

# Final Report

Captioning images with diverse objects (CVPR 2017)

MD SARWAR AHMED

Electronics and Communication Engineering, NIT Silchar

Date:01/08/2021

---

## ABSTRACT

*Image captioning is a difficult endeavour that combines computer vision and natural language processing. To achieve the goal of automatically describing an object, a variety of ways have been presented.*

Image Captioning refers to the process of generating textual description from an image – based on the objects and actions in the image. This process has many potential applications in our day to day real life. A noteworthy one would be to save captions of images so we can retrieve them easily at later stages just on the basis of this description.

But, generally the models based on image-caption generation through the recurrent neural networks (RNN) and (later) long-short term memory (LSTM) are more predominant. However, the capacity of the recent captioning models to scale and represent concepts not seen in paired image-text corpora is limited though.

## INTRODUCTION

The **Novel Object Captioner** (NOC) is a deep visual semantic captioning model that can explain a large number of object categories that are not even represented in existing paired picture caption datasets.

We are using external sources in our model, such as labelled images extracted from object recognition datasets and semantic knowledge extracted from unannotated text. We propose minimising a combined goal that can learn from these many data sources while also leveraging from the distributional semantic embeddings (Glove embeddings), allowing the model to generalise and describe novel things outside conventional image-caption datasets.

We show that our model uses semantic information to generate captions for hundreds of object categories in the ImageNet object recognition dataset that aren't seen in MSCOCO image-caption training data, as well as many that are only seen once in a while. Both automatic and human evaluations reveal that our model exceeds previous work in terms of being able to deduce and efficiently.

## MOTIVATION

A quick glance into the pre existing image captioning models is sufficient for us to understand and also, describe what is happening in the big picture, i.e. the process of automatically generating any kind of textual description from an artificial system is the task of image captioning.

Even though this image captioning task is straightforward – the generated output is expected to describe in a single sentence what is shown in the image – the objects present, their properties, the actions being performed and the interaction between the objects, etc. But even to replicate this behaviour in an artificial system is a huge task, as with any other image processing problem. Hence, complex and more advanced techniques such as Deep Learning (OpenCV and NLP techniques) are used to solve the task.

### Methodology to Solve the Task:

The task of image captioning can be divided into two modules logically – one is a type of visual (or lexical) classifier or commonly, an image based model – which extracts the features and nuances out of our image, and the other is a language based model – which translates the features and objects given by our image based model to a natural sentence.

For our image based model (viz encoder) –  
We usually rely on a **Convolutional Neural Network** model.

Usually, a pretrained CNN extracts the features from our input image.

The feature vector is linearly transformed to have the same dimension as the input dimension of the RNN/LSTM network. This network is trained as a language model on our feature vector.  
For training our LSTM model, we predefine our label and target text.

For example, if the caption is “A man and a girl sit on the ground and eat.”, our label and target would be as follows –

Label – [ <start>, A, man, and, a, girl, sit, on, the, ground, and, eat, . ]

Target – [ A, man, and, a, girl, sit, on, the, ground, and, eat, ., <end> ]

This is done so that our model understands the start and end of our labelled sequence.

And for our language based model (viz decoder) – we rely on an **LSTM Network**.

Some examples of possible textual descriptions of a given image of a girl with her father:

1. *A man and a girl sit on the ground and eat .*
2. *A man and a little girl are sitting on a sidewalk near a blue bag eating .*
3. *A man wearing a black shirt and a little girl wearing an orange dress share a treat .*

The model which we saw above was just the tip of the iceberg as it has just preliminarily produced some descriptions relating to a provided image.

There has been a lot of research done and still going on this topic of generating human-like captions.

Currently, the state-of-the-art model in image captioning is **Microsoft's CaptionBot**<sup>1</sup>.

A few ideas which we can further use to build a better image captioning model would be as:

**Adding in more data** – Of course, this is the usual tendency of any Deep Learning model that the more data we provide to our model, the better it will perform.

**Using Attention models** – Sequence to Sequence modelling with Attention using attention models (soft or hard attention depending on circumstances), help us in fine tuning our model performance.

**Moving on to bigger and better techniques** – There are a few techniques which researchers have been investigating – such as using reinforcement learning for building end-to-end deep learning systems, using deep compositional captioners, **novel object captioners**, novel attention models for visual sentinel, etc.

## Novel Object Captioner (NOC)

Our NOC version includes a language version that leverages distributional semantic embeddings skilled on unannotated textual content and integrates it with a visible popularity version.

We introduce auxiliary loss functions (targets) and together teach special additives on more than one recorded sources, to create a visible description version which then, concurrently learns an impartial item popularity version, also in addition to a language version.

We begin by first training a LSTM-primarily based totally on language version (LM) for the sole purpose of sentence generation. Our LM includes dense representations for phrases from distributional embeddings (GloVe) pre-skilled on outside textual content corpora.

Simultaneously, we additionally teach a modern visible popularity community to offer confidences over phrases withinside the vocabulary given an photo. This decomposes our version into discrete textual and visible pipelines which may be skilled completely in the use of unpaired textual content and unpaired photo records.

To generate descriptions conditioned on photo content, we integrate the predictions of our language and visible popularity networks through summing (element-wise) textual and visible confidences over the vocabulary of phrases.

During training, we introduce auxiliary photo-specific ( $L_{IM}$ ), and textual content-specific ( $L_{LM}$ ) targets together with the paired photo-caption ( $L_{CM}$ ) loss.

---

<sup>1</sup> (link : [www.captionbot.ai](http://www.captionbot.ai))

These loss functions, whilst skilled together, affect our version to now no longer most effectively produce affordable photo descriptions, however additionally expect visible principles in addition to generating some cohesive textual content (language modeling).

We shall first discuss briefly about the **auxiliary targets** and the **joint training**, (education of the model), after which we illustrate how we can leverage **embeddings** upskilled with outside textual content /annotations in order to compose descriptions of approximately more than one or more novel objects.

---

Our inspiration for acquainting helper goals is with the sole objective of figuring out how to portray pictures without losing the capacity to perceive more items. Commonly, picture subtitling models often consolidate a visual classifier pre-prepared/trained on a source space (ImageNet dataset) and afterward tuning it to the objective space (the picture subtitle/caption dataset).

In any case, we can say that a significant amount of data from the source dataset can be stifled if comparable data is absent when tweaking, driving the model or organization to neglect (over-compose loads) for objects not present in the objective space.

This is tricky in our situation in which the model depends on the source datasets, primarily MS-COCO dataset, to gain proficiency with a huge assortment of visual ideas not present in the objective dataset. Be that as it may, with pre-preparing just as the model keeps up with its capacity to perceive a more extensive assortment of items and is urged to portray objects which are absent in the objective dataset at test time.

For the simplicity of our work, we dynamic away the subtleties of the language and the visual models and first portray the joint preparation targets or destinations of the total model, for example, viz. the content explicit loss, the picture explicit loss, and the picture subtitle loss, or more precisely text-specific loss, the image-specific loss, and the image-caption loss respectively.

## 1. Picture explicit : **Image specific Loss**

Our visual acknowledgment model is a neural organization parameterized by ' $\theta_I$ ' and is prepared on object acknowledgment datasets. Dissimilar to regular visual acknowledgment models that are prepared with a solitary mark on a grouping task, for the assignment of picture inscribing a picture model that has high certainty over different visual ideas happening in a picture all the while would be ideal.

Thus, we decide to prepare our model utilizing different marks with a multi-modal loss. In the event that ' $I$ ' means a mark and ' $z_I$ ' signifies the double ground-truth an incentive for the name, then, at that point the target for the visual model is given by the cross-entropy Loss (LIM).

## 2. Caption explicit : **Text specific Loss**

LSTM Networks are used in our language model. The parameters of this network are denoted by ' $\theta_L$ ', while the activation of the network's final layer is denoted by ' $f_{LM}$ '. In a given sequence of words  $w_0, w_1, w_2, \dots, w_{t-1}$ , the language model is trained to predict the next word  $w_t$ . The softmax loss LLM, which is equivalent to the maximum-likelihood method, is used to optimise this.

### 3. Image-Caption explicit : **Image-Caption Loss**

The objective of the picture subtitling model is to create a sentence adapted on a picture (I).

NOC predicts the following word in an arrangement,  $w_t$ , moulded on recently produced words ( $w_0, \dots, w_{t-1}$ ) and a picture (I), by adding initiations from the profound language model, which works over past words, and the profound picture model, which works over a picture.

We indicate these last (added) enactments by  $f_{CM}$ .

Then, at that point, the likelihood of foreseeing the following word is given by,

$$P(w_t | w_0, \dots, w_{t-1}, I)$$

$$= S(w_t)(f_{CM}(w_0, \dots, w_{t-1}, I; \theta))$$

$$= S(w_t)\{(f_{LM}(w_0, \dots, w_{t-1}(L; \theta_L)) + f_{IM}(I; \theta_I))\}$$

Given sets of pictures and descriptions, the caption model upgrades the boundaries of the basic language model ( $\theta_L$ ) and picture model ( $\theta_I$ ) by limiting the subtitle model loss  $L_{CM} : L_{CM}(w_0, \dots, w_{t-1}, I; \theta_L, \theta_I)$

$$= -[\text{Summation}(t) \{\log(S(w_t)(f_{CM}(w_0, \dots, w_{t-1}, I; \theta_L, \theta_I)))\}]$$

---

### 4. Training explicit : **Joint Training Loss**

While numerous past approaches have been effective on picture captioning/ subtitling by pre-preparing the picture and language models and tuning the subtitle model alone (viz. CVPR 2015 LRCN; CVPR 2016 DCC), it is quite insufficient to produce descriptions of objects just found in outer information sources.

To cure this, we propose to prepare the picture/ visual model, language model, and caption model all the while on various information sources. The NOC model's last target at the same time limits the three individual corresponding destinations:

$$L = L_{CM} + L_{IM} + L_{LM} \text{ (Element wise sum)}$$

By sharing the loads of the subtitle model's organization with the picture organization and the language organization, the model can be prepared all the while on autonomous picture just information, unannotated text information, just as matched picture inscription information.

Thus, this type of optimisational co-upgrading of various goals helps the model in perceiving classes outside of the matched picture sentence information.

Our **language model** comprises of the accompanying segments: a nonstop lower dimensional implanting space for words ( $W_{glove}$ ), a solitary repetitive (recurrent network) LSTM covered up layer, furthermore, two direct change layers where the subsequent layer ( $WT_{glove}$ ) maps the vectors to the size of the vocabulary.

At last, a final Softmax activation layer is utilized on the yield layer to create a standardized maximum likelihood appropriation. The cross-entropy loss which is identical to the greatest probability is utilized as the preparation or training objective.

Notwithstanding our joint target  $L$ , we likewise utilize semantic embeddings in our language model to assist with producing sentences while depicting novel articles. In particular, the underlying info installing space ( $W_{glove}$ ) is utilized to address the information (one-hot) words into semantically significant thick fixed-length vectors.

While the last change layer ( $WT_{glove}$ ) inverts the planning of a thick vector back to the full jargon with the assistance of a Softmax initiation work.

These distributional embeddings share the property that words that are semantically comparable have comparative vector portrayals. The instinctive justification utilizing these embeddings in the information and yield change layers is to help the language model treat words inconspicuous in the picture text corpus to (semantically) comparative words that have recently been seen to empower compositional sentence length, for example we can urge it to utilize new or even uncommon word in a sentence depiction dependent on the visual certainty.

Another differently foremost element of our version of image captioner model is the **image model IM**.

Here, we will attempt to hire the VGG-16 CNN (convolutional neural network) because the visible popularity in the image community can now be depicted by the confidences provided by the visual recognition network. We alter the very last layers of the community to include the multi-label loss  $L_{IM}$ , to expect the visible confidences over a couple of labels within the complete vocabulary. The rest of the classification network stays unchanged.

Finally, we take an detail clever sum (element wise sum) of the visible and language outputs, one can think about this because the language version generating a clean possibility distribution over words (primarily based totally on GloVe parameter sharing) after which the photo signal “selecting” amongst those primarily based totally at the visible proof whilst summed with the language version beliefs.

To generate descriptions conditioned on photo content, we integrate the predictions of our language and visible popularity networks through summing (detail-clever) textual and visible confidences over the vocabulary of words.

During training, we introduce auxiliary photo-specific ( $L_{IM}$ ), and text-specific ( $L_{LM}$ ) goals together with the paired photo-caption ( $L_{CM}$ ) loss. These loss functions, whilst educated jointly, have an impact on our version to now no longer best produce affordable photo descriptions, however additionally expect visible standards in addition to generate cohesive text (language modeling).

**Contextual Summary** of the auxiliary targets (before the joint training):

- a. Training deep visual classifier or an image model (IM) : unpaired ImageNET images
- b. Training deep language model (LM) : unpaired text annotations from Wikipedia, BNC, etc
- c. Combining IM and LM into a caption model (CM), which is actually trained on the paired image sentence MSCOCO dataset.

**Main takeaways** from NOC (as opposed to our previous DCC<sup>2</sup>, LRCN models for image captioning):

1. Concept of transfer learning is replaced by joint training of the IM, LM, CM
  2. Concept of “forgetting<sup>3</sup>” is being tried to counter with semantic embeddings
  3. Ability of model to caption MSCOCO images perfectly even when samples from IM are withheld.
- 

## Training and inference

- In our framework, the vision module is the VGG-16 network without fully connected layers, which is pre-trained on ImageNet. We use conv5\_3 feature map to compute attention features. The size of the hidden layer in the prediction module is 1024, and  $\lambda=1e-5$ . We apply Adam optimizer with mini-batch size 10 to train our model. For VGG-16, we set the learning rate to  $1e-5$ , and for other parameters, the learning rate starts from  $1e-3$  and decays every 50k steps. The stopping criteria is based on the validation loss. The weights are initialized by the truncated normal initializer with  $\text{stddev} = 0.01$ . The model was implemented in PyTorch.
- During training, as image-caption pairs are given alongwith the joint training of the IM and the LM, the convolution structures are applied in the normal way, and the loss function for each sentence is the cross-entropy.
- During inference, the caption is generated given the image using a feed-forward process. The caption is initialized as zero padding and a start-token  $\langle S \rangle$ , and is fed as the input sentence to the model to predict the probability of the next word. The predicted word is appended to the caption, and the process is repeated until the ending token  $\langle /S \rangle$ , is predicted, or the maximum length is reached.

## Experiments

In the proposed NOC model experiments are done on three datasets.

### Dataset and Experimental setup

**MSCOCO:** MSCOCO is the most popular dataset for image captioning, comprising 82,783 training and 40,504 validation images. Each image has 5 human annotated captions. Here split the images into 3 datasets, consisting of 5,000 validation and 5,000 testing images, and 113,287 training images. And vocabulary contains 10,000 words.

- We extract sentences from Gigaword, the British National Corpus (BNC), UkWaC, and **Wikipedia**. Stanford CoreNLP 3.4.2<sup>4</sup> was used to extract tokenizations. This dataset was used to train the LSTM language model. For the dense word representation in the network, we use GloVe [20] pre-trained on 6B tokens of external corpora including Gigaword and Wikipedia.
- To create our LM vocabulary we identified the 80,000 most frequent tokens from the combined external corpora. We refined this vocabulary further to a set of 72,700 words that also had **GloVe** embeddings.

---

<sup>2</sup> L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In CVPR, 2016. [7]

<sup>3</sup> [Catastrophic Forgetting in Neural Networks. Kirkpatrick et al. PNAS 2017]

Catastrophic forgetting occurs because **when many of the weights where "knowledge is stored" are changed, it is unlikely for prior knowledge to be kept intact.**

<sup>4</sup> C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford corenlp natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, 2014.

- **Image Caption data:** To empirically evaluate the ability of NOC to describe new objects we use the training and test set from. This dataset is created from **MSCOCO** by clustering the main 80 object categories using cosine distance on word2vec (of the object label) and selecting one object from each cluster to hold out from training. The training set holds out images and sentences of 8 objects (bottle, bus, couch, microwave, pizza, racket, suitcase, zebra), which constitute about 10% of the training image and caption pairs in the MSCOCO dataset. Our model is evaluated on how well it can generate descriptions about images containing the eight held-out objects.
- **COCO held out objects.** Table below compares the F1 score achieved by NOC to the previous best method, DCC on the 8 held-out COCO objects. NOC outperforms DCC (by 10% F1 on average) on all objects except “couch” and “microwave”. The higher F1 and METEOR demonstrate that NOC is able to correctly recognize many more instances of the unseen objects and also integrate the words into fluent descriptions.

**Metrics:** F1(Utility), Meteor (Fluency);

**Baselines:** LRCN<sup>5</sup>, DCC.

## Results

Model	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg. F1	Avg. METEOR
DCC	4.63	29.79	<b>45.87</b>	<b>28.09</b>	64.59	52.24	13.16	79.88	39.78	21.00
NOC (ours)	<b>17.78</b>	<b>68.79</b>	25.55	24.72	<b>69.33</b>	<b>55.31</b>	<b>39.86</b>	<b>89.02</b>	<b>48.79</b>	<b>21.32</b>

Table 1. MSCOCO Captioning: F1 scores (in %) of NOC (our model) and DCC [7] on held-out objects not seen jointly during image-caption training, along with the average F1 and METEOR scores of the generated captions across images containing these objects.

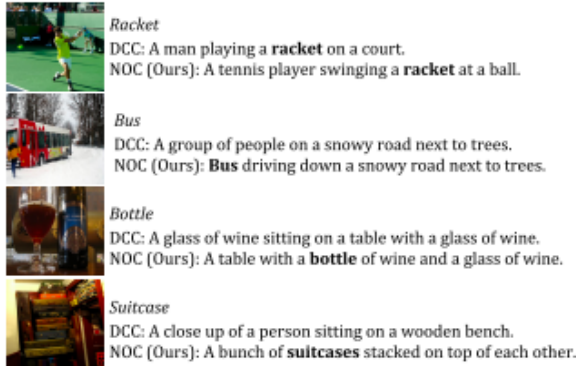


Figure 3. COCO Captioning: Examples comparing captions by NOC (ours) and DCC [7] on held out objects from MSCOCO.

	Image	Text	Model	METEOR	F1
1		Baseline (no transfer)	LRCN	19.33	0
			DCC	19.90	0
2	Image Net	Web Corpus	DCC	20.66	34.94
			NOC	17.56	36.50
3	COCO	Web Corpus	NOC	19.18	41.74
4	COCO	COCO	DCC	21.00	39.78
			NOC	<b>21.32</b>	<b>48.79</b>

Table 2. Comparison with different training data sources on 8 held-out COCO objects. Having in-domain data helps both the DCC [7] and our NOC model caption novel objects.

Although NOC and DCC use the same CNN, NOC is both able to describe more categories, and correctly integrate new words into descriptions more frequently.

DCC can fail either with respect to finding a suitable object that is both semantically and syntactically similar to the novel object, or with regard to their language model composing a sentence using the object name, but, in NOC the former never occurs (i.e. we don’t need to explicitly identify similar objects), reducing the overall sources of error.

<sup>5</sup> 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.



## Observations

When performing joint training and considering the overall optimization objective as the sum of the image specific loss, the text-specific loss and image-caption loss, we can define the objective more generally as:

$$L = L_{CM} + \alpha(L_{IM}) + \beta(L_{LM})$$















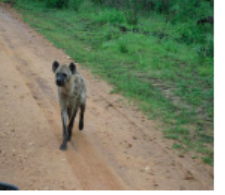
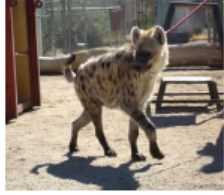
where  $\alpha$  and  $\beta$  are hyper-parameters which determine the weighting between different losses. In our experiments setting  $\alpha = 1$  and  $\beta = 1$  provided the best performance on the validation set. Other values of  $(\alpha, \beta) \in \{(1, 2), (2, 1)\}$  resulted in lower F1 and METEOR scores.

Metric	Model	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg.
F1	DCC	4.63	29.79	<b>45.87</b>	<b>28.09</b>	64.59	52.24	13.16	79.88	39.78
	NOC (ours)	<b>17.78</b>	<b>68.79</b>	25.55	24.72	<b>69.33</b>	<b>55.31</b>	<b>39.86</b>	<b>89.02</b>	<b>48.79</b>
METEOR	DCC	18.1	<b>21.6</b>	<b>23.1</b>	22.1	22.2	20.3	18.3	22.3	21.00
	NOC (ours)	<b>21.2</b>	20.4	21.4	21.5	21.8	<b>24.6</b>	18.0	21.8	<b>21.32</b>

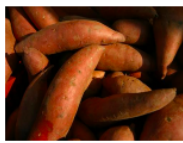
**Table 2.** MSCOCO Captioning: F1 and METEOR scores (in %) of NOC (our model) and DCC on the held-out objects not seen jointly during image-caption training, along with the average scores of the generated captions across images containing these objects.

Model	F1 (%)	METEOR (%)
DCC with word2vec	39.78	21.00
DCC with GloVe	38.04	20.26
NOC (ours, uses GloVe)	<b>48.79</b>	<b>21.32</b>

**Table 3.** DCC and NOC both using GloVe on MSCOCO dataset.

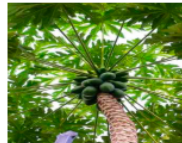
Earthenware			Caribou		
	A couple of <b>earthenware</b> sitting on top of a wooden table.	A <b>earthenware</b> sitting on a table with a plate of food.		A <b>caribou</b> that is standing in the grass.	A <b>caribou</b> that is laying in the grass.
Warship			Snowbird		
	A large <b>warship</b> is on the water.	A group of people standing around a large white <b>warship</b> .		A <b>snowbird</b> bird perched on a branch of a tree.	A <b>snowbird</b> bird sitting on a rock in the middle of a small tree.
Flounder			Lychee		
	A large <b>flounder</b> is resting on a rock	A man is holding a large <b>flounder</b> on a beach.		A bowl filled with lots of <b>lychee</b> and lychee.	A man holding a <b>lychee</b> and lychee tree.
Verandah			Hyena		
	A large building with a <b>verandah</b> and tropical plants in it.	A table with a <b>verandah</b> area and chairs.		A <b>hyena</b> dog walking across a dirt road.	A <b>hyena</b> standing on a dirt area next to a building.

## Vegetables



**NOC:** A bunch of **yam** are laying on a table.

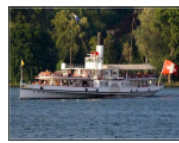
**DCC:** A person holding a knife and a knife.



**NOC:** A tree with a bunch of **papaya** hanging on it.

**DCC:** A **papaya** tree with a **papaya** tree.

## Water



**NOC:** A **steamship** boat is sailing in the water.

**DCC:** A boat is docked in the water.



**NOC:** A man standing on a boat holding a **snapper** in his hand.

**DCC:** A man standing on a boat with a man in the background.

## Clothing



**NOC:** A woman standing next to a woman holding a **boa**.

**DCC:** A man holding a pink umbrella in a pink **boa**.



**NOC:** A woman in **corset** posing for a picture.

**DCC:** A woman holding a red and white **corset** on a woman.

## Clothing



**NOC:** A man wearing a suit and tie with a **tweed** jacket.

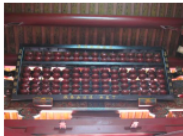
**DCC:** A man wearing a suit and tie in a suit.



**NOC:** A man wearing a hat and wearing **topcoat**.

**DCC:** A man wearing a suit and tie in a suit.

## Misc.



**NOC:** A **abacus** sitting on a wooden shelf with a **abacus**.

**DCC:** A **abacus** with a lot of different types of food.



**NOC:** A young child is holding a **drumstick** in a kitchen.

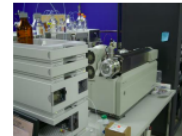
**DCC:** A little girl is **drumstick** with a toothbrush in the background.

## Misc.



**NOC:** A copier desk with a **copier** machine on top of it.

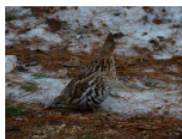
**DCC:** A laptop **copier** sitting on top of a table.



**NOC:** A **spectrometer** is sitting in a **spectrometer** room.

**DCC:** A white and white photo of a white and black photo of a white.

## Birds



**NOC:** A **grouse** is standing on a dirt ground.

**DCC:** A **grouse** is standing in the middle of a small pond.



**NOC:** A **shorebird** bird standing on a water pond.

**DCC:** A **shorebird** bird standing in the water near a body of water.

## Outdoors



**NOC:** A **volcano** view of a mountain with clouds in the background.

**DCC:** A man is sitting on a bench in the middle of a large **volcano**.



**NOC:** A **brownstone** building with a clock on the side of it.

**DCC:** A red and white **brownstone** in a city street.

## Water Animals



**NOC:** A **swordfish** sitting on a wooden bench in a city.

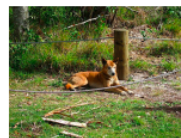
**DCC:** A man is sitting on a bench in the water.



**NOC:** A **crocodile** floats through the water edge of a body of water.

**DCC:** A large **crocodile** in a body of water.

## Animals



**NOC:** A **dingo** dog is laying in the grass.

**DCC:** A dog laying on a wooden bench next to a fence.



**NOC:** A small white and grey **tarantula** is sitting on a hill.

**DCC:** A black and white photo of a person on a white surface.

## Food



**NOC:** A plate of food with **hollandaise** sauce and vegetables.

**DCC:** A plate of food with a fork and a **hollandaise**.



**NOC:** A close up of a plate of food with **falafel**.

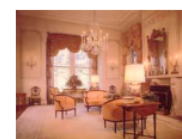
**DCC:** A plate of food with a fork and a **falafel**.

## Scenes



**NOC:** A woman standing in front of a **cabaret** with a large discotheque.

**DCC:** A woman standing in a room with a red and white background.



**NOC:** A **parlour** room with a table and chairs.

**DCC:** A large room with a large window and a table.

## Conclusions

While Table. 1 presents the F1 scores comparing the DCC model and our NOC model for each of the eight held-out objects in the test split, whereas, the Table. 2 supplements this by also providing the individual meteor scores for the sentences generated by the two models on these eight objects.

In the case of NOC, we sampled sentences and picked one with lowest log probability. Using beam search with a beam width of 1 produces sentences with METEOR score 20.69 and F1 of 50.51.

One aspect of difference between NOC and DCC is that NOC uses GloVe embeddings in it's language model whereas DCC uses word2vec embeddings to select similar objects for transfer. In order to make a fair comparison of DCC with NOC, it is also important to consider the setting where both models use the same word-embedding as seen in above presented Table 3.

---

One interesting future direction for us from here can be to create a model that can learn on new image-caption data after it has already been trained. This would be very similar to a predefined model<sup>6</sup>, in which to come to observe: after **an initial NOC model** has already been trained we might want to add more objects to the vocabulary, and train it on a few more image-caption pairs.

The key takeaway from here actually would be to improve the captioning model by re-training only on the new data instead of training on all the data from scratch which saves a lot of our time and machine power consumption in terms of either processing or yield generations.

---

<sup>6</sup> J. Mao, W. Xu, 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 ICCV, 2015