Two Novel Approaches in Generation and Retrieval of Image Descriptions

Akhil Chaturvedi **Dhanashree Balaram** **Eti Rastogi** **Ilaï Deutel** **Kevin Chon**
Carnegie Mellon University
{aschatur, dbalaram, erastogi, ideutel, khchon}@andrew.cmu.edu

Abstract

Automatically generating accurate and helpful textual descriptions from images has emerged as a very challenging problem involving computer vision and natural language processing. Two problems with many current image captioning models are lack of adjective rich attribute integration in captions, and use of computationally expensive recurrent networks for training a language model. In this project, we develop a novel model for adjectives-enriched captions, solving the issue of sparse attributes. Additionally, we propose a new caption retrieval method which instead of recurrent networks, uses a new word embedding weighting scheme to generate sentence embeddings for captions solving the problem of computationally efficiency.

1 Introduction

Image captioning is an important problem in Artificial Intelligence and can be used in tasks such as image based messaging, aiding the visually impaired, etc. Earlier works do not exhibit high quality attributes (adjectives) in their generated captions. As can be seen in fig. 1, the generated captions do not do justice to the image attributes. Hence, there is a need to map images to their captions in a joint space, exploiting the mutual relationship between image representations and attributes for enhancing image description generation. It is imperative that we generate captions that have adjectives like 'big', 'small', 'fluffy', 'broad', etc.. since these adjectives would help the user understand the whole scene better.

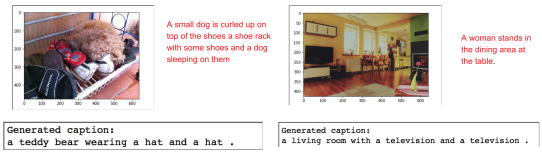


Figure 1: Failure cases from our baseline implementation of Google's Show and Tell [22]

Most recent image captioning methods involve training a complex Long Short Term Memory (LSTM) [7] or Gated Recurrent Unit (GRU) [2] model to generate descriptive sentences. These deep language models implicitly represent the embedding of a sentence or paragraph document in an embedding hyperspace from information that the image provides as features. These models usually have a large number of parameters that take a lot of time to train and require huge computing power [9] which can be a problem for low compute resource conditions.

In this paper, we introduce two methods for image captioning. One using LSTM-based attribute-boosted caption generation to solve the problem of lack of attributes in captions, and one using SIF-based caption retrieval from images to come up with an approach which forgoes the complexity

of LSTMs for training language models and is much less compute intensive. In this first method, which we call Att-Net, we design a feed forward neural network that learns the relationship between attributes and objects in an image. Our second approach presents an approach where we extend the idea of unsupervised encoded sentence embeddings called Smooth Inverse Frequency Model (SIF) [1] to an image caption retrieval approach and does not involve training a recurrent neural network.

2 Related Work

Research to improve image captioning branches out in three directions [23, 27]. The first direction consists in template-based methods. Pre-defined templates are filled with detected objects, attributes, scenes and actions. *BabyTalk* [12] generates tuples <object, attribute> from the image, as well as triplets <preposition, object_1, object_2> (for instance, preposition can be *near*, *against*, *besides*, ...). A conditional random field is constructed using pre-defined blocks (the template) and incorporating the previously generated unary potentials, which allows for labelling prediction and sentence generation from the most likely labelling. Other research works use hidden markov models [26], syntactic trees [15] or a constrain model on generated phrases [13].

The second direction consists in language-based methods. Language-based models learn probability distributions of captions conditioned on images. Show and Tell [22] famously describes a network architecture using a LSTM [7], as briefly described in section 3. The results are drastically improved by adding attention, which is validated by the state-of-the-art performance obtained at the time of publication on Flickr8k [8], Flickr30k [28] and MS COCO [25]. Dai *et al*. [3] describe a novel approach which is, to our knowledge, the first attempt to use a Conditional Generative Adversarial Network (CGAN) for image captioning. A generator, producing image descriptions, and an evaluator, assessing how well a description looks *real*, are jointly trained.

The third direction consists in retrieval-based methods. This approach produces image captions by copying captions from other images. For instance, Devlin *et al*. [4] explore a variety of nearest neighbor approaches for image captioning, explained in section 3.3. Each generated caption has a human-level quality and the model perform as well as several recent language-based model when measured using automatic evaluation metrics, but human studies have shown that methods generating novel captions are preferred [4].

A few authors have tried to improve language-based models using attributes. *Skeleton Key* [24] presents a method that decomposes the image into a skeleton sentence and attention using a *Ske-LSTM*, and generates a list of attributes for each word of the skeleton sentence using a *Attr-LSTM*. A newer framework, LSTM-A [27], uses Multiple Instance Learning for attributes detection, following Fang *et al*.'s method [5]. It was the leader on the MS COCO Captioning Leaderboard until November 13, 2017 [1]. More details about this paper are given in section 3.1.

3 Proposed Approaches

The problem of image captioning can be formalized as maximizing the probability of predicting a ground truth sentence, given an image and the model parameters [22]. Let I be an image, $W = (w_0, \ldots, w_N)$ a list of N words corresponding to the correct caption, and θ symbolizes the parameters of the model. We aim to maximize the probability of the correct description given I:

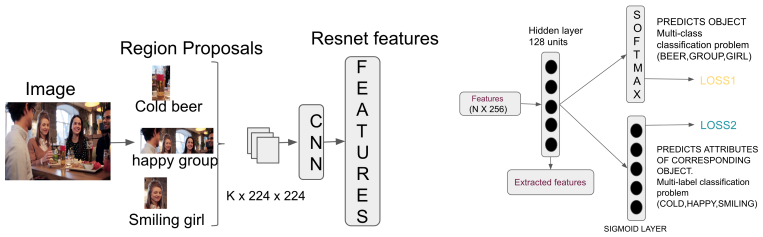
$$\theta^* = \text{argmin}_\theta \left(-\log p(W \mid I, \theta)\right) = \text{argmin}_\theta \left(-\sum_{t=0}^{N} \log(w_t \mid I, w_0, ...w_{t-1})\right) \tag{1}$$

3.1 Att-Net

We divide our problem of bolstering captions with attributes into two main tasks: generating attribute embeddings (figure 2) and generating image captions (figure 3).

We propose a model, Att-Net which generates a feature space rich in objects (nouns) and attributes (adjectives). Each object can have multiple attributes. We aim at finding a space where we can find similar objects with similar attributes close together. This relationship between an object and its

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



(a) Feature extraction network (b) Att-Net

Figure 2: Feature extraction and Att-Net architectures

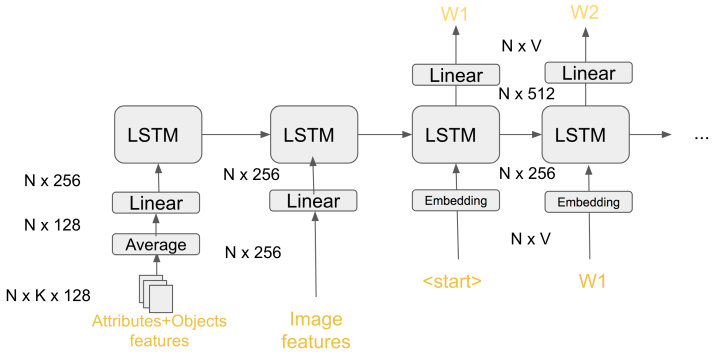


Figure 3: Caption generator model architecture

attributes can be learned by learning a multi-class multi-label classification task and extracting the penultimate layer of the network. It can be seen as an alternative to the Fang *et al.*'s weakly-supervised Multiple Instance Learning-based approach to extract objects/attributes vectors [5], which is used in the LSTM-A framework, the current state-of-the-art approach to image captioning [27]. Baseline LSTM-A models use the output probabilities of the most common objects and attributes for a given image as the attributes vector input.

Our approach for feature generation is different. Let $\mathcal{V}^A$ be the attributes vocabulary. We are given an image I, a sequence of objects $O = (o_1, \ldots, o_K)$ and their attributes $A = (a_1, \ldots, a_K)$, where $a_i \in \{0,1\}^{|\mathcal{V}^A|}$ and $\forall j \in \{1, \ldots, |\mathcal{V}^A|\}, a_i^j = \mathbb{1}_{\mathcal{V}_j^A \text{ describes } o_i}$. We decompose I into smaller images $\{\tilde{I}_i\}_{i \in \{1,\ldots,K\}}$, where $\tilde{I}_i$ represents object o_i and its set of attributes a_i. Each of the $\tilde{I}_i$ is fed into a deep neural network with two output layers, one with softmax activation representing the probability $\mathbb{P}_i^{obj}$ of each common object being represented in $\tilde{I}_i$, one with sigmoid activation representing the probability $\mathbb{P}_i^{attr}$ of each common attributes being represented in $\tilde{I}_i$. The total loss for image I is:

$$\mathcal{L}_{Att-Net} = \frac{1}{K} \sum_{i=1}^{K} \left(-\log \mathbb{P}_i^{obj}(o_i) - \frac{1}{|\mathcal{V}^A|} \sum_{j=1}^{|\mathcal{V}^A|} \left[a_i^{(j)} \log \mathbb{P}_i^{attr}(a_i^{(j)}) + (1 - a_i^{(j)}) \log(1 - \mathbb{P}_i^{attr}(a_i^{(j)}) \right] \right)$$

This correspond to a multi-class classification task with a log-softmax cross entropy loss for the object, and a multi-label classification task with a sigmoid binary cross entropy loss for the attributes. We use the Visual Genome dataset [11] for training Att-Net as it provides annotated objects and attributes for each bounding box of an image. We use a pre-trained ResNet-101 model [6] to extract image features, which are fed into our network, as shown on fig. 2a. During testing, a Regional Proposal Network[19, 14] pre-trained on Visual Genome[2] is applied on an image to extract objects bounding boxes.

The caption generator model takes attributes and image features as input in order to generate a caption which is enriched with attributes. We used a model similar to LSTM-A_3, from Yao *et al.* [27]. The

[2] `https://github.com/yikang-li/MSDN`

decoder consists of a LSTM model which tries to predict a word at every time-step conditioned on the previous hidden state and the previously generated words. Give an image I with K bounding boxes, we extract the K images using the bounding boxes coordinates and feed each of them to Att-Net. We obtain K object-attribute features, extracted from the penultimate layer of Att-Net. These features are then averaged and projected into a new space before being fed to an LSTM as the first input. The rest of the LSTM corresponds to the Show and Tell architecture [22]: the input at the next time step is image features, and captions words are subsequently used as inputs. Given image feature vectors f, averaged attribute-object features a and a caption W, the LSTM updating procedure is as:

$$x^{-2} = T_a a \qquad x^{-1} = T_f f$$

$$\forall t \in \{0, \ldots, n-1\}, x^t = T_e w_t \qquad \forall t \in \{-2, \ldots, n-1\}, h^t = \text{LSTM}(x^t)$$

where T_a, T_f and T_e are the attributes, images and textual transformation matrices.

The expected outputs for $t \in \{0, \ldots, n-1\}$ are w_{t+1}. A graphical representation of the model is shown on fig. 3. The caption generator is trained using equation 1.

3.2 Att-GAN

The paper *Towards Diverse and Natural Image Descriptions via a Conditional GAN* [3] describes a novel approach which is, to our knowledge, the first attempt to use a Conditional Generative Adversarial Network for image captioning. This approach is interesting for various theoretical reasons. If we assume that the evaluator E is perfect, it makes more sense to evaluate the generator G using the evaluator (is the generated caption good enough to fool E?) rather than using current methods, which are based on absolute similarity between generated and true captions. A good evaluator would also provide us with an interesting evaluation metrics, as detailed in section 3.2.2.

3.2.1 General Design

Generator: We use a Att-Net-based LSTM model as a generator as the one described in section 3.1. The generator network is pre-trained for 100 epochs with the same loss objective as mentioned earlier.

Evaluator : The aim of the evaluator is to differentiate between true captions and captions generated by the generator model. Our evaluator model generates a reward in the range $[0, 1]$. A reward of 0 indicates that the generator was not at all able to fool the evaluator, whereas reward of 1 means that the generator was completely successful in fooling the evaluator. The evaluator uses the same decoder architecture as the generator but with different parameters to generate captions. Given an image features f and a caption W, let h_d refer to the final hidden state generated by evaluator LSTM model over caption W. The reward generated by the evaluator is given as $r = \sigma(f \cdot h_d)$, where σ is the sigmoid function.

Let W_I^t be the true description for an image I and W_I^g be the description generated by the generator over an image I. We pretrain our evaluator model on the following loss :

$$\mathcal{L}_{\mathcal{E}} = -log(W_I^t) - log(1 - W_I^g)$$

The first term forces the evaluator to learn about true captions whereas the second term forces the evaluator to distinguish between the true descriptions and the generated ones, which in turn provide useful feedbacks to the generator pushing it to generate more realistic captions.

Adversarial Training : In adversarial training. we train both the generator and the discriminator simultaneously. The training objective of the generator is to fool the discriminator into classifying W_I^g as true caption. This can be achieved by making the generator learn to generate captions which are also capable of receiving a higher reward from the evaluator. The loss function for the generator is given by :

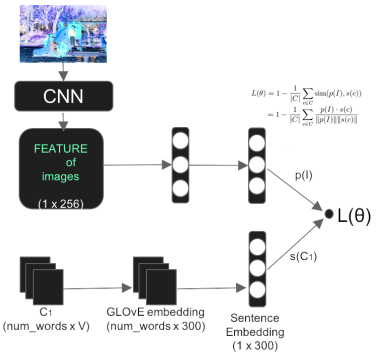
$$\mathcal{L}_{\dashv\lceil\sqsubseteq} = \sum_{t=1}^{T} [-log\mathbb{P}(w_t|w_{t-1}, h_{t-1})] - log(W_I^g)$$

3.2.2 Evaluation metrics

This model naturally provides an interesting evaluation metric, $E - GAN$ [3], which is the score given by the evaluator for a given input caption. The authors claim that "$E - GAN$ metrics are

Data: Word embeddings $\{v_w : w \in V\}$, a set of sentences S, parameter a and estimated probabilities $\{p(w) : w \in V\}$ of the words
Result: Sentence embeddings $\{v_W : W \in S\}$
for *sentence* $W \in S$ **do**
  $v_s \leftarrow \frac{1}{|W|} \sum_{w \in W} \frac{a}{a+p(w)}$;
Form a matrix X whose columns are $v_W : W \in S$ and let u be it's first singular vector ;
for *sentence* $W \in S$ **do**
  $v_W \leftarrow v_W - uu^T v_W$;

(a) Sentence Embedding algorithm using SIF



(b) Cosine loss-based multi-layer perceptron.

Figure 4: Our proposed approach to train image embeddings using the SIF method and the cosine-loss based multi-layer perceptron.

more consistent with human evaluations, while BLEU only favors those that significantly overlap with the training samples in detailed wording". Using $E - GAN$ as a complementary metric to n-grams-based methods seems promising, but we will need to be aware of the inherent bias (since the metrics depends directly on the model training and on the training data) that remains.

3.3 Caption Retrieval using Smooth Inverse Frequency Weightings

A recent work by Devlin *et al.* has shown that retrieval-based image captioning can perform as well as many generative approaches [4]. For a given test image I, image features are computed, using either GIST [16] or the `fc7` layer of a VGG16 neural network [20] (trained on the ImageNet task or fine-tuned for the classification of most commonly occurring words in image captions). k nearest-neighbor images are found among the training images based on the proximity of the selected image features. A set $C_k(I)$ of captions of these neigbors images is created, and a "consensus caption" is chosen among these: the caption that has the highest average lexical similarity to the the other captions in $C_k(I)$, where sim is either BLEU [17] or CIDEr [21]:

$$c^* = \text{argmax}_{c \in C_k(I)} \sum_{c' \in C_k(I)} sim(c, c') \tag{2}$$

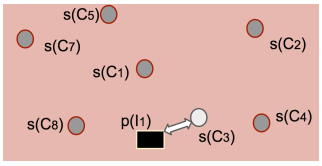
The main contribution of ours in the retrieval approach is the following (1) Integrate the SIF weighting strategy to get an unsupervised sentence embedding for captions that describe an image. (2) Project image features into a deep network which maps it to it's respective sentence embedding in the language hyperspace (words and sentences). Our idea of SIF embeddings is adapted from [1] on weighted average of word embeddings with removal of the projection of the average vectors on their first principal component. The weight of the word w is given by $a/(a + p(w))$ where a is a parameter and $p(w)$ is the (estimated) word frequency. SIF is highly analogous to TF-IDF weights.

Algorithm 4a shows how to compute SIF embeddings from sentences. We run the algorithm for each sentence in our caption corpus. For word embeddings, we use publicly available pre-trained GLOvE embeddings [18].

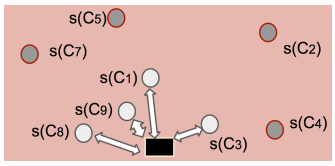
We take image features as an input to a multilayer perceptron and train a hidden embedding minimizing the cosine distance to the image's respective sentence caption. The image features are extracted from the penultimate hidden layer of the pre-trained ResNet-101 [6], that we mentioned in section 3.1. The features are mapped into a 256 dimension space using a affine transform with a sigmoid activation. The whole procedure of training image embeddings to map to the sentence hyperspace is shown in fig. 4b. Again, the loss for the MLP is the cosine distance between the image and the SIF sentence embeddings. Specifically, the loss is described by the following equation:

$$cos\_similarity(I, W) = \frac{p(I).s(W)}{||p(I)||_2 ||s(W)||_2} \qquad cos\_distance(I, W) = 1 - cos\_similarity(I, W)$$

Where $p(I)$ and $s(W)$ are the vectors which represent the image embedding and the sentence embedding respectively. We now have a rich hyperspace of sentence embeddings and a way to map



(a) Top-1 nearest neighbor



(b) Consensus caption, Top-4 nearest neighbors

Figure 5: Caption retrieval strategies, shown in a common 2-dimensional embedding space.

images to the sentence space such that they are as close to their respective captions as possible. To retrieve a caption for each test image, we map it to the 300 dimensional sentence space. We retrieve either the closest caption to that image embedding (fig. 5a) or choose the consensus of k closest captions to that image (fig. 5b), using equation 2 with the cosine similarity as sim.

4 Experimental methodology

4.1 Att-Net

MS COCO [25] has been used widely in image captioning tasks, which makes comparisons with existing baselines easier. Visual Genome [11] was chosen for its availability of information about objects embedded in an image and the object's specific attributes. We believe that this dataset has the ability to train our Att-Net better, given the bounding boxes with object-attribute labels. As mentioned in section 1, our research problem is that of generating image captions that are rich in objects and attributes. In order to test whether this goal is achieved, we compare the model with our implementation of Show and Tell [22]. Show and Tell is simply the caption generator described in section 3.1 and on fig. 3 with no objects/attributes input (the LSTM starts with input x^{-1} instead of x^{-2}).

We also run comparison experiments with *Boosting image captioning with attributes* [27]. The paper describes two components: an attribute detector [5], and the LSTM. We have run the pre-trained attribute detector on the MS COCO validation data and implemented the LSTM-A_3 model. Since our proposed model Att-Net replaces the attribute detector, we find that it is a valid comparison to make, in order to test our novelty. Due to limited computational power, we conducted only a few experiments to decide the hyper-parameters of Att-Net. Experiments on different learning rates, weight decay and optimizers were carried out. We found that the the loss function converged the best using Adam optimizer [10] with a learning rate of 0.0004 (decreased by a factor of 0.8 every 3 epochs). Pre-processing of the data involved cropping objects from each image, resizing the extracted images and mapping every object to its many attributes. Ground truth labels were of the form [object,[attribute$_1$, ..., attribute$_L$]] for each bounding box. To reduce the complexity of the model, we only consider the 100 most frequent objects and the 100 most frequent attributes. Att-Net was trained on images from the Visual Genome dataset [11]. For the validation set, as well as for test set, 5000 images common to MSCOCO and Visual Genome were selected at random. The rest of the available images in Visual Genome and MSCOCO were used for training the caption generator (97,000).

4.2 SIF - Retrieval

Our research question is of validating the SIF sentence embedding as an efficient language model with image caption retrieval. In our experiments to answer this question, we use the MS COCO datasetas our image captioning corpus. We use the default breakdown on the MS COCO website for the train and validation sets of 88,000, and 40,000 images respectively. We take a chunk of 5,000 images from the validation set to the considered as our test images. On the captions that come from the train corpus, we train the SIF method and generate the rich sentence hyperspace from which the test image will retrieve the captions. To train the multilayer perceptron to get image embedding from the image features and sentence embeddings, we use a single hidden layer of size 300, trained with the adam optimizer for 12 epochs with a learning rate of 0.007(optimal as seen through several manual experiments), on the cosine loss between sentence and image embedding. Using stochastic gradient descent optimizer with momentum of 0.9, we achieve similar results in 28 epochs. Hence,

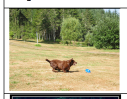
Image	Ground Truth Captions	Generated	Analysis
	- a dog is running in the field behind a Frisbee - a playful dog chasing down a frisbee that was brown - a brown dog running towards a frisbee in the yard. - a dog chasing a frisbee in a field. - this large brown dog is chasing a frisbee in this big field.	A dog in grass a with frisbee the	The dog, grass, and frisbee are correctly identified. The dog's location in grass and possessing a frisbee is also correctly identified. Grammar could be better without the extra 'a' and misplaced 'the'.
	- a young boy prepares to fly a kite - a child playing in a field with a frisbee. - a boy plays with a kite in a yard. - a child in a field with a kite - a young boy trying to get a kite to fly	A boy flying a kite a in field	The boy, kite, and field are correctly identified. The action of flying the kite is correctly identified. Grammar is slightly off with extra 'a'.
	- a young boy is eating a snack outside. - a little boy putting food in his mouth. - a little boy eating a bread ball outside. - a small boy putting small pieces of food in his mouth - a young blonde haired boy eating food while standing on a grass covered field.	A boy a eating donut a of.	The boy and donut are correctly identified. The action of the eating the donut is correctly identified. Donut is a reasonable assumption here in this model, given the object in the boy's mouth is a bread ball. Grammar could be better without the extra 'a's.
	- a shaggy dog sticks its head out of a moving car. - photograph of an outdoor arena that looks neat. - a dog that is sticking its head out the window. - a brown dog hanging its head out of a car door window. - a dog riding in a car looking out the window	A brown bear a in of and car	The color of the animal and location of the animal is properly identified. The animal is a dog, not a bear; however, this picture mainly displays its head. Because of the color and texture of the head, it can appear to resemble a bear. Grammar could be better without the extra 'a', 'of', 'and'.

Figure 6: Captions Generated from Att-Net and Language model collectively trained

	Show and Tell	Boosting with Image Captioning	Att-Net
CIDEr	0.855	1.002	0.000
Bleu_4	0.277	0.326	0.000
Bleu_3	0.279	0.430	0.000
Bleu_2	0.404	0.567	0.001
Bleu_1	0.579	0.734	0.004
ROUGE_L	0.396	0.540	0.151
METEOR	0.232	0.254	-

Figure 7: Metrics comparison of Att-Net with Baselines

even though adam converged faster, SGD makes it easier to tune to sensitive changes in the caption retrieval. An image feature is trained five times for its 300 dimension embedding because MS COCO gives five captions per image.

5 Results and Discussion

5.1 Att-Net

The metric results of our proposed Att-Net model are summarized in figure 7. Our metric results are poor, since most generated captions are grammatically incorrect. We compared our metrics with that of *Show and Tell* [22] and *Boosting Image Captioning with attributes* [27]. It can be seen that the caption generator does not produce well structured sentences . A possible reason for this could be that it overfits while jointly training it with Att-Net at the 4th epoch. We aim to look more into solving this issue in the future. As we can see in figure 6, our Att-Net does not predict captions that are attributes-rich. Although the captions don't capture good grammar, most objects are captured correctly. According to our results, we see that object-attribute pairings do not help in re-iterating attributes. This could be because the loss function is biased towards the larger loss value generated in the output layer of Att-Net. So, instead of adding the loss functions, with equal weightage we could possibly normalize the loss function. This could help because Softmax and sigmoid outputs are in different scales. Another reason for the failure of this model could be that we averaged the object-attribute features before feeding them into the LSTM.

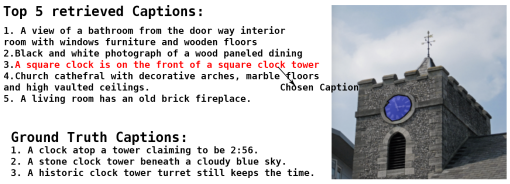
We implemented the GAN-based image captioning model we mentioned in section 3.2 but it took 3 days to run a single epoch, hence it didn't converge. Also, we did not have enough computing power to run experiments on GANs. Since we did not achieve suitable results for Att-Net it follows that Att-GAN would not have produced good results, had it converged. The complex architecture of Att-Net and the Language model is most likely the reason for the high computation time required.

5.2 SIF based Caption Retrieval

In table 8, we have shown some of the captions that were generated by our SIF based caption retriever. This method works remarkably well for object categories that are well represented in the dataset such as vehicles, horses, giraffes. However, it failed sometimes when describing the specific action or significance of the identified objects at hand. One important point to address was choosing the right caption out of the top K captions that were closest to the test image in the hyperspace. Figure

Image	Ground Truth Captions	Generated
	- a train engine carrying many carts down a track. - a train coming down the tracks in the city. - a passenger train glides around a curve on the outer track. - a train coming to a stop on the tracks out side. - the very long train is travelling down the tracks.	A commuter train traveling down some train tracks.
	- two men are on horses that are reared back. - two horses and their riders on some grass - two costumed horse riders make their horses stand on their hind legs. - a couple of men riding horses on top of a green field. - a pair of horses performing tricks in a field.	A couple men riding on horses next to a bunch of cattle.
	- a city street has an intersection sign on a pole. - street signs near tall buildings on the corner of greenwich st and vesey st. - a street sign on a pole on a street. - a greenwich street and vesey street sign are hanging on a pole. - two green and white street signs with building in background.	Road signage displayed on post at roadway intersection with pedestrian overpass.
	- two zebras seem to be embracing in the wild. - a couple of giraffe standing next to some trees. - two giraffes hang out near trees and nuzzle up to each other. - a couple of giraffe snuggling each other in a forest. - the two giraffes appear to be hugging each other.	Several giraffes and antelopes graze in a savannah grasslands setting.
	- a close up of a motorcycle parked on the sidewalk next to a door. - a motorcycle parked on the side of a building with mural - rusted red motorcycle parked beside a red building. - an old motorcycle parked outside a rustic building. - an old motorcycle rests on its kickstand by a door, in front of a wall with a mural of a tree.	A route 66 harley davidson motorcycle bike shop.

Figure 8: Captions Generated from the SIF based caption retrieval model



(a) Importance of choosing the right caption

Strategy	Bleu_4	Bleu_3	Bleu_2	Bleu_1	Rouge_L	METEOR	CIDEr
Top 15	**0.079**	**0.135**	0.238	0.434	0.321	0.129	0.273
Top 10	0.069	0.124	0.235	0.4330	0.322	0.130	0.275
Top 5	0.073	**0.135**	**0.246**	**0.443**	**0.327**	**0.136**	**0.281**
Top 3	0.066	0.122	0.232	0.433	0.325	0.130	0.253
Top 1	0.065	0.120	0.226	0.423	0.312	0.129	0.260
Show and Tell	0.191	0.297	0.404	0.579	0.396	0.195	0.600

(b) Metrics score results for different K values

Figure 9: Results for the SIF caption retrieval method

9a shows the importance of this. As shown, the third captions goes best with the test image even though it is not the top ranked retrieved caption (the first one is). So, instead of just the first caption we are using the consensus caption out of the top K retrieved captions. To optimize the value of K, we computed common metric scores, e.g Bleu, Meteor and cider for different values of K and chose the K value which gave the highest score as shown in table 9b. In the same table, we also show our results compared to the baseline show and tell mentioned previously.

6 Conclusion and Future Work

In this paper we proposed two different approaches to image captioning, one was to produce attributes-rich captions while the second aimed at reducing complexity by foregoing LSTM-based architectures. The first approach gave us insight into a joint image-text embedding space. We can delve more into this complex space by exploring different loss functions and architectures in the future. The second approach showed the efficacy of a retrieval based captioning method that got its captions' embeddings not by training with an LSTM but with the simple SIF method by [1]. Retrieval based methods realistically will have a ceiling to their performing given the size of the training space. We can use the SIF sentence embeddings in a join loss function even for generator based caption method and thus training captions with another parameter with semantic meaning.

References

[1] S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. 2016.

[2] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[3] B. Dai, D. Lin, R. Urtasun, and S. Fidler. Towards diverse and natural image descriptions via a conditional gan. *arXiv preprint arXiv:1703.06029*, 2017.

[4] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

[5] 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.

[6] 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.

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

[8] 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.

[9] R. Jozefowicz, W. Zaremba, and I. Sutskever. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2342–2350, 2015.

[10] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[12] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[13] R. Lebret, P. O. Pinheiro, and R. Collobert. Phrase-based image captioning. *arXiv preprint arXiv:1502.03671*, 2015.

[14] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and caption regions. *CoRR*, abs/1707.09700, 2017.

[15] 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*, pages 747–756. Association for Computational Linguistics, 2012.

[16] A. Oliva and A. Torralba. Torralba, a.: Building the gist of a scene: The role of global image features in recognition. progress in brain research 155, 23-36. 155:23–36, 02 2006.

[17] 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.

[18] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[19] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] 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.

[22] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[23] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. *arXiv preprint arXiv:1704.06972*, 2017.

[24] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. *CoRR*, abs/1704.06972, 2017.

[25] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.

[26] 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*, pages 444–454. Association for Computational Linguistics, 2011.

[27] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. *CoRR*, abs/1611.01646, 2016.

[28] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.