Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition

Yufei Wang[1]   Zhe Lin[2]   Xiaohui Shen[2]   Scott Cohen[2]   Garrison W. Cottrell[1]

[1]University of California, San Diego   [2]Adobe Research

{yuw176, gary}@ucsd.edu   {zlin, xshen, scohen}@adobe.com

Abstract

Recently, there has been a lot of interest in automatically generating descriptions for an image. Most existing language-model based approaches for this task learn to generate an image description word by word in its original word order. However, for humans, it is more natural to locate the objects and their relationships first, and then elaborate on each object, describing notable attributes. We present a coarse-to-fine method that decomposes the original image description into a skeleton sentence and its attributes, and generates the skeleton sentence and attribute phrases separately. By this decomposition, our method can generate more accurate and novel descriptions than the previous state-of-the-art. Experimental results on the MS-COCO and a larger scale Stock3M datasets show that our algorithm yields consistent improvements across different evaluation metrics, especially on the SPICE metric, which has much higher correlation with human ratings than the conventional metrics. Furthermore, our algorithm can generate descriptions with varied length, benefiting from the separate control of the skeleton and attributes. This enables image description generation that better accommodates user preferences.

1. Introduction

The task of automatically generating image descriptions, or image captioning, has drawn great attention in the computer vision community. The problem is challenging in that the description generation process requires the understanding of high level image semantics beyond simple object or scene recognition, and the ability to generate a semantically and syntactically correct sentence to describe the important objects, their attributes and relationships.

The image captioning approaches generally fall into three categories. The first category tackles this problem based on retrieval: given a query image, the system searches for visually similar images in a database, finds and transfers the best descriptions from the nearest neighbor captions for the description of the query image [11, 20, 26, 34].
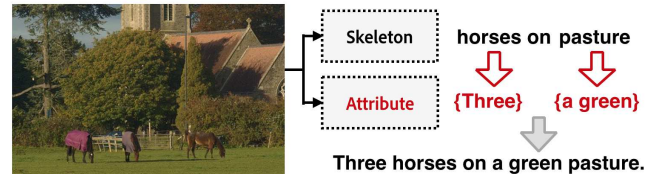

Figure 1: Illustration of the inference stage of our coarse-to-fine captioning algorithm with skeleton-attribute decomposition. First, the skeleton sentence is generated, describing the objects and relationships. Then the objects are revisited and the attributes for each object are generated.

The second category typically uses template-based methods to generate descriptions that follow predefined syntactic rules[17, 25, 28, 14, 46, 32]. Most recent work falls into the third category: language model-based methods [16, 42, 45, 12, 31, 23]. Inspired by the machine translation task [37, 3, 7], an image to be described is viewed as a "sentence" in a source language, and an Encoder-Decoder network is used to translate the input to the target sentence. Unlike machine translation, the source "sentence" is an image in the captioning task. Therefore, a natural encoder is a Convolutional Neural Network (CNN) instead of a Recurrent Neural Network (RNN).

Starting from the basic form of a CNN encoder-RNN decoder, there have been many attempts to improve the system. Inspired by their success in machine translation, Long-short Term Memory (LSTM) networks are used as the decoder in [42, 12]. Xu *et al.* [45] add an attention mechanism that learns to attend to parts of the image for word prediction. It is also found that feeding high level attributes instead of CNN features yields improvements [47, 44].

Despite the variation in approaches, most of the existing LSTM-based methods suffer from two problems: 1) they tend to parrot back sentences from the training corpus, and lack variation in the generated captions [10]; 2) due to the word-by-word prediction process in sentence generation, attributes are generated before the object they refer to. Mixtures of attributes, subjects, and relations in a complete sentence create large variations across training samples, which can affect training effectiveness.

In order to overcome these problems, in this paper, we propose a coarse-to-fine algorithm to generate the image description in a two stage manner: First, the skeleton sentence of the image description is generated, containing the main objects involved in the image, and their relationships. Then, the objects are revisited in a second stage using attention, and the attributes for each object are generated if they are worth mentioning. The flow is illustrated in Figure 1. By dealing with the skeleton and attributes separately, the system is able to generate more accurate image captions.

Our work is also inspired by a series of Cognitive Neuroscience studies. During visual processing such as object recognition, two types of mechanisms play important roles: first, a fast subcortical pathway that projects to the frontal lobe does a coarse analysis of the image, categorizing the objects [5, 15, 18], and this provides top-down feedback to a slower, cortical pathway in the ventral temporal lobe [40, 6] that proceeds from low level to high level regions to recognize an object. The exact way that the top-down mechanism is involved is not fully understood, but Bar [4] proposed a hypothesis that low spatial frequency features trigger the quick "initial guesses" of the objects, and then the "initial guesses" are back-projected to low level visual cortex to integrate with the bottom-up process.

Analogous to this object recognition procedure, our image captioning process also comprises two stages: 1) a quick global prediction of the main objects and their relationship in the image, and 2) an object-wise attribute description. The objects predicted by the first stage are fed back to help the bottom-up attribute generation process. Meanwhile, this idea is also supported by object-based attention theory. Object based attention proposes that the perceptual analysis of the visual input first segments the visual field into separate objects, and then, in a focal attention stage, analyzes a particular object in more detail [33, 13].

The main contributions of this paper are as follows: First, we are the first to divide the image caption task such that the skeleton and attributes are predicted separately. Second, our model improves performance consistently against a very strong baseline that outperforms the published state-of-the-art results. The improvement on the recently proposed SPICE [1] evaluation metric is significant. Third, we also propose a mechanism to generate image descriptions with variable length using a single model. The coarse-to-fine system naturally benefits from this mechanism, with the ability to vary the skeleton/attribute part of the captions separately. This enables us to adapt image description generation according to user preferences, with descriptions containing a varied amount of object/attribute information.

2. Related Work

Existing image captioning methods Retrieval-based methods search for visually similar images to the input image, and find the best caption from the retrieved image captions. For example, Devlin *et al*. in [11] propose a K-nearest neighbor approach that finds the caption that best represents the set of candidate captions gathered from neighbor images. This method suffers from an obvious problem that the generated captions are always from an existing caption set, and thus it is unable to generate novel captions.

Template-based methods generate image captions from pre-defined templates, and fill the template with detected objects, scenes and attributes. Farhadi *et al*. [17] use single ⟨object, action, scene⟩ triple to represent a caption, and learns the mapping from images and sentences separately to the triplet meaning space. Kulkarni *et al*. [25] detect objects and attributes in an image as well as their prepositional relationship, and use a CRF to predict the best structure containing those objects, modifiers and relationships. In [27], Lebret *et al*. predict phrases from an image, and combine them with a simple language model to generate the description. These approaches heavily rely on the templates or simple grammars, and so generate rigid captions.

Language model-based methods typically learn the common embedding space of images and captions, and generate novel captions without many rigid syntactical constraints. Kiros and Zemel [22] propose multimodal log-bilinear models conditioned on image features. Mao *et al*. [31] propose a Multimodal Recurrent Neural Network (MRNN) that uses an RNN to learn the text embedding, and a CNN to learn the image representation. Vinyals *et al*. [42] use LSTM as the decoder to generate sentences, and provide the image features as input to the LSTM directly. Xu *et al*. [45] further introduce an attention-based model that can learn where to look while generating corresponding words. You *et al*. [47] use pre-generated semantic concept proposals to guide the caption generation, and learn to selectively attend to those concepts at different time-steps. Similarly, Wu *et al*. [44] also show that high level semantic features can improve the caption generation performance.

Our work is also a language-model-based method. Unlike approaches to LSTM-based methods that try to feed a better image representation to the language model, we focus on the caption itself, and show how breaking the original word order in a natural way can yield better performance.

Analyzing the sentences for image captioning Parsing of a sentence is the process of analyzing the sentence according to a set of grammar rules, and generating a rooted parse tree that represents the syntactic structure of the sentence [24]. There is some language-model-based work that parses the captions for better sentence encoding. For example, Socher *et al*. [36] proposed the Dependency Tree-RNN, which uses dependency trees to embed sentences into a vector space, and then performs caption retrieval with the embedded vector. Unfortunately, the model is unable to generate novel sentences.
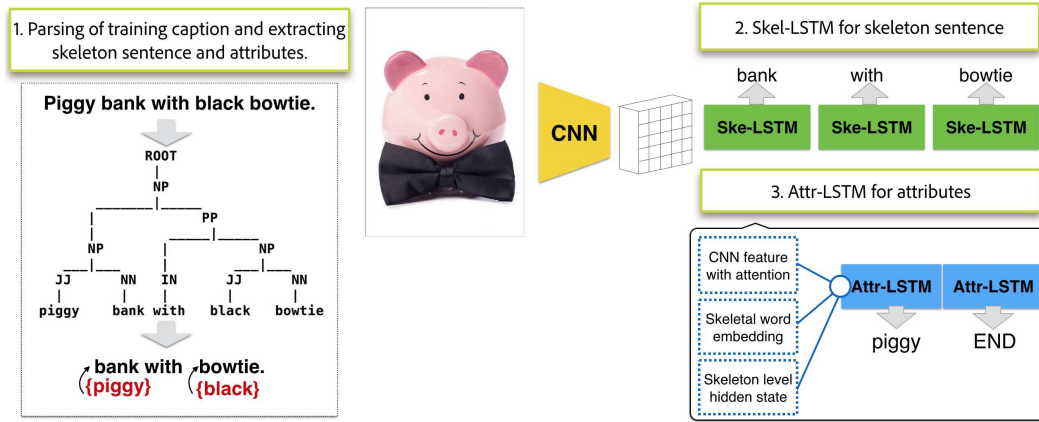
Figure 2: The overall framework of the proposed algorithm. In the training stage, the training image caption is decomposed into the skeleton sentence and corresponding attributes. A Skel-LSTM is trained to generate the skeleton based on the main objects and their relationships in the image, and then an Attr-LSTM generates attributes for each skeletal word.

The work that is closest to our own is the hierarchical LSTM model proposed by Tan and Chan [39]. They view captions as a combination of noun phrases and other words, and try to predict the noun phrases (together with the other words) directly with an LSTM.The noun phrases are encoded into a vector representation with a separate LSTM. In the inference stage, K image-relevant phrases are generated first with the lower level LSTM. Then, the upper level LSTM generates the sentence that contains both the "noun phrase" token and other words. When a noun phrase is generated, suitable phrases from the phrase pool are selected, and then used as the input to the next time-step. This work is relevant to ours in that it also tries to break the original word order of the caption. However, it directly replaces the phrases with a single word "phrase token" in the upper level LSTM without distinguishing those tokens, although the phrases can be very different. Also, the phrases in an image are generated ahead of the sentence generation, without knowing the sentence structure or the location to attend to.

Evaluation metrics Evaluation of image caption generation is as challenging as the task itself. Bleu [35], CIDEr [41], METEOR [9], and ROUGE [29] are common metrics used for evaluating most image captioning benchmarks such as MS-COCO and Flickr30K. However, these metrics are very sensitive to n-gram overlap, which may not necessarily be a good way to measure the quality of an image description. Recently, Anderson *et al*. [2] introduced a new evaluation metric called SPICE that overcomes this problem. SPICE uses a graph-based semantic representation to encode the objects, attributes and relationships in the image. They show that SPICE has a much higher correlation with human judgement than the conventional evaluation metrics.

In our work, we evaluate our results using both conventional metrics and the new SPICE metric. We also show how unimportant words like "a" impact scores on conventional metrics.

3. The Proposed Model

The overall framework of our model is shown in Figure 2. In the training stage, the ground-truth captions are decomposed into the skeleton sentences and attributes for the training of two separate networks. In the test stage, the skeleton sentence is generated for a given image, and then attributes conditioned on the skeleton sentence are generated. They are then merged to form the final generated caption.

3.1. Skeleton-Attribute decomposition for captions

To extract the skeleton sentence and attributes from a training image caption, we use the Stanford constituency parser [24, 30]. As shown in Figure 2, the parser constructs a constituency tree from the original caption, while the nodes hierarchically form phrases of different types. The common phrase types are Noun phrase (NP), Verb phrase (VP), Prepositional phrase (PP), and Adjective phrase (AP).

To extract the objects in the skeleton sentence, we find the lowest level NP's, and keep the last word within the phrase as the skeletal object word. The words ahead of it within the same NP are attributes describing this skeletal object. The lowest level phrases of other types are kept in the skeleton sentence.

Sometimes, it is difficult to decide whether all the words except for the last one in a noun phrase are attributes. For example, the phrase "coffee cup" is a noun-noun compound. Should we keep "coffee cup" as a single entity, or use "coffee" as a modifier? In this work, we don't distinguish noun-noun compounds from other attribute-noun word phrases, and treat "coffee" as the attribute of "cup". Our experience is that the coarse-to-fine network can learn the correspondence, although strictly speaking they are not attribute-object pairs.

3.2. Coarse-to-fine LSTM

We use the high level image features extracted from a CNN as the input to the language model. For the decoder part, our coarse-to-fine model consists of two LSTM submodels: one for generating skeleton sentences, and the other for generating attributes. We denote the two submodels as Skel-LSTM and Attr-LSTM respectively.

Skel-LSTM The Skel-LSTM predicts the skeleton sentence given the image features. We adopt the soft attention based LSTM in [45] for the Skel-LSTM. Spatial information is maintained in the CNN image features, and an attention map is learned at every time step to focus attention to predict the current word.

We denote the image features at location $(i, j) \in L \times L$ as $v_{ij} \in \mathbb{R}^D$. The attention map at time step t is represented as normalized weights $\alpha_{ij,t}$, computed by a multilayer perceptron conditioned on the previous hidden state h_{t-1}.

$$\alpha_{ij,t} = \text{Softmax}(\text{MLP}(v_{ij}, h_{t-1})) \qquad (1)$$

Then, the context vector z_t at time t is computed as:

$$z_t = \sum_{i,j} \alpha_{ij,t} v_{ij} \qquad (2)$$

The context vector is then fed to the current time step LSTM unit to predict the upcoming word.

Unlike [45], in our model, the attention map $\alpha_{ij,t}$ is not only used to predict the current skeletal word, but also to guide the attribute prediction: the attributes corresponding to a skeletal word describe the same skeletal object, and the attention information we get from Skel-LSTM can be reused in the Attr-LSTM to guide where to look.

Attr-LSTM After the skeleton sentence is generated, the Attr-LSTM predicts the attribute sequence for each skeletal word. Rather than predicting multiple attribute words separately for one object, the Attr-LSTM can predict the attribute sequence as a whole, naturally taking care of the order of attributes. The Attr-LSTM is similar to the model in [42], with several modifications.

The original input sequence of the LSTM in [42] is:

$$x_{-1} = \text{CNN}(I) \qquad (3)$$

$$x_t = W_e y_t, t = 0, 1, ..., N - 1 \qquad (4)$$

where I is the image, $\text{CNN}(I)$ is the CNN image features as a vector without spatial information, W_e is the learned word embedding, and y_t is the ground-truth word encoded as a one-hot vector. y_0 is a special start-word token.

In our coarse-to-fine framework, attribute generation is conditioned on the skeletal word it is describing. Therefore, apart from the image features, the Attr-LSTM should be informed by the current skeletal word. On the other hand, the context of the skeleton sentence is also important to give the Attr-LSTM a global understanding of the caption,

rather than just focusing on the single current skeletal word. We experimented with feeding the skeletal hidden activations from different time steps into the Attr-LSTM, including the previous time step, the current time step, and the final time step, and found that the current time step hidden activations yield the best result. Moreover, as mentioned in Skel-LSTM, rather than using global image features as the input, we use attention-based image features to encourage the attribute predictor to focus on the current skeletal word.

We formulate the input of Attr-LSTM at the first time step as a multilayer network that fuses different sources of information into the embedding space:

$$x_{-1} = \text{MLP}(W_I z_T + W_t s_T^{skel} + W_h h_T^{skel}) \qquad (5)$$

where T is the time step of the current skeletal word, $z_T \in \mathbb{R}^D$ is the attention weighted average of the image features, $s_T^{skel} \in \mathbb{R}^{m_s}$ is the embedding of the skeletal word at time T, $h_T^{skel} \in \mathbb{R}^{n_s}$ is the hidden state in the Skel-LSTM of dimension n_s. m_s and n_s are dimensionality of the Skel-LSTM word embedding, and the LSTM units, respectively. W_l, W_t, W_h are learned parameters. The remaining input to Attr-LSTM is the same as Equation 4. The Attr-LSTM framework is illustrated in Figure 2.

In the training stage, the ground truth skeleton sentence is fed into the Skel-LSTM, and s_T^{skel} is the ground truth skeleton word embedding. In test stage, s_T^{skel} is the embedding of predicted skeleton word.

Attention refinement for attribute prediction Optionally, we can refine the attention map acquired in the Skel-LSTM for better localization of the skeletal word, thus improving the attribute prediction. The attention map α is a pre-word α that is generated before the word is predicted. It can cover multiple objects, or can even be in a different location from the predicted word. Therefore, a refinement of the attention map after the prediction of the current word can provide more accurate guidance for the attribute prediction.

The LSTM unit at time step T outputs the word probability prediction $P_{attend} = (p_1, p_2, ..., p_Q)$, where Q is the vocabulary size in Skel-LSTM. In addition to the single weighted sum feature vector z_T, we can also use the feature vector v_{ij} in each location as input to the Skel-LSTM. Thus, for each of the L^2 locations, we can get the probability of word prediction P_{ij}. We can use the spatial word probability to refine the attention map α:

$$\alpha_{post(ij)} = \frac{1}{Z} P_{attend}^T \cdot P_{ij} \qquad (6)$$

where Z is the normalization factor so that $\alpha_{post(ij)}$ sums to one. The refined post-word α is proportional to the similarity between P_{attend} and P_{ij}. In Figure 3, we illustrate the attention refinement process.

Fusion of Skeleton-Attributes After attributes are predicted for all the skeletal words, attributes are merged into
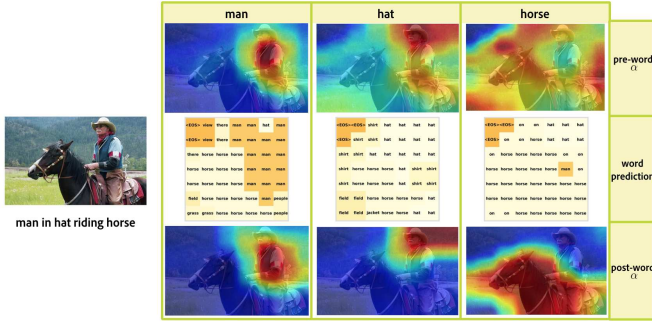
Figure 3: Illustration of attention refinement process. Due to limited space, only three object words are shown from the predicted caption "man in hat riding horse". For each word, the attention map, predicted words for each location, and refined attention map are shown. We provide more examples in the supplementary material.

the skeleton sentence just before the corresponding skeletal word, and the final caption is formed.

3.3. Variable-length caption generation

Due to the imperfections in the current parser approach that we use, there are some cases where the parsing result is noisy. Most of the time, the noise is from incorrect noun phrase recognition, and short skeleton sentences with one or several missing objects. This leads to a shorter skeleton prediction in the Skel-LSTM on average, thus eventually causes shorter predictions for the full sentence.

To overcome this problem, we designed a simple yet effective trick to vary the length of the generated sentence. Without modifying the trained network, In the inference stage of either Skel-LSTM or Attr-LSTM, we modify the sentence probability with a length factor:

$$\log(\hat{P}) = \log(P) + \gamma \cdot l \tag{7}$$

Where P is the probability of a generated sentence, and $\hat{P}$ is the modified sentence probability. l is the length of the generated sentence. γ is the length factor to encourage or discourage longer sentences. Note that the modification is performed during generation of the each word rather than performed after the whole sentence is generated. It is equivalent to adding γ to each word log probability except for the end-of-sentence token $\langle EOS \rangle$ when sampling the next word from the word probability distribution. This trick of sentence probability modification works well together with beam search.

Our coarse-to-fine algorithm especially benefits from this mechanism, since it can be applied to either Skel-LSTM or Attr-LSTM, resulting in varied information in either objects, or the description of those objects. This allows us to generate captions according to user preference on the complexity of captions and amount of information in captions.

4. Experiments

In this section, we describe our experiments on two datasets to test our proposed approach.

4.1. Datasets

We perform experiments on two datasets: the popular benchmark MS-COCO, and Stock3M, a new dataset with much larger scale and more natural captions.

MS-COCO has 123,287 images. Each image is annotated with 5 human generated captions, with an average length of 10.36 words. We use the standard training/test/validation split that is commonly used by other work [47, 44], and use 5000 images for testing, and 5000 images for validation.

MS-COCO is a commonly used benchmark for image captioning tasks. However, there are some issues with the dataset: the images are limited and biased to certain content categories, and the image set is relatively small. Moreover, the captions generated by AMT workers are not particularly natural. Therefore, we collected a new dataset: Stock3M. Storck3M contains 3,217,654 user uploaded images with a large variety of content. Each image is associated with one caption that is provided by the photo uploader on a stock website. The caption given by the photo uploader is more natural than those found in MS-COCO, and the dataset is 26 times larger in terms of number of images. The captions are much shorter than MS-COCO, with an average length of 5.25 words, but they are more challenging, due to a larger vocabulary and image content variety. We use 2000 images for validation and 8000 images for testing.

4.2. Experimental details

Preprocessing of captions We follow the preprocessing procedure in [21] for the captions, removing the punctuation and converting all characters to lower case. For MS-COCO, we discard words that occur fewer than 5 times in skeleton sentences, and fewer than 3 times in attributes. This results in 7896 skeleton, and 5199 attribute words. In total, there are 9535 unique words. For the baseline method that processes the full sentences, a similar preprocessing procedure is applied to the full sentences. Words that occur less than 5 times are discarded, resulting in 9567 unique words.

For Stock3M, due to the larger vocabulary size, we set the word occurrence thresholds to 30 for skeleton and 5 for attributes respectively. This results in 11047 skeleton and 12385 attribute words, with a total of 14290 unique words. In the baseline method that processes full sentences, the occurrence threshold is 30, resulting in 13788 unique words.

Image features and training details for MS-COCO It has been argued that high level features such as attributes are better as input to caption-generating LSTMs [47, 44].

Our empirical finding is that by simply adopting a better network architecture that provides better image features, and fine-tuning the CNN within the caption dataset, the features extracted are already excellent inputs to the LSTM. We use ResNet-200 [19] as the encoder model. Images are resized to 256×256 and randomly cropped to 224×224. The layer before the average pooling layer and classification layer is used for the image features. and it outputs features with size $2048 \times 7 \times 7$, maintaining the spatial information.

Our system is implemented in Torch [8]. We fine-tune the CNN features as follows: first, the CNN features are fixed, and an LSTM is trained for full sentence generation. After the LSTM achieves reasonable results, we start fine-tuning the CNN with learning rate 1e-5. The fine-tuned CNN is then used for both Skel-LSTM and Attr-LSTM. The parameters for the Decoder network are as follows: word embedding is trained from scratch, with a dimension of 512. For Skel-LSTM, we set the learning rate 0.0001, and the hidden layer dimension 1800. For Attr-LSTM, the learning rate is 0.0004, and the hidden layer is 1024-dimensional. Adagrad is used for training. The learning rate is cut in half once after the validation loss stops dropping.

Image features and training details for Stock3M We use GoogleNet [38] fine-tuned on Stock3M as the CNN encoder, and add an embedding module after the 1024-dimensional output of GoogleNet $pool5/7 \times 7s1$ layer.

Stock3M is different from MS-COCO in that the images mostly contain single objects, and the captions are more concise than MS-COCO. The average length of Stock3M captions is about half that of MS-COCO. Hence, we did not observe improvement with the attention mechanism, because there are fewer things to focus on. For simplicity, we use the LSTM in [42] for Skel-LSTM. Consequently, for Attr-LSTM, there is no attention input in the -1 time step. We will show that even without attention, the coarse-to-fine algorithm improves substantially over baseline.

Parameters in the testing stage For both Skel-LSTM and Attr-LSTM, we use a beam search strategy, and adopt length factor γ as explained in Section 3.3. The beam size and value of γ are chosen using the validation set, and are provided in supplementary material.

4.3. Results

Evaluation metrics Apart from the conventional evaluation metrics that are commonly used: Bleu [35], CIDEr [41], METEOR[9], and ROUGE [29], we use the recently proposed SPICE metric [2], which is not sensitive to n-grams and builds a scene graph from captions to encode the objects, attributes and relationships in the image. We emphasize our performance on this metric, because it has much higher correlation with human ratings than the other conventional metrics, and it shows the performance specific to different types of information, such as different types of attributes, objects, and relationships between objects.

Baseline In order to demonstrate the effectiveness of our method, we also present a baseline result. The baseline method is trained and tested on full caption sentences, without skeleton-attribute decomposition. For each dataset, we use the same network architecture as in the Skel-LSTM architecture, and use the same hyper-parameters and the same CNN encoder as in our proposed coarse-to-fine method.

Quantitative results We report both SPICE in Table 1 and conventional evaluation metrics in Table 2.

First, it is worth noting that our baseline method is a very strong baseline. In Table 2, we compare our method with published state-of-the-art methods. Our baseline method already outperforms the state-of-the-art by a considerable margin, indicating the importance of a powerful image feature extractor. By just fine-tuning the CNN with the simple baseline algorithm, we outperform the approaches with augmentation of high level attributes [47, 44]. The baseline already ranks 3rd - 4th place on the MS-COCO CodaLab leaderboard[1]. Note that we use no augmentation tricks such as ensembling, or scheduled sampling [43], which can improve the performance further. We provide our submission to the leaderboard in the supplementary material.

SPICE is an F-score of the matching tuples in predicted and reference scene graphs. It can be divided into meaningful subcategories. In Table 1 we report the SPICE score as well as the subclass scores of objects, relations and attributes. In particular, size, color, and count attributes are reported. Table 1 shows consistent improvement over baseline for the two datasets, and this extends to the subcategories. The cardinality F-score for Stock3M is not reported here because there are too few images with this type of attribute to have a meaningful evaluation: there are only 78 cardinality attributes out of 8000 test images.

In Table 2, we also show the comparison between the proposed method and baseline method on conventional evaluation metrics. As shown, there is no significant improvement over baseline on most of the conventional metrics on MS-COCO. This is due to an intrinsic problem with the conventional metrics: they overly rely on n-gram matching. The proposed coarse-to-fine algorithm breaks the original word order of the training captions, and thus weakens the objective of predicting exact n-grams as in the training captions. There is even a small drop on BLEU-3 and BLEU-4 on MS-COCO against the baseline. To investigate if the two methods indeed have similar performance as reflected in those conventional metrics, we conducted further analysis of the results.

We preprocess the ground-truth and predicted captions to

---

[1]https://competitions.codalab.org/competitions/3221

Table 1: Performance of our proposed method and the baseline method on SPICE measurement, for the two datasets. We also include the results on different semantic concept subcategories.

	Model	SPICE	Precision	Recall	Object	Relation	Attribute	Size	Color	Cardinality
MS-COCO	Baseline	0.188	0.508	0.117	0.350	0.048	0.098	0.045	0.132	0.039
	Ours	**0.196**	**0.529**	**0.123**	**0.363**	**0.050**	**0.110**	**0.073**	**0.170**	**0.064**
Stock3M	Baseline	0.157	0.173	0.166	0.250	0.049	0.077	0.129	0.135	-
	Ours	**0.172**	**0.190**	**0.185**	**0.276**	**0.061**	**0.081**	**0.144**	**0.151**	-

Table 2: Performance of our proposed methods and other state-of-the-art methods on MS-COCO and Stock3M. Only scores that were reported in the papers are shown here.

Datasets	Models	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
MS-COCO	NIC [42]	-	-	-	0.277	0.237	-	0.855
	LRCN [12]	0.669	0.489	0.349	0.249	-	-	-
	Toronto [45]	0.718	0.504	0.357	0.250	0.230	-	-
	ATT [47]	0.709	0.537	0.402	0.304	0.243	-	-
	ACVT [44]	0.74	0.56	0.42	0.31	0.26	-	0.94
	Baseline	0.742	0.577	0.442	0.340	0.268	0.552	1.069
	Ours	0.742	0.577	0.440	0.336	0.268	0.552	1.073
	Baseline (w/o a)	0.664	0.481	0.351	0.258	0.245	0.485	0.949
	Ours (w/o a)	**0.673**	**0.489**	**0.355**	**0.259**	**0.247**	**0.489**	**0.966**
Stock3M	Baseline	0.236	0.133	0.079	0.050	0.108	0.233	0.720
	Ours	**0.245**	**0.138**	**0.083**	**0.052**	**0.110**	**0.239**	**0.724**
	Baseline (w/o a)	0.233	0.134	0.082	0.053	0.108	0.235	0.737
	Ours (w/o a)	**0.246**	**0.140**	**0.086**	**0.055**	**0.111**	**0.241**	**0.738**

remove all the a's in the captions. This is motivated by the observation that 15% of words in the MS-COCO captions are a. This function word affects the n-gram match greatly, though it conveys very little information in the MS-COCO like captions. Therefore, by removing the a's in the captions, we obtain a measure that is not influenced by the n-grams using a, and hence is more focused on content words. The performance evaluation on the same datasets with a removed is shown in Table 2 as "Baseline/Ours (w/o a)". It can be seen that consistent improvement is achieved with our coarse-to-fine method.

In Table 2, we also present the performance of our coarse-to-fine method as well as the baseline method on Stock3M evaluated on conventional metrics. In Stock3M, the frequency of the word a is only 2.5%, therefore it has no big impact on the relative performance of the two methods. We can see consistent improvement on all the metrics.

4.4. Analysis of generated descriptions

Generating variable-length captions. In the coarse-to-fine algorithm, a length factor is applied to the Skel-LSTM and Attr-LSTM separately to encourage longer skeleton/attribute generation in order to generate captions that have similar length to the training captions. However, we can further manually tune the length factor value to control the length of skeleton/attribute of the generated captions. In Figure 4, we show some test examples from Stock3M and MS-COCO . For each of the images, four cap-

tions are generated with four pairs of (skeleton, attribute) length factor values: (-1, -1), (1.5, -1), (-1, 1.5), (1.5, 1.5). The four value pairs represent all combinations of encouraging less/more information in skeleton/attributes. Attributes are marked in red in the generated caption. We can see how the length factor works together with beam search to get syntactically and semantically correct captions. The amount of object/attribute information naturally varies with the length of the skeleton/attributes.

We can certainly apply the same trick on the baseline method using different length factor values. For comparison, in Figure 4 (red box), we show the four captions generated from baseline method also using four different length factor values: $\gamma \in \{-1, -0.5, 0.5, 1.5\}$. As illustrated, although the captions generated by the baseline model can also have different lengths, they are much less flexible and useful than the ones generated by our coarse-to-fine model. This is because the coarse-to-fine model can decompose the caption into skeletons and attributes, and have separate requirements for objects and attributes according to user preference: the user may prefer descriptions that only describe main objects but in more detail; or he/she may prefer descriptions that contain all the objects in the images, but cares less about the object attributes.

Post-word α helps with attribute prediction The results we show in Table 1 and 2 for the proposed coarse-to-fine model adopts attention refinement for attribute prediction in the Attr-LSTM on MS-COCO. Here, we further
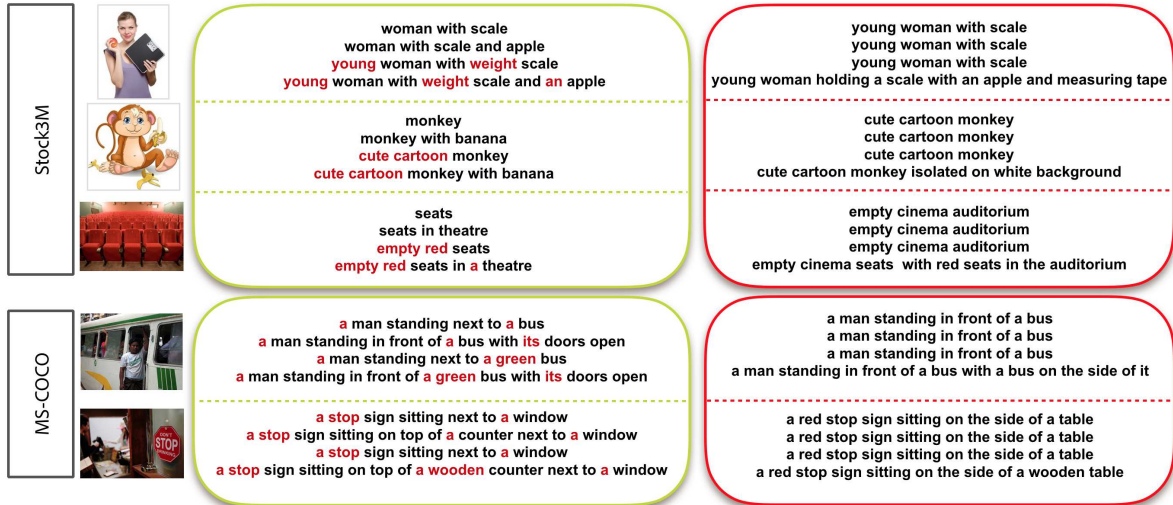
Figure 4: Examples of predicted titles for image examples from Stock3M and MS-COCO. Four titles are generated from our coarse-to-fine model (middle, in green box) and baseline model (right, in red box) respectively. For the coarse-to-fine model, four pairs of length factor value γ for skeletal title and attributes are (-1, -1), (1.5, -1),(-1, 1.5), (1.5, 1.5) respectively. For the baseline method, the γ's are -1, -0.5, 0.5, 1.5 respectively.

Table 3: Comparison of our proposed method with and without post-word α attention on MS-COCO.

Model	Attribute	Color	Size	Cardinality
Pre-word α	0.107	0.167	0.069	0.063
Post-word α	**0.110**	**0.170**	**0.073**	**0.064**

validate the effectiveness of the post-word α refinement approach in Table 3, by comparing the result without attention refinement (Pre-word α) with the result with attention refinement (Post-word α). The post-word α only refines the attended area for attribute prediction, therefore we only show the improvement of SPICE score on attribute subcategories. The performance on other categories is unchanged. We see consistent improvement across different types of attributes, especially on color and size. This shows that a good attention map can improve attribute prediction.

The ability to generate unique and novel captions It has been pointed out that the current LSTM based method has a problem generating sentences that have not been seen in the training set, and generates the same sentences for different test images [10]. This means that the LSTM dynamics are caught in a rut of repeating the sequences it was trained on for visually similar test images, and is less capable of generating unique sentences for a new image with an object/attribute composition that is not seen in the training set. With the skeleton-attribute decomposition, we claim that our algorithm can generate more unique captions, and can give more accurate attributes even when the attribute-object pattern is new to the system. As shown in Table 4, our coarse-to-fine model increases the percentage of generated unique captions by 3%, and increases the percentage of novel captions by 8%.

Table 4: Percentage of generated unique sentences and captions seen in training captions for the baseline method and our coarse-to-fine method. The statistics are gathered from the test set of MS-COCO containing 5000 images.

Model	Unique captions	Seen in training
Baseline	63.96%	56.06%
Coarse-to-fine	66.96%	47.76%

Qualitative result of generated captions In the supplementary material, we show more qualitative examples of generated captions from our coarse-to-fine model and baseline model.

5. Conclusion

In this paper, we propose a coarse-to-fine model for image caption generation. The proposed model decomposes the original image caption into a skeleton sentence and corresponding attributes, and formulates the captioning process in a natural way in which the skeleton sentence is generated first, and then the objects in the skeleton are revisited for attribute generation. We show with experiments on two challenging datasets that the coarse-to-fine model can generate better and more unique captions over a strong baseline method. Our proposed model can also generate descriptive captions with variable lengths separately for skeleton sentence and attributes, and this allows for caption generation according to user preference.

In future work, we plan to investigate more complicated skeleton/attribute decomposition approaches, and allow for attributes that appear after the skeletal object. It is also of interest to design a model that automatically decides on the length of generated caption based on the visual complexity of the image.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 2

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 3, 6

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. 1

[4] M. Bar. A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, 2003. 2

[5] J. Bullier. Integrated model of visual processing. *Brain Research Review*, 2001. 2

[6] J. Bulliera and L. G. Nowakb. Parallel versus serial processing: new vistas on the distributed organization of the visual system. *Current Opinion in Neurobiology*, 1995. 2

[7] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014. 1

[8] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011. 6

[9] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *In Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014. 3, 6

[10] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works. *CoRR*, abs/1505.01809, 2015. 1, 8

[11] J. Devlin, S. Gupta, R. B. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *CoRR*, 2015. 1, 2

[12] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 1, 7

[13] J. Duncan. Selective attention and the organization of visual information. *Journal of Experimental Psychology*, 1984. 2

[14] D. Elliott and F. Keller. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1292–1302, 2013. 1

[15] A. K. Engel, P. Fries, and W. Singer. Dynamic predictions: Oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, 2001. 2

[16] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[17] 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 *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 15–29, Berlin, Heidelberg, 2010. Springer-Verlag. 1, 2

[18] C. D. Gilbert, M. Sigman, and R. E. Crist. The neural basis of perceptual learning. *Neuron*, 31(5), 2001. 2

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *arXiv prepring arXiv:1506.01497*, 2015. 6

[20] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899, May 2013. 1

[21] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137, 2015. 5

[22] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In T. Jebara and E. P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603. JMLR Workshop and Conference Proceedings, 2014. 2

[23] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. 1

[24] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. 2, 3

[25] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2891–2903, Dec 2013. 1, 2

[26] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi. Treetalk: Composition and compression of trees for image descriptions. 1

[27] R. Lebret, P. H. O. Pinheiro, and R. Collobert. Phrase-based image captioning. *CoRR*, abs/1502.03671, 2015. 2

[28] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. 1

[29] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10, 2004. 3, 6

[30] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. 3

[31] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*, 2015. 1, 2

[32] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé, III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 747–756, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. 1

[33] U. Neisser. *Cognitive Psychology*. 1967. 2

[34] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011. 1

[35] 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*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 3, 6

[36] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2:207–218, 2014. 2

[37] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014. 1

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9, 2015. 6

[39] Y. H. Tan and C. S. Chan. phi-lstm: A phrase-based hierarchical LSTM model for image captioning. *CoRR*, abs/1608.05813, 2016. 3

[40] K. Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 1996. 2

[41] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575, 2015. 3, 6

[42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014. 1, 2, 4, 6, 7

[43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *CoRR*, abs/1609.06647, 2016. 6

[44] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel. What value do explicit high level concepts have in vision to language problems. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 2016. 1, 2, 5, 6, 7

[45] 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 D. Blei and F. Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2048–

2057. JMLR Workshop and Conference Proceedings, 2015. 1, 2, 4, 7

[46] Y. Yang, C. L. Teo, H. Daumé, III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 444–454, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. 1

[47] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. *CoRR*, abs/1603.03925, 2016. 1, 2, 5, 6, 7