Project Report:
Image captioning with region attention

Alexandre Zouaoui
alexandre.zouaoui@telecom-paristech.fr

Abstract

This report reviews the Neural Baby Talk [6] (NBT) approach to image captioning. We investigate the proposed model and evaluate it on the Flickr30k dataset both quantitatively and qualitatively. We highlight the strengths and weaknesses of NBT. Lastly, we present NBT results on a new dataset consisting in unusual visual relations (UnRel) [8].

1. Introduction

As a field, Computer Vision and Object Recognition has seen astonishing progress at a steady pace ever since Convolutional Neural Networks (CNN) [4] made a comeback in the ImageNet competition [10] 2012. One could argue that image classification - the task of labelling an image based on the entity present in it - has essentially been "solved". One could also claim that object detection - the task of finding objects location and their label in an image - is getting convincing results. At this point, it is natural to ask such models to caption images. In other words, to describe an image based on its content. This is something a 3 years-old child can do.

Simply put, captioning an image requires a model to process an input image so as to output a sentence - i.e. a sequence of words. Fortunately, together with advances in Computer Vision and Object Recognition, came improvements in Natural Language Processing (NLP) that are also due to Neural Networks - Recurrent Neural Networks (RNN) in that case. Naturally, recent approaches have sought to combine a Vision Model and a Language Model into a single end-to-end differentiable model that would take an input image, embed it using a CNN and use this feature representation to output a sequence of words using a RNN.

Such models have seen promising results but are held back by a new set of challenges that were non-existant in classification and detection tasks.

1. RNN are prone to overfitting.
2. Evaluating sequences is a non-trivial problem.
3. Annotations are expensive and can be ambiguous.

For these reasons above there is a clear need to develop methods that are more robust to novel situations and can adapt to out-of-training-domain images, as well as models that are flexible enough not to rely on fully-supervised settings where the training captions are tightly linked with elements in the image. Additionally, how to best evaluate image captioning models is still an active research area in NLP.

One major flaw of the recent methods that combine CNN and RNN lies in the fact that they are acting as a black box. One reasonable property one should ask for is the ability for the network to actually map words in the caption that refer to entities to regions in the image. This mechanism is called *attention*. It would seem unthinkable to a human to describe an image without being able to point to objects contained in the image and that are to be part of the generated sentence. This approach has the benefit of leveraging the existing object detection models that have been shown to work decently well. In essence, NBT [6] imitates *Baby Talk* as it generates a sequence whose tokens can either be visual - i.e. correspond to a region in the image - or textual - i.e. do not correspond to a region in the image but is part of the learned vocabulary. This flexible model exhibits desirable properties that we discuss in the next section.

2. Presentation of the methods

In this section we present the main contributions of the NBT paper [6] on the modelling part.

We mentioned above how NBT could be seen as a neural slot-filling model - the slots being filled with visual words that correspond to objects grounded in the image.

2.1. Mathematical setup

More formally, given an image I, the goal is to generate a sequence of words $y = \{y_1, \ldots, y_T\}$. To be able to ground entities in the image, a set of N image regions $r_I = \{r_1, \ldots, r_N\}$ are extracted. The model parameters θ

are derived by maximizing the log-likelihood of the correct caption:

$$\theta^* = \arg\max_{\theta} \sum_{(I,y)} \log p(y|I;\theta) \quad (1)$$

The specific image region at timestep t is introduced as a latent variable r_t and the chain rule on joint probabilities is used, such that $p(y|I;\theta)$ in 1 can be written:

$$\prod_{t=1}^{T} p(y_t|r_t, y_{1,\ldots,t-1}; I; \theta) p(r_t|y_{1,\ldots,t-1}; I; \theta) \quad (2)$$

Under this framework, a word at timestep t is either a visual word, y^{vis} - meaning it is tied to a specific image region in r_I - or a textual word, y^{txt} that belongs to the vocabulary of the language model but is not grounded in the image. For modelling purposes, textual words y^{txt} are tied to a default region $\tilde{r}$.

The generated sentence template is filled with visual words y^{vis} that correspond to the output of an object detector. To refine the coarse label obtained by a typical object detection network, the authors seek to determine the plurality of the word and its fine-grained class when available.

We can now derive the cross-entropy objective to be minimized during training:

$$\begin{aligned} L(\theta) = -\sum_{t=1}^{T} \log \Big(& p(y_t^*|\tilde{r}, y_{1,\ldots,t-1}^*; I; \theta) \mathbb{1}_{\{y_t^* = y^{txt}\}} \\ & + p(b_t^*, s_t^*|r_t, y_{1,\ldots,t-1}^*; I; \theta) \\ & \times (\frac{1}{m} \sum_{i=1}^{m} p(r_t^i|y_{1,\ldots,t-1}^*; I; \theta)) \mathbb{1}_{\{y_t^* = y^{vis}\}} \Big) \end{aligned} \quad (3)$$

The first term in 3 indicates the textual word y^{txt} probability. The second term corresponds to the averaged target region r_t probability multiplied by the caption refinement probability $p(b_t^*, s_t^*|r_t, y_{1,\ldots,t-1}^*; I; \theta)$ where b_t^* and s_t^* are the target ground truth plurality and fine-grained class described above. Note that this setting allows weakly supervised learning, in situations where the caption words are not directly linked to bounding boxes in the image.

2.2. Architecture

The vision model consists in Faster RCNN and ResNet-101 to retrieve and proccess the region proposals. The language model consists in an attention model with two LSTM layers. The refinement model consists in two single layer MLPs using ReLU activations that are respectively used for fine-grained classification and plurality.

The authors provide an implementation on GitHub: (`https://github.com/jiasenlu/NeuralBabyTalk`)

3. Results

In this section we present our replication results on the Flickr30k dataset [9].

The Flickr30k dataset contains 31 783 images, each described by 5 captions, in which bounding boxes are tied to words in the caption. There are 2 567 unique words and 480 categories. One of the reasons for choosing Flickr30k to replicate the paper results over COCO [2] is that I noticed a nicer overlap in the base categories of UnRel [8] and Flickr30k. There are indeed only 9 missing object categories from UnRel out of 41 when using Flickr30k categories (compared to 16 when using COCO). I figured it would be easier to use Flickr30k vocabulary to later produce captions on the UnRel dataset.

3.1. Quantitative results

As underlined in the introduction 1, evaluation in image captioning is tricky. There are a variety of metrics inherited from NLP that all shed some light on the performance of the model but are not as clear-cut as an accuracy metric for image classification. Fortunately they seem to be correlated.

Score metrics in NLP convey different viewpoints. BLEU-n score [7], from the Machine Translation community, looks for the proportion of matched n-grams between the target and the generated output. It is a precision-based metric, contrarily to ROUGE [5], which is a recall-based metric from the summarization community. Another metric that better correlates with human judgement is METEOR [3], which uses sentence-level similarity scores and universal parameters on lemmatized words. CIDEr [11] is based on consensus but its authors have shown that having 5 captions might not be sufficient to measure accurately how a majority of humans would describe an image. Finally, SPICE [1] focuses on semantics by evaluating captions on scene graphs whose nodes are either objects, relations or attributes, using a $F1$-score.

It is always enlighting to compare the whole NBT model results to an oracle model where the vision model delivers the ground truth bounding boxes and their corresponding label in the image. This allows us to see how the language model performs on its own as well as assess how much of an improvement we could get if object detectors were working flawlessly.

We observe in 1 that improvement on object detectors can still be made. However the relative improvement suggests that the language model is still far from delivering meaningful sentences on average.

3.2. Qualitative results

At this point it is important to qualitatively analyze the results obtained on Flickr30k. Figures 2 and 3 provide insight on what the model is able to do. The gray boxes indicate that an object was detected by the vision model but

	BLEU-1 [7]	BLEU-4 [7]	METEOR [3]	ROUGE [5]	CIDEr [11]	SPICE [1]
NBT Authors	**69.0**	**27.1**	**21.7**	-	**57.5**	**15.6**
Replication	**69.0**	27.0	21.6	**48.3**	57.0	15.5
NBT Authors + *oracle*	**72.0**	**28.5**	23.1	-	64.8	19.6
Replication + *oracle*	71.8	28.4	**23.3**	**50.1**	**65.2**	**19.9**

Figure 1. Replication results on Flickr30k. Bottom 2 rows correspond to results obtained using an *oracle* vision model as described in 3

it was never used as a visual word by the language model when generating the caption. The takeaway may be that the model is flexible enough so that it can still output a relevant caption when no object has been detected (see column 2 in 2). On the other hand, it sometimes fails to incorporate an accurate detection into the caption (see column 3 in 2).

When it comes to comparing NBT with an oracle vision model, we notice in 3 that the model can indeed leverage the true objects to generate the caption. However, we observe that the gains are quite marginal, which is in line with the reported results in 1, since the language model might only use a fraction of the available detected objects. Conversely, the language model may output the correct entity although it was not detected by the vision model as seen above. This flexibility casts a shadow on the desirable property we mentioned in 1 that is to require the model to ground the captions objects in image regions as a human would do.

4. Implementation details

In this section we discuss the currently available NBT implementation. The NBT implementation pipeline is complex as it requires to combine the vision and language models.

4.1. NBT for custom images

The authors are aware their current implementation cannot be used on custom images (see Github issues # 10 and # 27). Doing so requires having access to a vocabulary as well as region proposals. Our idea was to re-use Flickr30k [9] vocabulary and apply it to caption UnRel [8] images given the ground truth proposals. More details about the reverse engineering and the code used can be found here: (`https://github.com/inzouzouwetrust/RECVIS_final_project`). We report the results on 115 annotated images in 4, 5 and 6.

4.2. Disclaimer

We mentioned during our presentation that the authors might be using a dubious split where the validation set was also the testing set. Upon closer look, it appears to be an uneducated guess. We were misguided by a series of printing statements on the data loading part that are the same regardless of whether we are training or evaluating the model. It appears that the validation data loader uses the **test** split by default. However that is not sufficient to claim that the authors used the test split as a validation split during their training. We are out of time to retrain from scratch and investigate deeper. That being said, if someone were to use the default values provided in the code to train from scratch, he would be using the **test** split as the validation split, which is prohibited in standard machine learning settings.

4.3. Replication

To replicate the results published by the authors, we used the pre-trained NBT model on Flickr30k [9]. We use a GPU instance configured on Amazon Web Services using the GitHub Student Pack credits. As the official Flickr30k [9] homepage is down, it required us to host the dataset on a AWS cloud storage bucket using a local copy downloaded last December.

5. Conclusion

In this review we have seen the clear advantages of NBT, as it leverages object detectors to populate a template sentence with entities found in the image. Doing so partially tackles the compositionality issues of RNNs used in the language generation model. Moreover, NBT is loose enough to work on weakly supervised settings.

However, this flexibility comes at the cost of accountability. We would expect the detected entities to be used somewhere in the output but it is not always the case. As a result, NBT is able to create sentence without any detection but may miss correct detections that can sometimes be generated by the language model regardless.

One idea would be to reverse the model so as it would produce captions using all the detected entities in the image. However that requires some modelling prowess.

Another approach consists in building on visual relations model from Peyre & al. [8] to include attributes of the detected objects together with the existing triplets. A naive way to do it would be to train fine-grained classifiers that could tell the difference between a *small brown* dog and a *big white* dog.

One last area of improvement for the vision model consists in replacing the bounding boxes with segmentation or polygons as the former arguably provides a *weaker* representation.

References

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

[2] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

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

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[5] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. page 10, 01 2004.

[6] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *CVPR*, 2018.

[7] K. Papineni, S. Roukos, T. Ward, and W.-j. Zhu. Bleu: a method for automatic evaluation of machine translation. 10 2002.

[8] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Weakly-supervised learning of visual relations. In *ICCV*, 2017.

[9] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[11] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

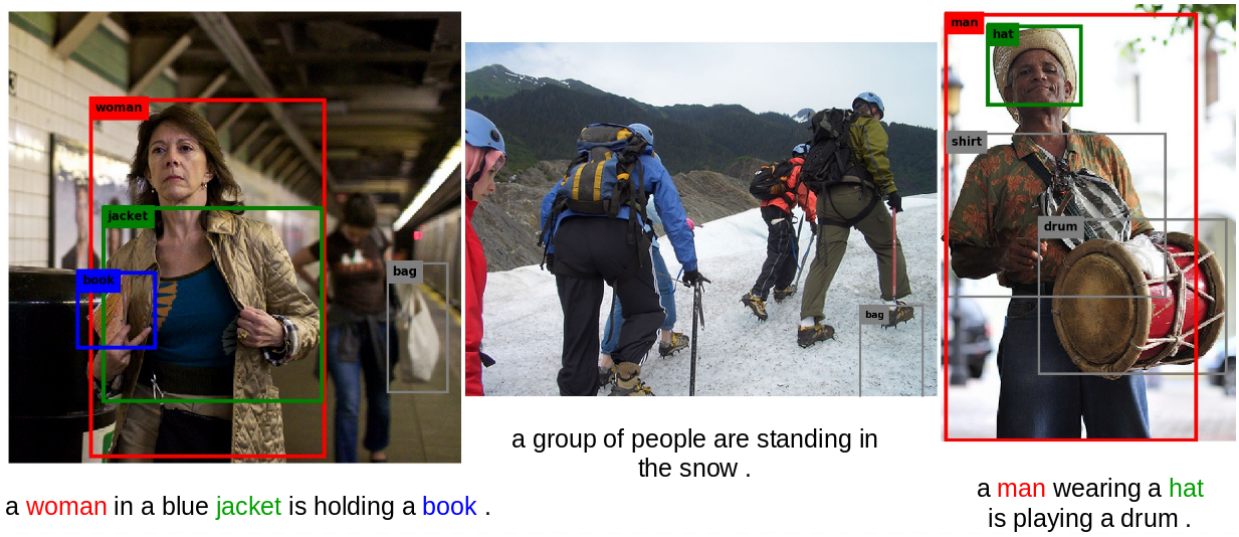

Figure 2. Captioning results on Flickr30k [9] testing split. Note that the *gray boxes* indicate that an object was detected by the vision model but it was never used as a visual word by the language model when generating the caption. **Left**: the sentence generated matches the ground truth except for the attribute of the jacket object. **Middle**: the caption is reasonable despite an erroneous detection that was not used in the end. **Right**: the caption is accurate despite the *drum* object being detected but not used by the language model. The language model generated it on its own, without grounding the *drum* in an image region.

Figure 3. Comparing NBT to an oracle version. Left side is the output from NBT. Right side is the output from the language model of NBT using the ground truth bounding boxes and their label (i.e. the *oracle*). **Left**: we observe how having access to the ground truth locations and labels can improve the generated caption despite a labelling error. Here the man is holding a *ski stick*, not a *ski*. However he indeed appears to be skiing, hence the oracle being correct, contrarily to NBT which detects a *ball*. **Right**: Here NBT generates a reasonable caption despite lacking detections from the vision model. On the other hand, the language model only uses one detection provided by the oracle. In addition, the oracle version hallucinates a *restaurant* which is not visible. This shows why even when the oracle is provided the output sentence might not be ideal.

	BLEU-1 [7]	BLEU-4 [7]	METEOR [3]	ROUGE [5]	CIDEr [11]	SPICE [1]
NBT + *oracle*	32.3	2.9	14.7	33.0	12.0	2.4

Figure 4. Results on 115 annotated images from the UnRel dataset [8]. Here the vision model delivers the ground truth bounding boxes and detection labels (i.e. *oracle*). Each image was described thrice: general description, attributes oriented and most salient detail. The massive difference between BLEU-1, ROUGE scores and the rest might be due to having the ground truth detections, hence having correct unigrams in the image but not the entire meaning.

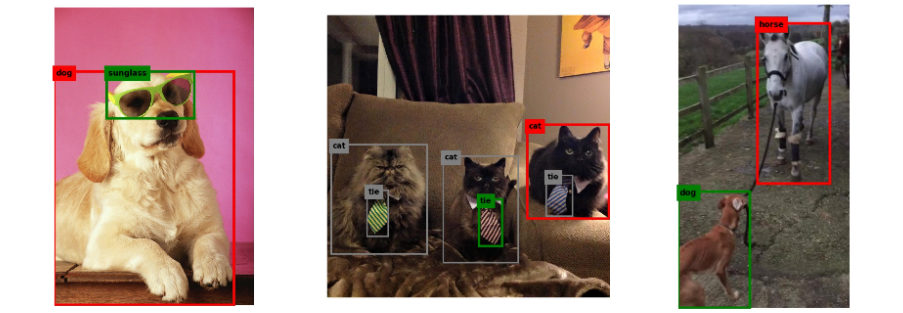

Figure 5. Captioning results on UnRel [8] with *oracle* detections using pre-trained NBT [6] model on Flickr30k [9]. Note that the *gray boxes* indicate that an object was detected by the vision model but it was never used as a visual word by the language model when generating the caption. **Left**: the generated caption is flawless except for the fact that the dog is not sitting on a couch but on the floor. This highlights the language model compositionality issues. **Middle**: Here two out of three cats are dismissed despite being available. Note that the cat and tie pair used by NBT are not actually related. **Right**: Here the captioning results does not capture what is actually pictured even though the entities are present in the sentence.

Figure 6. Captioning results on UnRel [8] with *oracle* detections using pre-trained NBT [6] model on Flickr30k [9]. **Left**: The elephant (here manually mapped to *animal*) is not used to generate the caption. **Middle**: Here the counting is wrong but the general idea is accurate. We notice how the added *plurality* model can be an issue since only one detection is used to generate the plural noun. **Right**: Again, only one person is described. The *blue* attribute relates to the undetected person. Comically the shoes are fused with the jean. All these pictures have made it clear that NBT modelling is not sufficient to tackle unsual relations as it still suffers from the language model overfitting on Flickr30k [9].