale datasets and the application-like Caption Bot contain a range of scenes and objects, they fail to cover the content of drug-related items.

Table 2 shows the comparison of image captioning results from different models (without beam search).

| Model            | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | CIDEr | METEOR |
|------------------|--------|--------|--------|--------|-------|--------|
| MS-COCO          | 0.169  | 0.073  | 0.000  | 0.000  | 0.237 | 0.203  |
| Flickr30k        | 0.103  | 0.024  | 0.000  | 0.000  | 0.010 | 0.117  |
| Flickr8k         | 0.125  | 0.044  | 0.000  | 0.000  | 0.061 | 0.145  |
| Caption Bot      | 0.079  | 0.024  | 0.000  | 0.000  | 0.059 | 0.095  |
| DrunaliaCap      | 0.569  | 0.451  | 0.359  | 0.226  | 2.973 | 0.559  |

## Conclusion

The results indicate that these datasets contain very few or no drug paraphernalia images and annotations.
In summary, the scores of each metric and the intuitive output of each method show that images and annotations of drug paraphernalia are not covered by the methods other than DrunaliaCap, which demonstrates the necessity of constructing the new dataset. In addition, the models trained on the DrunaliaCap dataset generate descriptions of a certain quality for drug-related paraphernalia.

**FIGURE 7.** Evaluation of functional captions with different beam sizes indicated by the x-axis. The red line DIK-\( n \) represents the average number of nonduplicate words from the DIK in the generated captions per image and corresponds to the left y-coordinate axis. The purple, green and blue lines represent the BLEU-4, METEOR and CIDEr scores (one sentence per image with the highest log probability), respectively, and correspond to the y-coordinate axis on the right.

**D. GENERATING IMPORTANT KNOWLEDGE OF DRUG-RELATED PARAPHERNALIA**

Figure 7 shows the trends of the results under different evaluation criteria with increasing beam size. As shown in the figure, an increase in the beam size does not help improve the scores of BLEU-4, METEOR or CIDEr. However, as the beam size increases from 1 to 4, the average number of words with important knowledge (denoted as DIK-\( n \), corresponding to \( n \) in Algorithm 1) increases significantly from 1.85 to 3.60. When the beam size is greater than 4, the growth trend of DIK-\( n \) becomes much slower.

We further investigated the distribution of the words from the DIK in the generated sentences. Figure 8 shows some representative examples. The words from the DIK (in blue) corresponding to the factual nouns (in red) are distributed in different captions generated by the beam search algorithm. The number of nonduplicate words with important knowledge can be calculated by looking up the key words (words in red) in the DIK according to Algorithm 1.

In summary, our proposed method described in Section III.C showed its effectiveness in evaluating and optimizing the generated captions and preventing them from missing important knowledge of drug paraphernalia, which is a different but important perspective not measured by other metrics.

**V. DISCUSSION**

**A. BIASES IN THE EXPERIMENTS**

The models trained on datasets other than DrunaliaCap achieved extremely low scores. This is because these datasets contain few or no images of drug paraphernalia as well as factual and functional descriptions, and these limitations are why we constructed the new dataset. However, we must clarify that this biased distribution of images and annotations only verifies the necessity of constructing the new dataset for DrunaliaCap. It does not prove that the other datasets have any defects as image captioning benchmarks for testing algorithmic performance.

**B. GENERATION OF VALUABLE NON-FACTUAL INFORMATION**

Existing non-factual captioning studies tend to use sequence generation metrics such as BLEU, METEOR, and CIDEr to evaluate and optimize image captioning models. Scores obtained under these metrics highly depend on the similarity between the generated captions and the ground truth annotations. Unlike factual captioning, in the task of non-factual captioning, words and phrases containing valuable information (sentiment, personality, writing style, important knowledge, etc.) for a certain object in an image may be distributed in the ground truth annotations of different images. The extent to which the generated captions contain all such valuable information may be ignored by existing evaluation metrics. In this paper, we propose a method to evaluate and optimize functional captions to prevent them from missing important knowledge. However, our method partly relies on human engineering to maintain the vocabularies. Unsupervised or semisupervised solutions for digging and generating valuable non-factual information are worthy of future effort.

**C. FUTURE ALGORITHM EXPLORATION**

The backbone of the encoder-decoder captioning models in the proposed DrunaliaCap system is based on [5]. For the task of autogenerating factual and functional captions for drug paraphernalia, there is much room for future exploration from an algorithmic perspective. Some recent image captioning studies [21], [26], [31] have constructed variant LSTM language models to learn factual and non-factual knowledge in corpora. Some studies [21], [32]–[34] have allowed for learning non-factual knowledge in unpaired corpora via weakly supervised or unsupervised methods. These methods are expected (but not limited) to be used to train better models for autogenerating factual and functional captions of drug paraphernalia.

**D. FUTURE WORK AND POTENTIAL APPLICATIONS**

The images in the newly constructed DrunaliaCap dataset are close-ups of drug-related items and are annotated with factual and functional captions. Larger-scale datasets containing photos of drug paraphernalia in common life settings and other stylized text annotations are worth exploring in future research and engineering applications. Images of drug-related paraphernalia and corresponding text descriptions are common in drug prevention teaching classes, police enforcement reports, social networks and e-commerce websites. Autogenerated descriptions are expected to be used as references...
• Factual annotation:
  - a cigar
• Generated functional captions:
  - a cigar is often used for smoking marijuana
  - a cigar is often used to smoke marijuana
  - a cigar can be used to smoke marijuana
  - marijuana rolled into a cigar is called a blunt
• Generated knowledge in DIK:
  \{smoking, marijuana, smoke, blunt\} \in \mathcal{D}[\text{cigar}]

• Factual annotation:
  - a round base glass bong
• Generated functional captions:
  - the photo shows a round base glass bong for smoking crystal methamphetamine
  - the photo shows a round base glass bong for smoking powerful drugs like PCP or crystal methamphetamines
  - a round base glass bong is often used to filter drugs through water for smoking
  - a round base glass bong is often used to filter and smoke drugs through water
• Generated knowledge in DIK:
  \{smoking, crystal methamphetamines, PCP, filter, smoke, water\} \in \mathcal{D}[\text{bong}]

• Factual annotation:
  - a piece of tinfoil and a red lighter
• Generated functional captions:
  - the photo shows a piece of tinfoil and a red lighter to heat and inhale heroin
  - the photo shows a piece of tinfoil and a red lighter to heat and smoke crack cocaine
  - drug addicts often use a piece of tinfoil and a red lighter to heat and inhale heroin
  - drug addicts often place drugs onto a piece of tinfoil and heat it using a lighter
• Generated knowledge in DIK:
  \{heat, inhale, heroin, smoke, crack cocaine, place\} \in (\mathcal{D}[\text{tinfoil}] \cup \mathcal{D}[\text{lighter}])

FIGURE 8. Examples of generating important knowledge (words from the DIK) for drug paraphernalia with a beam size of 4. The red words represent nouns related to drug paraphernalia in the factual annotations, and the blue words contain important knowledge related to those nouns (see Algorithm 1 for details).

VI. CONCLUSION
We design a system with deep learning called DrunaliaCap for a newly proposed task of autogenerating factual and functional captions of images of drug-related paraphernalia. A new image captioning dataset is constructed to train deep learning-based models. We further propose a method to evaluate and optimize the generation of captions to prevent them from missing important knowledge. Experiments are conducted to validate the dataset and methods in our proposed DrunaliaCap system. We analyze the experimental results and describe the significance, limitations, future exploration, and potential applications of our work. To the best of our knowledge, we are the first to propose a deep learning-based system to autogenerate image captions for drug-related paraphernalia. Hopefully, this study will encourage future work in deep learning-based research on drug-related images and text.

REFERENCES
[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2014, pp. 740–755.

[2] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” Trans. Assoc. Comput. Linguistics, vol. 2, pp. 67–78, Dec. 2014.

[3] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” J. Artif. Intell. Rex., vol. 47, pp. 853–899, Aug. 2013.

[4] J. Strang, J. Bearn, M. Farrell, E. Finch, M. Gossop, P. Griffiths, J. Marsden, and K. Wolff, “Route of drug use and its implications for drug effect, risk of dependence and health consequences,” Drug Alcohol Rex., vol. 17, no. 2, pp. 197–211, Jun. 1998.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 2048–2057.

[6] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, May 2017.

[7] K. Tran, X. He, L. Zhang, and J. Sun, “Rich image captioning in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2016, pp. 49–56.

[8] C. C. Park, B. Kim, and G. Kim, “Attend to you: Personalized image captioning with context sequence memory networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 895–903.

[9] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, “The IAPR TC-12 benchmark: A new evaluation resource for visual information systems,” in Proc. Int. Workshop Image, vol. 2, 2006, pp. 1–55.

[10] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, Y. Wang, and Y. Wang, “AI challenger: A large-scale dataset for going deeper in image understanding,” 2017, arXiv:1711.06475. [Online]. Available: http://arxiv.org/abs/1711.06475.
