

Aesthetic Image Captioning from Weakly-Labelled Photographs

Koustav Ghosal

Aakanksha Rana

Aljosa Smolic

V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland.

Abstract

Aesthetic image captioning (AIC) refers to the multi-modal task of generating critical textual feedbacks for photographs. While in natural image captioning (NIC), deep models are trained in an end-to-end manner using large curated datasets such as MS-COCO, no such large-scale, clean dataset exists for AIC. Towards this goal, we propose an automatic cleaning strategy to create a benchmarking AIC dataset, by exploiting the images and noisy comments easily available from photography websites. We propose a probabilistic caption-filtering method for cleaning the noisy web-data, and compile a large-scale, clean dataset ‘AVA-Captions’, ($\sim 230,000$ images with ~ 5 captions per image). Additionally, by exploiting the latent associations between aesthetic attributes, we propose a strategy for training a convolutional neural network (CNN) based visual feature extractor, typically the first component of an AIC framework. The strategy is weakly supervised and can be effectively used to learn rich aesthetic representations, without requiring expensive ground-truth annotations. We finally showcase a thorough analysis of the proposed contributions using automatic metrics and subjective evaluations.

1. Introduction

Availability of large curated datasets such as MS-COCO [41] (100K images), Flickr30K [64] (30K images) or Conceptual Captions [74] (3M images) made it possible to train deep learning models for complex, multi-modal tasks such as natural image captioning (NIC) [81] where the goal is to factually describe the image content. Similarly, several other captioning variants such as visual question answering [5], visual storytelling [38], stylized captioning [56] have also been explored. Recently, the PCCD dataset (~ 4200 images) [11] opened up a new area of research of describing images aesthetically. Aesthetic image captioning (AIC) has potential applications in the creative industries such as developing smarter cameras or web-based applications, ranking, retrieval of images and videos *etc.* However in [11], only six well-known photographic/aesthetic attributes such as composition, color, lighting, *etc.* have

been used to generate aesthetic captions with a small curated dataset. Hence, curating a large-scale dataset to facilitate a more comprehensive and generalized understanding of aesthetic attributes remains an open problem.

Large-scale datasets have always been pivotal for research advancements in various fields [15, 41, 64, 67]. However, manually curating such a dataset for AIC is not only time consuming, but also difficult due to its subjective nature. Moreover, a lack of unanimously agreed ‘standard’ aesthetic attributes makes this problem even more challenging as compared to its NIC counterpart, where deep models are trained with known attributes/labels [41]. In this paper, we make two contributions. Firstly, we propose an automatic cleaning strategy to generate a large scale dataset by utilizing the noisy comments or aesthetic feedback provided by users for images on the web. Secondly, for a CNN-based visual feature extractor as is typical in NIC pipelines, we propose a weakly-supervised training strategy. By automatically discovering certain ‘meaningful and complex aesthetic concepts’, beyond the classical concepts such as composition, color, lighting, *etc.*, our strategy can be adopted in scenarios where finding clean ground-truth annotations is difficult (as in the case of many commercial applications). We elaborate these contributions in the rest of this section.

To generate a clean aesthetic captioning dataset, we collected the raw user comments from the Aesthetic Visual Analysis (AVA) dataset [58]. AVA is a widely used dataset for aesthetic image analysis tasks such as aesthetic rating prediction [44, 48], photographic style classification [25, 34]. However, AVA was not created for AIC. In this paper, we refer to the original AVA with raw user comments as AVA raw-caption. It contains $\sim 250,000$ photographs from dpchallenge.com and the corresponding user comments or feedback for each photograph (3 billion in total). Typically, in Dpchallenge, users ranging from casual hobbyists to expert photographers provide feedback to the images submitted and describe the factors that make a photograph aesthetically pleasing or dull. Even though these captions contain crucial aesthetic-based information from images, they cannot be directly used for the task of AIC. Unlike the well instructed and curated datasets [41], AVA raw-captions are unconstrained user-comments in the wild with typos,





Training Strategy				
(a) Noisy Data & Supervised CNN (NS)	i like the angle and the composition	i like the colors and the composition	i like the composition and the lighting	i like the composition and the bw
(b) Clean Data & Supervised CNN (CS)	i like the idea , but i think it would have been better if the door was in focus .	i like the colors and the water . the water is a little distracting .	i like the way the light hits the face and the background .	i like this shot . i like the way the lines lead the eye into the photo .
(c) Clean Data & Weakly Supervised CNN (CWS)	i like the composition , but i think it would have been better if you could have gotten a little more of the building	i like the composition and the colors . the water is a little too bright .	this is a great shot . i love the way the light is coming from the left .	i like the composition and the bw conversion .

Figure 1. Aesthetic image captions. We show candidates generated by three different frameworks discussed in this paper: **(a)** For NS, we use an ImageNet trained CNN and LSTM trained on noisy comments **(b)** For CS, we use an ImageNet trained CNN and LSTM trained on compiled AVA-Captions dataset **(c)** For CWS, we use a weakly-supervised CNN and LSTM trained on AVA-Captions

grammatically inconsistent statements, and also containing a large number of comments occurring frequently without useful information. Previous work in AIC [11] acknowledges the difficulty of dealing with the highly noisy captions available in AVA.

In this work, we propose to clean the raw captions from AVA by proposing a probabilistic n-gram based filtering strategy. Based on word-composition and frequency of occurrence of n-grams, we propose to assign an informativeness score to each comment, where comments with a little or vague information are discarded. Our resulting clean dataset, **AVA-Captions** contains $\sim 230,000$ images and $\sim 1.5M$ captions with an average of ~ 5 comments per image and can be used to train the Long and Short Term Memory (LSTM) network in the image captioning pipeline in the traditional way. Our subjective study verifies that the proposed automatic strategy is consistent with human judgement regarding the informativeness of a caption. Our quantitative experiments and subjective studies also suggest that models trained on AVA-Captions are more diverse and accurate than those trained on the original noisy AVA-Comments. It is important to note that our strategy to choose the large-scale AVA raw-caption is motivated from the widely used image analysis benchmarking dataset, MSCOCO, which is now used as an unified benchmark for diverse tasks such as object detection, segmentation, captioning, *etc.* We hope that our cleaned dataset will serve as a new benchmarking dataset for various creative studies and aesthetics-based applications such as aesthetics based image enhancement, smarter photography cameras, *etc.*

Our second contribution in this work is a weakly supervised approach for training a CNN, as an alternative to the standard practice. The standard approach for most image captioning pipelines is to train a CNN on large annotated datasets *e.g.* ImageNet [15], where rich and discriminative

visual features are extracted corresponding to the physical properties of objects such as cars, dogs *etc.* These features are provided as input to an LSTM for generating captions. Although trained for classification, these ImageNet-based features have been shown to translate well to other tasks such as segmentation [42], style-transfer [22], NIC. In fact, due to the unavailability of large-scale, task-specific CNN annotations, these ImageNet features have been used for other variants of NIC such as aesthetic captioning [11], stylized captioning [56], product descriptions [82], *etc.*

However, for many commercial/practical applications, availability of such datasets or models is unclear due to copyright restrictions [24, 37, 83]. On the other hand, collecting task-specific manual annotations for a CNN is expensive and time intensive. Thus the question remains open if we can achieve better or at least comparable performance by utilizing easily available weak annotations from the web (as found in AVA) and use them for training the visual feature extractor in AIC. To this end, motivated from weakly supervised learning methods [18, 69], we propose a strategy which exploits the large pool of unstructured raw-comments from AVA and discovers latent structures corresponding to meaningful *photographic concepts* using Latent Dirichlet Allocation (LDA) [10]. We experimentally observe that the weakly-supervised approach is effective and its performance is comparable to the standard ImageNet trained supervised features.

In essence, our contributions are as follows:

1. We propose a caption filtering strategy and compile AVA-Captions, a large-scale and clean dataset for aesthetic image captioning (Sec 3).
2. We propose a weakly-supervised approach for training the CNN of a standard CNN-LSTM framework (Sec 4)

3. We showcase the analysis of the AIC pipeline based on the standard automated metrics (such as BLEU, CIDEr, SPICE *etc.* [2, 62, 78]), diversity of captions and subjective evaluations which are publicly available for further explorations (Section 6).

2. Related Work

Due to the multi-modal nature of the task, the problem spans into many different areas of image and text analysis and thus related literature abound. However, based on the primary theme we roughly divide this section into four areas as follows:

Natural Image Captioning: While early captioning methods [21, 27, 29, 61, 75] followed a dictionary look-up approach, recent parametric methods [1, 4, 9, 19, 20, 23, 30, 31, 35, 50, 52–54, 77] are generative in the sense that they learn a mapping from visual to textual modality. Typically in these frameworks, a CNN is followed by a RNN or LSTM [4, 19, 20, 31, 35, 52–54, 81], although fully convolutional systems have been proposed by Aneja *et al.* [3] recently.

Image Aesthetics: Research in understanding the perceptual and aesthetic concepts in images can be divided into the model-based [6, 14, 16, 32, 36, 47, 60, 72] and the data-driven [34, 44, 45, 48, 49, 51] approaches. While model-based approaches rely on manually hard-coding the aspects such as the Rule of Thirds, depth of field, colour harmony, etc., the data driven approaches usually train CNNs on large-scale datasets and either predict an overall aesthetic rating [44, 45, 49] or a distribution over photographic attributes [25, 34, 44, 45].

Learning from Weakly-Annotated / Noisy Data: Data dependency of very deep neural nets and the high cost of human supervision has led to a natural interest towards exploring the easily available web-based big data. Typically in these approaches, web-crawlers collect easily available noisy multi-modal data [8, 12, 79] or e-books [17] which is jointly processed for labelling and knowledge extraction. The features are used for diverse applications such as classification and retrieval [68, 76] or product description generation [82].

Aesthetic Image Captioning: To the best of our knowledge, the problem of aesthetic image captioning has been first and only addressed by Chang *et al.* in [11]. The authors propose a framework which extracts features covering seven different aspects such as general impression, composition and perspective, color and lighting, etc. and generate meaningful captions by fusing them together. They compile the photo critique captioning dataset (PCCD) with $\sim 4K$ images and $\sim 30K$ captions. While their method is purely supervised and the network is trained using strong labels, we adopt a weakly-supervised approach to train our network with indirect labels. Additionally, AVA-Captions is a


Image	Comments	Scores
	Photo Quality : Awesome	9.62
	I love the colors here	1.85
	I like the trees in the background and the softness of the water.	28.41
	The post processing looks great with the water, but the top half of the photo doesn't work as well.	47.44

Figure 2. Informativeness of captions. We suggest the readers to check the supplementary material for more comments and the corresponding scores.

significantly bigger (~ 60 times) dataset with $\sim 240K$ and $\sim 1.3M$ images and captions, respectively. The scale of AVA allows training deeper and more complex architectures which can be generalized to PCCD as well. We demonstrate this later in Table 1b.

3. Caption Filtering Strategy

In AVA raw-caption, we observe two main types of undesirable captions. First, there are captions which suffer from generic noise frequently observed in most text corpora, especially those compiled from social media. They include typing errors, non-English comments, colloquial acronyms, exclamatory words (such as “wooooo”), extra punctuation (such as “!!!!”), etc. Such noise can be handled using standard natural language processing techniques [43].

Second, we refer to the *safe* comments, which carry a little or no useful information about the photograph. For example, in Figure 2, comments such as “Photo Quality : Awesome” or “I love the colors here” provide a valid but less informative description of the photograph. It is important to filter these comments, otherwise the network ends up learning these less-informative, *safe* captions by ignoring the more informative and discriminative ones such as “The post processing looks great with the water, but the top half of the photo doesn’t work as well.” [11].

To this end, we propose a probabilistic strategy for caption filtering based on the informativeness of a caption. Informativeness is measured by the presence of certain n-grams. The approach draws motivation from two techniques frequently used in vision-language problems — word composition and term-frequency - inverse document frequency (TF-IDF).

Word Composition: Bigrams of the “descriptor-object” form often convey more information than the unigrams of the objects alone. For example, “post processing” or “top half” convey more information than “processing” or “half”. On the other hand, the descriptors alone may not always be sufficient to describe a complete concept and its mean-

ing is often closely tied to the object [59]. For example, “sharp” could be used in two entirely different contexts such as “sharp contrast” and “sharp eyes”. This pattern is also observed in the 200 bigrams (or ugly and beautiful attributes) discovered from AVA by Marchesotti *et al.* [58] such as “nice colors”, “beautiful scene”, “too small”, “distracting background”, *etc.* Similar n-gram modelling is found in natural language processing as adjective-noun [57, 73, 75] or verb-object [71, 84] compositions.

TF-IDF: The other motivation is based on the intuition that the key information in a comment is stored in certain n-grams which occur less frequently in the comment corpus such as “softness”, “post processing”, “top half” *etc.* A sentence composed of frequently occurring n-grams such as “colors” or “awesome” is less likely to contain useful information. The intuition follows from the motivation of commonly used TF-IDF metric in document classification, which states that more frequent words of a vocabulary are less discriminative for document classification [66]. Such hypothesis also forms a basis in the CIDEr evaluation metric [78] widely used for tasks such as image captioning, machine translation, *etc.*

Proposed “Informativeness” Score: Based on these two criteria, we start by constructing two vocabularies as follows: for unigrams we choose only the nouns and for bigrams we select “descriptor-object” patterns *i.e.* where the first term is a noun, adjective or adverb and the second term is a noun or an adjective. Each n-gram ω is assigned a corpus probability P as:

$$P(\omega) = \frac{C_\omega}{\sum_{i=1}^D C_i} \quad (1)$$

where the denominator sums the frequency of each n-gram ω such that $\sum_{i=1}^D P(\omega_i) = 1$, where D is the vocabulary size, and C_ω is the corpus frequency of n-gram ω . Corpus frequency of an n-gram refers to the number of times it occurs in the comments from all the images combined. This formulation assigns high probabilities for more frequent words in the comment corpus.

Then, we represent a comment as the union of its unigrams and bigrams *i.e.*, $S = (S_u \cup S_b)$, where $S_u = (u_1 u_2 \dots u_N)$ and $S_b = (b_1 b_2 \dots b_M)$ are the sequences of unigrams and bigrams, respectively. A comment is assigned an informativeness score ρ as follows:

$$\rho_s = -\frac{1}{2} \left[\log \prod_i^N P(u_i) + \log \prod_j^M P(b_j) \right] \quad (2)$$

where $P(u)$ and $P(b)$ are the probabilities of a unigram or bigram given by Equation 1. Equation 2 is the average of the negative log probabilities of S_u and S_b .

Essentially, the score of a comment is modelled as the joint probability of n-grams in it, following the simplest

Markov assumption *i.e.* all n-grams are independent [33]. If the n-grams in a sentence have higher corpus probabilities then the corresponding score ρ is low due to the negative logarithm, and vice-versa.

Note that the score is the negative logarithm of the product of probabilities and longer captions tend to receive higher scores. However, our approach does not *always* favour long comments, but does so only if they consist of “meaningful” n-grams conforming to the “descriptor-object” composition. In other words, randomly long sentences without useful information are discarded. On the other hand, long and informative comments are kept. This is also desirable as longer comments in AVA tend to be richer in information as expert users are specifically asked to provide detailed assessment which is referred to as *critique club effect* in [55].

We label a comment as informative or less-informative by thresholding (experimentally kept 20) the score ρ . Some sample scores are provided in Figure 2. The proposed strategy discards about 1.5M (55%) comments from the entire corpus. Subsequently, we remove the images which are left with no informative comments. Finally, we are left with 240,060 images and 1,318,359 comments, with an average of 5.58 comments per image. We call this cleaner subset as **AVA-Captions**. The proposed approach is evaluated by human subjects and the results are discussed in Figure 6 and Section 6.3.4.

4. Weakly Supervised CNN

Although the comments in AVA-Captions are cleaner than the raw comments, they cannot be directly used for training the CNN *i.e.* the visual feature extractor. As discussed in Sec 1, the standard approach followed in NIC and its variants is to use an ImageNet trained model for the task. In this section, we propose an alternative weakly supervised strategy for training the CNN from scratch by exploiting the *latent* aesthetic information within the AVA-Captions. Our approach is motivated from two different areas: visual attribute learning and text document clustering.

4.1. Visual and Aesthetic Attributes

Visual Attribute Learning is an active and well-studied problem in computer vision. Instead of high-level object/scene annotations, models are trained for low-level attributes such as “smiling face”, “open mouth”, “full sleeve” *etc.* and the features are used for tasks such as image-ranking [63], pose-estimation [85], fashion retrieval [80], zero-shot learning [28], *etc.* Similarly, our goal is to identify aesthetic attributes and train a CNN. A straightforward approach is to use the n-grams from comments (Sec 3) and use them as aesthetic attributes. However, there are two problems with this approach: Firstly, the set of n-grams is huge ($\sim 25K$) and thus training the CNN directly using



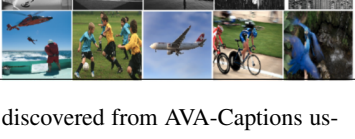
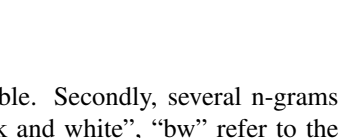
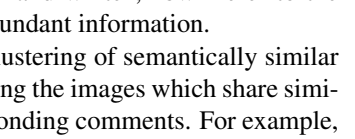
Topics	Images
“Cute-Expression”, “Face”, “Ear”	
“Landscape”, “Sky”, “Cloud”	
“Action Shot”, “Sport”, “Great Action”	
“Black and white”, “Tone”, “Contrast”	
“Motion Blur”, “Move- ment”, “Shutter Speed”	

Figure 3. Some topics / labels discovered from AVA-Captions using LDA.

them as labels is not scalable. Secondly, several n-grams such as “grayscale”, “black and white”, “bw” refer to the same concept and carry redundant information.

Therefore, we apply a clustering of semantically similar n-grams and thereby grouping the images which share similar n-grams in their corresponding comments. For example, portraits are more likely to contain attributes such as “cute expression”, “face” *etc.* and landscape shots are more likely to share attributes such as “tilted horizon”, “sky”, “over-exposed clouds” *etc.* Essentially, the intuition behind our approach is to discover clusters of photographic attributes or topics from the comment corpus and use them as labels for training the CNN. In text document analysis, it is a common practice to achieve such grouping of topics from a text corpus using a technique called Latent Dirichlet Allocation [10].

4.2. Latent Dirichlet Allocation (LDA)

LDA is an unsupervised generative probabilistic model, widely used for topic modelling in text corpora. It represents text documents as a probabilistic mixture of topics, and each topic as a probabilistic mixture of words. The words which co-occur frequently in the corpus are grouped together by LDA to form a topic. For example, by running LDA on a large corpus of news articles, it is possible to discover topics such as “sports”, “government policies”, “terrorism” *etc* [39].

Formally stated, given a set of documents $D_i = \{D_1, D_2 \dots D_N\}$, and a vocabulary of words $\omega_i = \{\omega_1, \omega_2 \dots \omega_M\}$, the task is to infer K latent topics $T_i = \{T_1, T_2, \dots T_K\}$, where each topic can be represented as a collection of words (term-topic matrix) and each document can be represented as a collection of topics (document-topic matrix). The term-topic matrix represents the probabilities of each word associated with a topic and the document-topic matrix refers to the distribution of a document over the K

latent topics. The inference is achieved using a variational Bayes approximation [10] or Gibb’s sampling [65]. A more detailed explanation can be found in [10].

4.3. Relabelling AVA Images

We regard all the comments corresponding to a given image as a document. The vocabulary is constructed by combining the unigrams and bigrams extracted from the AVA-Captions as described in Section 3. In our case: $N = 230,698$ and $M = 25,000$, and K is experimentally fixed to 200. By running LDA with these parameters on AVA-Captions, we discover 200 latent topics, composed of n-grams which co-occur frequently. The method is based on the assumption that the visual aesthetic attributes in the image are correlated with the corresponding comments and images possessing similar aesthetic properties are described using similar words.

Even after the caption cleaning procedure, we observe that n-grams such as “nice composition” or “great shot” still occur more frequently than others. But, they occur mostly as independent clauses in bigger comments such as “*I like the way how the lines lead the eyes to the subject. Nice shot!*”. In order to avoid inferring topics consisting of these less discriminative words, we consider only those n-grams in the vocabulary which occur in less than 10% comments.

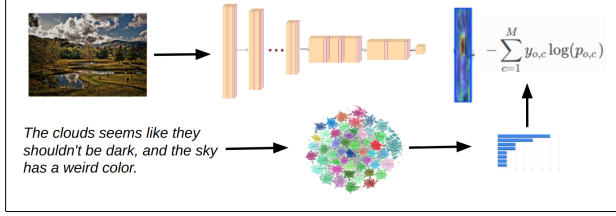
In Figure 3, we select 5 topics thus inferred and some of the corresponding images whose captions belong to these topics. It can be observed that the images and the words corresponding to each topic are fairly consistent and suitable to be used as labels for training the CNN.

4.4. Training the CNN

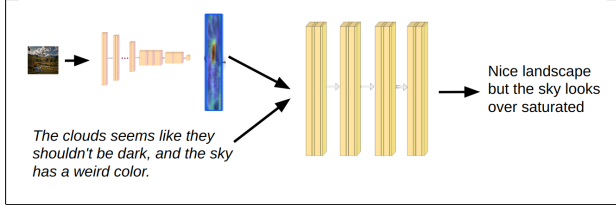
Given an image and its corresponding captions, we estimate the topic distribution D_T of the comments. The CNN is trained using D_T as the ground-truth label. We adopt the ResNet101 [26] architecture and replace the last fully connected layer with K outputs, and train the framework using cross-entropy loss [70] as shown in Figure 4a.

5. The Final Framework

We adopt the NeuralTalk2 [46] framework as our basis. Note, that our approach is generic and can be used with any CNN-LSTM framework for image captioning. In [46], visual features are extracted using an ImageNet trained ResNet101 [26] which are passed as input to an LSTM for training the language model using the ground-truth captions. For our framework, we use two alternatives for visual features (a) ImageNet trained (b) weakly supervised (Sec 4). The LSTM architecture is kept unchanged except hyper-parameters such as vocabulary size, maximum allowed length of a caption *etc.* The language model is trained using the clean and informative comments from the AVA-Captions dataset (See Figure 4b).



(a) **Weakly-supervised training of the CNN:** Images and comments are provided as input. The image is fed to the CNN and the comment is fed to the inferred topic model. The topic model predicts a distribution over the topics which is used as a label for computing the loss for the CNN.



(b) **Training the LSTM:** Visual features extracted using the CNN and the comment is fed as an input to the LSTM which predicts a candidate caption.

Figure 4. **Proposed pipeline**

6. Experiments

The experiments are designed to evaluate the two primary contributions: First, the caption cleaning strategy and second, the weakly-supervised training of the CNN. Specifically, we investigate: (a) the effect of caption filtering and the weakly supervised approach on the quality of captions generated in terms of accuracy (Sec 6.3.1) and diversity (Sec 6.3.2), (b) the generalizability of the captions learnt from AVA, when tested on other image-caption datasets (Sec 6.3.3), (c) subjective or human opinion about the performance of the proposed framework (Sec 6.3.4).

6.1. Datasets

AVA-Captions: The compiled AVA-Captions dataset is discussed in detail in Section 3. We use 230,698 images and 1,318,359 comments for training; and 9,362 images for validation.

AVA raw-caption: The original AVA dataset provided by Murray *et al.* [58] and the raw unfiltered comments are used to train the framework in order to observe the effects of caption filtering.

Photo Critique Captioning Dataset (PCCD): This dataset was introduced by [11] and is based on www.gurushots.com. Professional photographers provide comments for the uploaded photos on seven aspects: general impression, composition and perspective, color and lighting, subject of photo, depth of field, focus and use of camera, exposure and speed. In order to verify whether

the proposed framework can generate aesthetic captions for images beyond the AVA dataset we trained it with AVA-Captions and tested it with PCCD. For a fair comparison, we use the same validation set provided in the original paper.

6.2. Baselines

We compare three implementations: (a) **Noisy - Supervised (NS):** NeuralTalk2 [46] framework trained on AVA-Original. It has an ImageNet trained CNN, followed by LSTM trained on raw, unfiltered AVA comments. NeuralTalk2 is also used as a baseline for AIC in [11]. (b) **Clean - Supervised (CS):** The LSTM of the NeuralTalk2 is trained on AVA-Captions *i.e.* filtered comments. The CNN is same as NS *i.e.* Imagenet trained. (c) **Clean and weakly-supervised (CWS):** NeuralTalk2 framework, where the CNN is trained with weak-supervision using LDA and the language model is trained on AVA-Captions.

6.3. Results and Analysis

6.3.1 Accuracy

Most of the existing standards for evaluating image captioning such as BLEU (B) [62], METEOR (M) [7], ROGUE (R) [40], CIDEr (C) [78] etc. are mainly more accurate extensions of the brute-force method [13] *i.e.* comparing the n-gram overlap between candidate and reference captions. Recently introduced metric SPICE (S) [2] instead compares scene graphs computed from the candidate and reference captions. It has been shown that SPICE captures semantic similarity better and is closer to human judgement more than the rest. Traditionally, SPICE is computed between the candidate and all the reference captions. A variant of SPICE (which we refer to as S-1) is used in [11] where the authors compute SPICE between the candidate and each of the reference captions and choose the best. In this paper, we report both S and S-1.

From Table 1(a), we observe that both CS and CWS outperform NS significantly over all metrics. Clearly, training the framework with cleaner captions results in more accurate outputs. On the other hand, the performance of CWS and CS is comparable. We argue that this indicates that the proposed weakly-supervised training strategy is capable of training the CNN as efficiently as a purely supervised approach and extract meaningful aesthetic features. Additionally as mentioned in Sec 1, the proposed CWS approach has an advantage that it does not require expensive human annotations to train. Thus, it is possible to scale to deeper architectures, and thus learn more complex representations simply by crawling the vast, freely-available and weakly-labelled data from the web.

Method	B1	B2	B3	B4	M	R	C	S	S-1
NS	0.379	0.219	0.122	0.061	0.079	0.233	0.038	0.044	0.135
CS	0.500	0.280	0.149	0.073	0.105	0.253	0.060	0.062	0.144
CWS	0.535	0.282	0.150	0.074	0.107	0.254	0.059	0.061	0.144

(a) Accuracy

Method	Train	Val	S-1	P	R
CNN-LSTM-WD	PCCD	PCCD	0.136	0.181	0.156
AO	PCCD	PCCD	0.127	0.201	0.121
AF	PCCD	PCCD	0.150	0.212	0.157
CS	AVA-C	PCCD	0.144	0.166	0.166
CWS	AVA-C	PCCD	0.142	0.162	0.161

(b) Generalizability

Table 1. (a) **Results on AVA-Captions:** Both CS and CWS, trained on AVA-Captions perform significantly better than NS, which is trained on noisy data. Also, the performance of CWS and CS is comparable, which proves the effectiveness of the weakly supervised approach (b) **Generalization results on PCCD:** Models trained on AVA-C perform well on PCCD validation set, when compared with models trained on PCCD directly. We argue that this impressive generalizability is achieved by training on a larger and diverse dataset.

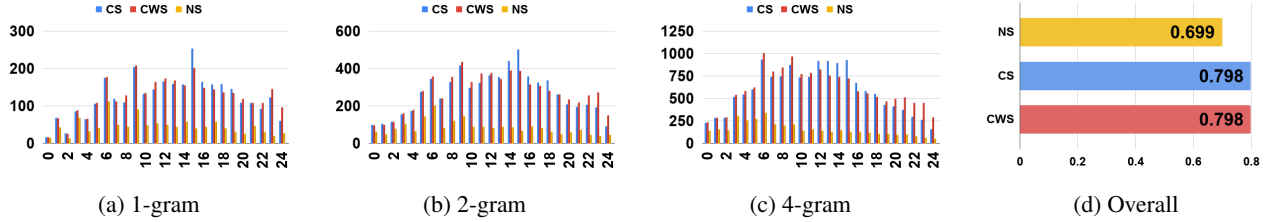


Figure 5. **Diversity:** Figures (a) - (c) report diversity of captions following [3]. The x -axes correspond to n -gram positions in a sentence. The y -axes correspond to the number of unique n -grams at each position, for the entire validation set. Figure (d) plots the overall diversity, as reported in [11]. We observe that the diversity of the captions increase significantly when the framework is trained on cleaner ground-truth *i.e.* AVA-Captions (CS or CWS) instead of AVA-Original (NS).

6.3.2 Diversity

Image Captioning pipelines often suffer from monotonicity of captions *i.e.* similar captions are generated for the validation images. This is attributed to the fact that the commonly used cross-entropy loss function trains the LSTM by reducing the entropy of the output word distribution and thus giving a *peaky* posterior probability distribution [3]. As mentioned earlier in Section 1, this is more pronounced in AIC due to the vast presence of the *easy* comments in the web. Diversity of the captions is usually measured by overlap between the candidate and the reference captions. We evaluate diversity following two state-of-the-art approaches [3, 11]. In [11], the authors define that two captions are different if the ratio of common words between them is smaller than a threshold (3% used in the paper). In [3], from the set of all the candidate captions, the authors compute the number of unique n -grams (1, 2, 4) at each position starting from the beginning up to position 13.

We plot diversity using [11] in Figure 5d. We compute using the alternative approach of [3] in Figure 5(a-c) but up to 25 positions since on an average the AVA captions are longer than the COCO captions. From both, we notice that diversity of NS is significantly lesser than CS or CWS. We observe that NS ends up generating a large number of “safe” captions such as “I like the composition and colours” or “nice shot” *etc.* We argue, that our caption filtering strategy reduces the number of useless captions from the data and thus the network learns more accurate and informative

components.

6.3.3 Generalizability

We wanted to test whether the knowledge gained by training on a large-scale but weakly annotated dataset is generic *i.e.* transferable to other image distributions. To do so, we train our frameworks on AVA-Captions and compare them with the models from [11], trained on PCCD. Everything is tested on the PCCD validation set. The models used by [11] are: (a) CNN-LSTM-WD is the NeuralTalk2 framework trained on PCCD. (b) Aspect oriented (AO) and (c) Aspect fusion (AF) are supervised methods, trained on PCCD. Note, that all the models are based on the NeuralTalk2 framework [46] and hence comparable in terms of architecture.

In Table 1(b), we observe that both CS and CWS outperform CNN-LSTM-WD and AO in S-1 scores. AF is still the best strategy for the PCCD dataset. Please note, both AO and AF are supervised strategies and require well defined “aspects” for training the network. Hence, as also pointed out in [11], it is not possible to train these frameworks on AVA as such aspect-level annotations are unavailable. However, we observe that both CS and proposed CWS, trained on AVA-Captions score reasonably well on PCCD. They are also generic strategies which can be easily mapped to other captioning tasks with weak supervision. We argue that the observed generalization capacity is due to training with a large and diverse dataset.

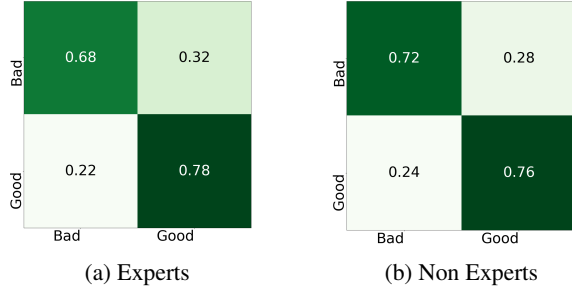


Figure 6. **Subjective evaluation of caption filtering:** The matrix compares our scoring strategy and human judgement for distinguishing a *good* and a *bad* caption. The rows stand for our output, and the columns represent what humans thought. We observe that the proposed caption filtering strategy is fairly consistent with what humans think about the informativeness of a caption.

6.3.4 Subjective (Human) Evaluation

Human judgement is still the touchstone for evaluating image captioning, and all the previously mentioned metrics are evaluated based on how well they correlate with the same. Therefore, we perform quality assessment of the generated captions by a subjective study. Our experimental procedure is similar to Chang *et al.* [11]. We found 15 participants with varying degree of expertise in photography (4 experts and 11 non-experts) to evaluate our framework. In order to familiarize the participants with the grading process, a brief training with 20 trials was provided beforehand. The subjective evaluation was intended to assess: (a) whether the caption scoring strategy (Equation 2) is consistent with human judgement regarding the same (b) the effect of cleaning on the quality of generated captions.

(a) Consistency of Scoring Strategy: We chose 25 random images from the validation set, and from each image, we selected 2 accepted and 2 discarded captions. During the experiment the subject was shown an image and a caption, and was asked to give a score on a scale of 100. In Figure 6a and 6b, we plot our predictions and human judgement in a confusion matrix. We find that our strategy is fairly consistent with what humans think as a good or a bad caption. Interestingly, with the experts, our strategy produces more false positives for bad captions. This is probably due to the fact that our strategy scores long captions higher, which may not always be the case and is a limitation.

(b) Effect of Caption Filtering: Similarly, 25 random images were chosen from the validation set. Each image had 3 captions, the candidates generated by NS, CS and CWS frameworks. During each trial, the subject was shown an image and one of the captions and asked to rate it into one of the categories - Good, Average and Bad. These categories follow from [11] and the participants were asked to rate a caption based on whether it conveyed enough information about a photograph. We observe in Table 2 the

Method	Experts				Non-Experts			
	Good (3)	Com (2)	Bad (1)	Avg	Good (3)	Com (2)	Bad (1)	Avg
NS	0	80	20	1.80	0	84	16	1.84
CS	8	84	8	2.0	28	68	4	2.24
CWS	4	80	16	1.88	20	72	8	2.12

Table 2. **Subjective comparison of baselines:** We observe that human subjects find CS and CWS to be comparable but both significantly better than NS. This underpins the hypothesis derived from the quantitative results that filtering improves the quality of generated captions and the weakly supervised features are comparable with the ImageNet trained features

percentage of good, common and bad captions generated by each method.

We observe that humans did not find any caption from NS to be good. Most of them were common or bad. This is due to its high tendency to generate the short, safe and common captions. Humans find CS to be performing slightly better than CWS which can probably be attributed to the lack of supervision during training the CNN. But as mentioned in Section 1, semi-supervised training is effective in practical scenarios due the easy availability of data and it might be worth investigating whether it is possible to improve its performance using more data and more complex representations. Additional qualitative results are provided in Figure 1 and also the supplementary material.

7. Conclusion

In this work, we studied aesthetic image captioning which is a variant of natural image captioning. The task is challenging not only due to its inherent subjective nature but also due to the absence of a suitable dataset. To this end, we propose a strategy to clean the weakly annotated data easily available from the web and compile AVA-Captions, the first large-scale dataset for aesthetic image captioning. Also, we propose a new weakly-supervised approach to train the CNN. We validated the proposed framework thoroughly, using automatic metrics and subjective studies.

As future work, it could be interesting to explore alternatives for utilizing the weak-labels and exploring other weakly-supervised strategies for extracting rich aesthetic attributes from AVA. It could also be interesting to extend this generic approach to other forms of captioning such as visual storytelling [38] or stylized captioning [56] by utilizing the easily available and weakly labelled data from the web.¹

¹This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776

References

- [1] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. Dont just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018. 3
- [2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 3, 6
- [3] J. Aneja, A. Deshpande, and A. G. Schwing. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5561–5570, 2018. 3, 7
- [4] L. Anne Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2016. 3
- [5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015. 1
- [6] T. O. Aydin, A. Smolic, and M. Gross. Automated aesthetic analysis of photographic images. *IEEE transactions on visualization and computer graphics*, 21(1):31–42, 2015. 3
- [7] S. Banerjee and A. Lavie. 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*, pages 65–72, 2005. 6
- [8] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010. 3
- [9] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pages 333–342. ACM, 2010. 3
- [10] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 2, 5
- [11] K.-Y. Chang, K.-H. Lu, and C.-S. Chen. Aesthetic critiques generation for photos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3514–3523, 2017. 1, 2, 3, 6, 7, 8
- [12] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416, 2013. 3
- [13] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5804–5812, 2018. 6
- [14] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. *Computer Vision–ECCV 2006*, pages 288–301, 2006. 3
- [15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 1, 2
- [16] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011. 3
- [17] S. K. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3277, 2014. 3
- [18] C. Doersch and A. Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017. 2
- [19] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 3
- [20] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 3
- [21] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010. 3
- [22] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 2
- [23] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, page 201422953, 2015. 3
- [24] D. Gershgorin. The data that transformed ai research and possibly the world, 2017. 2
- [25] K. Ghosal, M. Prasad, and A. Smolic. A geometry-sensitive approach for photographic style classification. 1, 3
- [26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [27] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 3
- [28] S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang. Learning hypergraph-regularized attribute predictors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 409–417, 2015. 4
- [29] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality

- similarity for multinomial data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2407–2414. IEEE, 2011. 3
- [30] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization. *arXiv preprint arXiv:1506.06272*, 2015. 3
- [31] J. Johnson, A. Karpathy, and L. Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016. 3
- [32] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011. 3
- [33] D. Jurafsky. *Speech & language processing*. Pearson Education India, 2000. 4
- [34] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell. Recognizing image style. In *BMVC 2014*, 2014. 1, 3
- [35] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 3
- [36] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 419–426. IEEE, 2006. 3
- [37] A. Kell. Where does data come from?, 2018. 2
- [38] R. Kiros. neural-storyteller, 2015. 1, 8
- [39] S. Li. Topic modeling and latent dirichlet allocation (lda) in python, 2018. 5
- [40] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 6
- [41] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1
- [42] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2
- [43] E. Loper and S. Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 3
- [44] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 457–466. ACM, 2014. 1, 3
- [45] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 990–998, 2015. 3
- [46] R. Luo. An image captioning codebase in pytorch, 2017. 5, 6, 7
- [47] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. *Computer Vision–ECCV 2008*, pages 386–399, 2008. 3
- [48] S. Ma, J. Liu, and C. Wen Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 3
- [49] L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 497–506, 2016. 3
- [50] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015. 3
- [51] G. Malu, R. S. Bapi, and B. Indurkha. Learning photography aesthetics with deep cnns, 2017. 3
- [52] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 3
- [53] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2533–2541, 2015.
- [54] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014. 3
- [55] L. Marchesotti, N. Murray, and F. Perronnin. Discovering beautiful attributes for aesthetic image analysis. *International journal of computer vision*, 113(3):246–266, 2015. 4
- [56] A. Mathews, L. Xie, and X. He. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8591–8600, 2018. 1, 2, 8
- [57] I. Misra, A. Gupta, and M. Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017. 4
- [58] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE, 2012. 1, 4, 6
- [59] T. Nagarajan and K. Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018. 4
- [60] P. Obrador, M. A. Saad, P. Suryanarayan, and N. Oliver. Towards category-based aesthetic models of photographs. In *International Conference on Multimedia Modeling*, pages 63–76. Springer, 2012. 3
- [61] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151, 2011. 3
- [62] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for

- Computational Linguistics, 2002. 3, 6
- [63] D. Parikh and K. Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, 2011. 4
 - [64] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1
 - [65] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577. ACM, 2008. 5
 - [66] J. Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142, 2003. 4
 - [67] A. Rana, P. Singh, G. Valenzise, F. Dufaux, N. Komodakis, and A. Smolic. Deep tone mapping operator for high dynamic range images. *Transaction of Image Processing*, 2019. 1
 - [68] A. Rana, J. Zepeda, and P. Perez. Feature learning for the image retrieval task. In *Asian Conference on Computer Vision (ACCV)*, pages 152–165. Springer, 2014. 3
 - [69] Z. Ren and Y. Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–771, 2018. 2
 - [70] R. Y. Rubinfeld and D. P. Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media, 2013. 5
 - [71] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR 2011*, pages 1745–1752. IEEE, 2011. 4
 - [72] J. San Pedro, T. Yeh, and N. Oliver. Leveraging user comments for aesthetic aware image search reranking. In *Proceedings of the 21st international conference on World Wide Web*, pages 439–448. ACM, 2012. 3
 - [73] R. Santa Cruz, B. Fernando, A. Cherian, and S. Gould. Neural algebra of classifiers. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 729–737. IEEE, 2018. 4
 - [74] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2556–2565, 2018. 1
 - [75] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1):207–218, 2014. 3, 4
 - [76] C. Sun, C. Gan, and R. Nevatia. Automatic concept discovery from parallel text and visual corpora. In *Proceedings of the IEEE international conference on computer vision*, pages 2596–2604, 2015. 3
 - [77] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 21(2):42–70, 2014. 3
 - [78] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 3, 4, 6
 - [79] S. Vittayakorn, T. Umeda, K. Murasaki, K. Sudo, T. Okatani, and K. Yamaguchi. Automatic attribute discovery with neural activations. In *European Conference on Computer Vision*, pages 252–268. Springer, 2016. 3
 - [80] J. Wang, Y. Cheng, and R. Schmidt Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2295–2304, 2016. 4
 - [81] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 1, 3
 - [82] T. Yashima, N. Okazaki, K. Inui, K. Yamaguchi, and T. Okatani. Learning to describe e-commerce images from noisy online data. In *Asian Conference on Computer Vision*, pages 85–100. Springer, 2016. 2, 3
 - [83] E. Zerman, A. Rana, and A. Smolic. Colornet - estimating colorfulness in natural images. In *The International Conference on Image Processing (ICIP)*, 2019. 2
 - [84] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017. 4
 - [85] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1644, 2014. 4


Supplementary Material: Aesthetic Image Captioning from Weakly-Labelled Photographs

We provide additional results regarding four different aspects discussed in the paper.


1. In Section 1, we provide more examples of the proposed caption filtering strategy and the datasets. In Table 1, we provide 5 randomly selected images from AVA and the corresponding original, unfiltered comments and the scores assigned to each comment by our algorithm. It is followed by Table 2, where we compare the statistics of AVA-Captions and PCCD datasets.
2. In Section 2, we provide additional results from re-labelling AVA images by topic modelling. In Table 3, we show top-10 n-grams from 10 randomly sampled topics and 16 random images which were labelled as that topic. In Section 2.2, we discuss “granularity” of the topics discovered *i.e.* whether the n-gram grouping makes sense. In Figure 1, we show how efficiently the CNN learns the discovered topics by plotting the confusions in prediction.
3. In Section 3, Table 4, we provide details about the CNN and LSTM architecture used.
4. In Section 4, Figure 2, we illustrate more qualitative results. We show 10 images, randomly selected from the validation set and the candidate captions generated by NS, CS and our approach. We also show how human subjects graded each caption.

1. Caption Filtering Results and Datasets


Table 1: **Images, Caption and scores:** Additional outputs of the caption cleaning strategy. We show images, corresponding captions and the score assigned by our algorithm to each caption. The false positives are highlighted in red, with explanations provided at the bottom of each image.

Images	Comments	Scores
	I appreciate the humor. Your efforts should be commended.	15.30
	Haha, both of you? That's something! I have to tell my girl-friend about that!	11.79
	It is nuts to get up between 4-5am. More power to you. :-)	8.59
	Colors are a little too green and there's no central composition.	33.58
	imo would have been better if you had desaturated the red channel to get rid of the glow on the phone (distracting)	47.96


(a) The image was submitted for a challenge titled “4:00-5:00 AM” *i.e.* the photographs submitted should have been captured during that time. It shows an empty bed of the photographer in low light. Barely visible, there is a phone in the bottom-left corner, which has a reddish-glow and is a bit distracting. Our strategy discarded the first three comments and kept the last two.

	Would have liked this bigger. I don't think you needed to try to show the entire long side of the house.	25.17
	There is little contrast (either in color or angle) to this image, so it feels a bit flat.	29.34
	very nice shot, could be better with a stronger contrast, but I like it nevertheless.	16.99
	very interesting shot.	5.47
	Why posting so small pictures?	7.53


(b) The image was submitted for a challenge titled “Abandoned Buildings”. It shows an abandoned hut, and the image is taken in greyscale. However, the resolution of the image is low and the small details in the image are hard to see. Our strategy accepted comment 1,2 and discarded 3,4,5. Comment 3 looks like a false negative. It carries information about the low contrast but the presence of too-frequent n-grams such as “nice shot” or “like” lowers the overall score.

Images	Comments	Scores
	Wow, 1000 shots! You deserve a medal for persistence. Great job.	42.03
	Beautiful lighting and colours....such a serene image. Great work!	38.33
	Simply cool, well done my friend.	16.57
	That's pretty awesome! beautiful.	17.70
	beautiful blues, flow and compositon	32.36
	Beautiful. The blue is perfect for the dark background and the sense of moving light is great.	49.86

(c) The title of the challenge is “abstract”. The image shows smoke captured using a slow shutter speed. It is mentioned in the description, that the photographer made 1000 attempts to take the correct shot, which is referred to in some comments. Comment 1 could be called a false positive. It does not say much about the photographic attributes. But the presence of unigrams such as “persistence”, “medal” highers the score.

	That horse looks nervous :)	3.65
	Damn freakin' stupendous. Love it!	12.39
	hahaha i love this one.....	4.39
	Very creative and well done. Lighting is fine. You might consider a smaller crop to remove wasted space a little all the way around, but particularly on the left side. That will focus more attention on your main subject and provide them more room in the composition without losing anything significant from the background.	102.04
	comp could be tighter, negative space on the left, color excellent, excellent texture, great lighting, good dof, great movement, very humorous, ok lines, overall a great attempt	123.53
	now thats funny I like it. the horse looks a bit scared, wait where are you going to put THAT?	12.46

(d) The challenge is titled “anachronism”. It is funny and shows a horse-rider filling the horse at a fuel station. We observe that comment 4,5 are detailed and long and represent the “critique club” comments. Of all the comments in AVA-Captions, these are the most informative ones. Our strategy assigns high score to these comments.

Images	Comments	Scores
	Aha, Snoopy, the star go-anywhere dog...	42.99
	I discovered the dog in second sight. now this iimage deserves one point more!	39.93
	If 'Fido' does his job well 'barking' we'll see a plenty of leafs in Spring Summer and Fall!	57.04
	I like it! I like it a lot. The perspective and use of your lens is superb.	36.95
	The dog just cracked me up.	3.36
	Love the colors in this shot. And the curved horizon gives the shot a neat earthy-globe feel.	58.91
	there is a dog in that there tree!	13.13
	ha! its like Where's Waldo! lol. beautiful texture throughout the photo	45.24
	That little terrier probably chewed the tree up... nice composition.	20.80

(e) The challenge is called “A single tree”. The image shows a dead tree and a dog sitting on the bottom right branch. While some accepted captions provide meaningful descriptions for the image, there are quite a few false-positives (red highlight). They are noisy annotations. Please note, that our strategy primarily aims to remove “safe” captions. The false positives observed here are noisy, but not safe. They are too-specific to this image and not generic descriptions such as “*very interesting shot*”. Thus, while our strategy successfully handles “safe” captions, it needs to be improved for the too specific cases.

Properties	AVA-Captions	PCCD
Number of images	240,060	3,840
Train/Val	230,698/9,362	3840/300
Number of captions	1,318,359	30,254
Captions per image $\mu(\sigma)$	5.48(4.86)	7.08(4.48)
Words per caption $\mu(\sigma)$	22.34(12.86)	18.58(10.96)
Longest caption (words)	137	163

Table 2. **Datasets** : A comparison of datasets used for the experiments.

A difference between AVA-Captions and PCCD lies in how the captions are pre-processed. In PCCD, a feedback on a particular aspect of an image is provided by a single user and the comment is a reasonably long paragraph with multiple sentences. During training in [1], the paragraph is split based on delimiters and each sentence is used as a separate caption. For AVA-Captions we preserve the original sentence structure and use the entire block as one caption. This is based on the observation that in a multi-line comment, sentences are not independent and together, they make much more sense. Due to this pre-processing, an average AVA caption is longer than an average PCCD caption whereas the average number of captions is more in PCCD (Table 2)

2. Relabelling AVA using LDA

2.1. 10 Random Topics (Top-10 N-grams) and Corresponding Images

Table 3: Top-10 N-grams from 10 randomly chosen topics and 16 randomly selected images corresponding to that topic. Here, we show additional examples relating to the discussion in Section 3.2 in the original paper. Note, that a “Topic” is an abstract concept made up of collection of n-grams. Therefore, there are no “names” for a topic. In the following figure, we show top-10 words (from the term-topic matrix) from each topic and 16 randomly selected images which were relabelled as that topic by applying LDA model on the corresponding captions. While some topic-image pairs are visually consistent, some are ambiguous (highlighted in red).

N-grams and Images

Topic 158: “ model”, “ pose”, “ hair”, “ skin”, “ skin tone”, “ body”, “ shoulder”, “ great pose”, “ beautiful model”, “ woman“



(a) The topic can be roughly labelled as “*fashion photography*”. The images are quite similar and coherent in that sense. (1,3) and (1,4) are false positives. They have comments that use the word “model” as in the model of the car.

Topic 108: “ depth”, “ field”, “ depth field”, “ great depth”, “ shallow depth”, “ nice depth”, “ good depth”, “ use depth”, “ little depth”, “ front“



(b) This topic consist of images for which “*depth of field*” plays an important role. Most of the photographs were captured using a shallow depth of field and has a blurred background. For others, the critics suggested using a shallower depth of field.

N-grams and Images

Topic 84: “space”, “negative space”, “use negative”, “empty space”, “space top”, “pepper”, “space left”, “space right”, “good use”, “much negative”



(c) “*Negative space*” refers to empty space in a photograph. It is an important compositional technique adopted by photographers to draw attention to the main subject by surrounding it with blank space. The images shown reflect this technique with a lot of empty areas in the image.

Topic 36: “rule”, “third”, “rule third”, “centered composition”, “example”, “good example”, “use rule”, “good use”, “use third”, “intersection”



(d) “*The Rule of Thirds*” is one the most commonly used rule in photography. It is applied by dividing the whole frame into a 3×3 grid and placing the subjects at the intersection or along of horizontal and vertical lines. In all the images, either the rule is applied or the subject is too centered and the critic urged to crop the picture such that it follows the rule.

Topic 101: “flash”, “bit harsh”, “lighting bit”, “clock”, “little bright”, “bright”, “much light”, “harsh”, “lighting harsh”, “light bit”



(e) We can roughly assign the label “*harsh or extreme lighting condition*” to this topic. It can also be observed that most of the images are quite under or over exposed and most of the comments suggest using a flash or less light.

N-grams and Images

Topic 192: “area”, “heh”, “loss”, “dark area”, “bright white”, “pier”, “white area”, “bright area”, “ladder”, “darker”



(f) As mentioned earlier, not all topics are consistent with the images. Some topics created by the LDA are not coherent enough and thus the images lack visual consistency. During training, these ambiguous topics lead to ambiguity. We choose $K = 200$ as it results in minimum perplexity or most coherent topics.

Topic 185: “reflection”, “water”, “nice reflection”, “reflection water”, “great reflection”, “color reflection”, “reflection nice”, “mirror”, “light reflection”, “reflection great”



(g) All the n-grams and images probably refer to the topic label “*water and reflection*”, a common strategy used to capture images by using reflections. Reflections add symmetry to a photograph and they are often applied in the case of images of rivers, especially at night.

Topic 132: “building”, “tower”, “structure”, “architecture”, “church”, “roof”, “arch”, “skyline”, “top”, “cityscape”



(h) The n-grams are similar and refer to the concept of a “*building*” and are also consistent with the corresponding images. Architectural photography is an important discipline of photography.

N-grams and Images

Topic 198: “concept”, “great concept”, “nice concept”, “good concept”, “good idea”, “concept good”, “nice idea”, “concept execution”, “cool concept”, “chip”



(i) This is another example of an ambiguous topic. n-grams such as “nice concept” or “good concept” probably co-occur together in case of photographs with an interesting story and reflect out of the box thinking by the photographer. However, the set of such concepts is non-exhaustive and could be anything. Thus the images lack visual consistency.

Topic 80: “line”, “power”, “nice line”, “leading line”, “curve”, “great line”, “power line”, “diagonal line”, “line color”, “line nice”



(j) The topic can be labelled “*leading or vanishing lines*” which is also a common technique used to guide the viewer to the main subject of the image. By using lines in the image such as roads, stairs, railings or railway tracks it is possible to guide the eyes of the viewer to the main subject at the end of the line. The images grouped together are visually quite consistent.

2.2. Granularity of Topics

In [3], the authors perform probabilistic Latent Semantic Analysis (p-LSA) on the raw AVA Comments and attempt a similar attribute discovery with $k = 50$. They report that p-LSA topics thus discovered were not granular enough *i.e.* the grouping was not photographically meaningful. We observe that simple LDA on raw comments also performs poorly. But, with some modifications it can be used to discover fairly meaningful topics. The key implementation steps are highlighted below.

1. Instead of running LDA on raw comments we perform LDA on AVA-Captions. Our cleaning strategy removes a major chunk of vague comments, resulting in more coherent and granular topics.
2. Unlike [3], we create the vocabulary by combining unigrams and bigrams. This strategy essentially groups concepts such as “line”, “nice line”, “leading line”, “great line”, “power line” together, resulting in more coherent concepts such as “*about leading lines*”.
3. We observe that the 200 bigrams (or ugly and beautiful attributes) discovered in [3] follow a typical pattern; in each bigram, an adverb, adjective or a noun is followed by an adjective or a noun. For example the top 5 beautiful bigrams

are “nice colors”, “beautiful scene”, “nice perspective”, “big congrats”, “so cute” and the top 5 ugly attributes are “too small”, “distracting background”, “snap shot”, “very dark”, “bad focus”. Based on this observation, we restrict the vocabulary to contain only unigrams and bigrams of certain patterns. For unigrams we choose the nouns. For bigrams we choose cases where the first term is a noun, adjective or adverb and the second term is a noun or an adjective. This helps us to get rid of less meaningful unigrams and bigrams such as “cool”, “superb”, “not seeing”, “not sure ” *etc.*

4. We restrict the vocabulary to only those words which occur in less than 10% of the comments. Thus, too frequent attributes or *photographic stopwords* such as “composition”, “shot”, “nice composition” *etc.* are ignored during topic modelling.

We provide the full list of topics, number of images each topic and the top-10 words for the reviewers’ perusal in a separate CSV file called “Topic_Words.txt”.

2.3. Training the CNN: Confusion Matrix

A CNN is trained with these topics as labels and the training performance was visualized using a confusion matrix in Figure 1. The rows represent the predicted labels and columns represent the topic labels.

3. Architecture

CNN		LSTM	
Parameter	Value	Parameter	Value
Name	ResNet101	Name	neuraltalk2[2]
Input Size	256×256	Model	fc
Output Size	200×1	Beam Size	2
Batch Size	64	Learning Rate	$1 \times e^{-5}$
Learning Rate	$1 \times e^{-5}$	Vocabulary Size	10048
Optimizer	ADAM	Training Images	230,698
Loss	Binary Cross Entropy	Validation Images	9,362
		Max length of comment	16
		Word Count Threshold	40

Table 4. **Architecture** : We report only those parameters which were modified from the implementation in [2]. The rest of the parameters have been set to default and can be found in [2]. We tried other available models too. But these parameters were chosen as a trade-off between performance and speed. Note, that our proposed strategy is generic and can be tried out using any of the models.

4. Qualitative Results

In this section, we show additional qualitative results and compare how the different baselines performed for the same photograph. In Figure 2, we show an image and three captions generated by NS, CS and CWS (in the same order), respectively. These images were also used during the subjective evaluation by humans. We report the average score obtained by each baseline. The errors are highlighted in red with additional discussion in the figure caption.

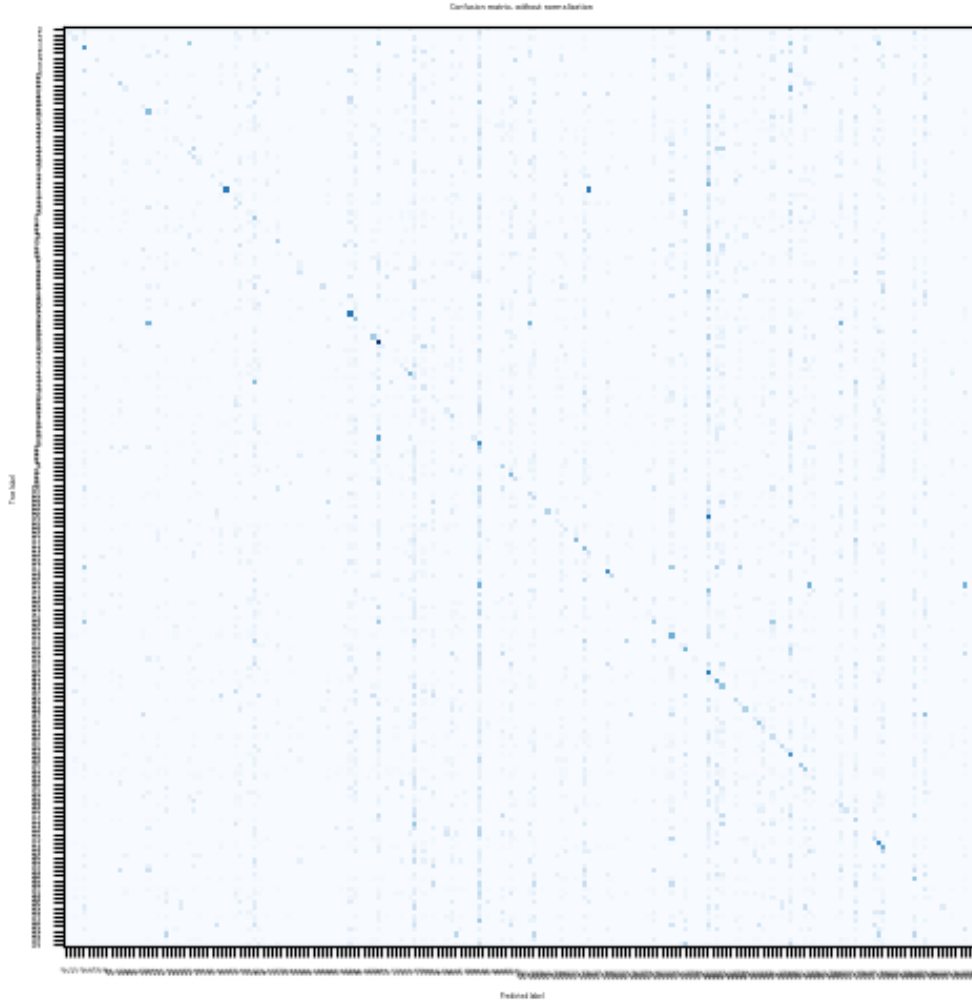


Figure 1. **Confusion Matrix with 200 Topics:** Note that in order to plot the confusion matrix, we select the most probable topic as the true topic for an image. But this is essentially a multi-class multi-label problem where an image may be assigned to multiple topics with equal probabilities. Therefore, the confusions are not always *strictly-hard* confusions. In fact the topics are highly overlapping and probably, a confusion matrix is not the best way to visualize the performance of such multi-class multi-label problems. However, a “visible” diagonal is still indicative of the fact that the attributes are granular enough for the CNN to learn certain photographic aspects.

References

- [1] K.-Y. Chang, K.-H. Lu, and C.-S. Chen. Aesthetic critiques generation for photos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3514–3523, 2017. 4
- [2] R. Luo. An image captioning codebase in pytorch, 2017. 9
- [3] L. Marchesotti, N. Murray, and F. Perronnin. Discovering beautiful attributes for aesthetic image analysis. *International journal of computer vision*, 113(3):246–266, 2015. 8







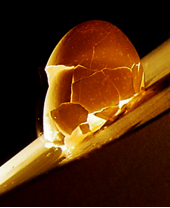


Images	Captions	Scores	Images	Captions	Scores
	<p>i like the angle and the composition 62.00</p> <p>i like the idea , but i think it would have been better if the door was in focus . 33.91</p> <p>i like the composition , but i think it would have been better if you could have gotten a little more of the building . 74.91</p>			<p>i like the composition and the dof 65.18</p> <p>i like the composition and the shallow dof . i like the composition and the dof . 57.73</p> <p>i like the idea , but i think it would have been better if the background was a little more in focus . 72.36</p>	
(a)			(b)		
	<p>i like the colors and the composition 55.91</p> <p>i like the colors and the water . the water is a little distracting . 41.27</p> <p>i like the composition and the colors . the water is a little too bright . 69.00</p>			<p>i like the composition and the bw 60.55</p> <p>i like this shot . i like the way the lines lead the eye into the photo . 69.82</p> <p>i like the composition and the bw conversion . 67.55</p>	
(c)			(d)		
	<p>i like the composition and the lighting 57.45</p> <p>i like the way the light hits the face and the background . 66.09</p> <p>this is a great shot . i love the way the light is coming from the left . 66.00</p>			<p>nice shot 9.00</p> <p>i like the composition and the colors , but i think it would have been better if you could have gotten a little closer to the subject . 46.00</p> <p>i like the angle and the angle of the shot . the sky is a little too bright . 65.64</p>	
(e)			(f)		
	<p>i like the composition and the lighting 54.00</p> <p>i like the way you used the shallow depth of field . i like the way you used the shallow dof . 63.45</p> <p>i like the idea , but i think it would have been better if you had a little more light on the keys . 65.36</p>			<p>i like the idea but the lighting is a bit harsh 51.36</p> <p>i like the idea , but i think it would have been better if the focus was a little sharper . 72.73</p> <p>i like the idea , but the lighting is a bit harsh . 58.09</p>	
(g)			(h)		
	<p>i like the lighting and the composition 58.00</p> <p>i love the simplicity of this shot . the lighting is perfect . 74.00</p> <p>i like the idea , but the lighting is a bit harsh. 35.91</p>			<p>i like the simplicity of this shot 58.00</p> <p>i like the simplicity of this shot . i like the simplicity of the shot . 53.82</p> <p>i like the idea , but i think it would have been better if the bird was in focus . 9.00</p>	
(i)			(j)		

Figure 2. **Qualitative Results:** Each image has three captions generated by NS, CS and CWS, in order; and the corresponding scores by human subjects. While some captions are satisfactory, some are problematic (highlighted in red). For example, in (b) and (g), the candidate by CS has repetition of “depth of field” and “dof”, which are the same concepts. We did not try merging the acronyms with the actual concepts in the vocabulary. In our candidate in (f) we notice the repetition of “angle”. This is a generic problem associated with any form of image captioning. We did not address this either. In (i) and (j), the captions that our model generates are wrong analyses of the photographs. In (i) the lighting is not harsh and in (j) there is no bird. Note, that the candidates generated by NS are short and mostly less informative. They are also graded poorly by human subjects. The “good”, “common” and “bad” categories were defined by quantizing the scores at the intervals of [0, 33, 66, 100].