A Meta Learning Approach to Novel Image Captioning

Eli Pugh
Stanford University
epugh@stanford.edu

Abstract

Image captioning for images that contain novel objects not seen in training data is a difficult yet very valuable task. A recently released novel object captioning challenge (nocaps) [1] aims to evaluate different algorithms at this task. While approaches to this task aim to evaluate on totally unseen data, it's also useful to have models that can adapt well to distribution shifts after only a few unseen examples. To this end, I propose a Model-Agnostic Meta Learning (MAML) [4] approach using a popular UpDown image captioning model [3]. This is a significant challenge, as it's often difficult to use MAML for generative tasks, especially with very large and complex models such as UpDown. I show that it's possible to train such a model with MAML, though it falls significantly short of SOTA performance on this task. I also show that it's possible to take a traditionally trained model as an initialization for MAML and still learn to improve accuracy on out-of-domain images. I make all code available at:
https://github.com/elipugh/nocaps-meta.

1 Introduction

Image captioning is a very classic multi-modal task that combines elements from object recognition, segmentation, relational modelling, and natural language generation. Due to significant advancements over the past few years in computer vision [10] and natural language understanding [11], performance on public datasets such as COCO [7] and Flickr30k [8] have improved dramatically. Despite this, these models often struggle when they encounter images that contain objects unseen in training data.

The recently released nocaps benchmark [1] is designed to evaluate models on this out-of-distribution task and encourage development of image captioning models that generalize well to new domains. Previous attempts at novel image captioning use object detection data integrated with image captioning approaches to fill in the blanks for unknown objects, but relationships between objects often depend on the objects themselves, so this leaves much to be desired.

In this paper, I focus instead on a slightly different challenge. I use meta-learning to be able to leverage a few examples of image-caption pairs with novel images to better classify more new images that contain those novel objects. This is useful because it's often not so difficult to obtain a few examples of image-caption pairs when adjusting to a new domain.

I will first explore prior approaches to this task in section 2. I then go into the task in more detail, as well as dataset analysis in section 3. In section 4 I outline implementation and more training details. In section 5 I list some different tests and results. In section 6 I provide some analysis. Finally I conclude and give future work in section 7.

Preprint. Under review.

2 Prior Work

2.1 Image Captioning

A large number of prior works on image captioning have been done and focus on using large datasets to pair images with descriptions. For example, (Hu et al. 2020) utilizes a VIsual VOcabulary(VIVO) pre-training method to first map similar objects into vectors to pair image-level tags with corresponding image region features. This is both easier than human labeling as well as auto-tagging. Then, results are fine-tuned in a way that the words initially unordered in a tag are ordered to form a caption. This can even be done for novel objects which does not work for a standard nocaps benchmark. The model can be generalized to work for other similar images, however, more classes significantly improve the performance and the model does not perform well when limited training examples are used. The same is true for (Anderson et al. 2018). They combine both the bottom-up and top-down mechanisms to approach image captioning. While results are strong, it also needs to be trained on many examples to be performed accurately.

2.2 Meta Learning

Other works on meta learning have focused on implementing MAML to speed up the process of imaging. Rather than requiring a large dataset of images, meta learning can be used to feed just a few examples so that the network is trained to learn and caption an image quickly. For instance, (Li et al. 2019) uses few-shot learning to imply relationships from only a few images and then apply those to generate a description given an input image. In (Dong et al. 2018), These are similar to my model, but on smaller scales. In these cases, they manually write a caption except for the missing fill-in-the-blank and use the model to try and find the correct word for an object they have not yet seen whereas my model will generate the entire caption.

2.3 UpDown and CBS

In addition, the Bottom Up and Top Down Attention (UpDown) model has been the basis of many image captioning models. Both [3] and [1] UpDown as a reference point for image captioning. UpDown generates the words and then Constrained Beam Search (CBS) modifies UpDown by selecting a few possibilities in a way that makes a tree so that it can choose the best one [2].

3 Challenge

3.1 Data

The nocaps challenge uses training data compiled from a few sources. For training, teams may use image-caption pairs from the COCO dataset [7], as well as bounding boxes and image-level tags from the Open images dataset [6]. The test data is from Open Images as well, and these images contain nearly 400 objects that are not seen in COCO training data.

Open Images contains over 1.9 million images with over 600 object classes in total, and an average of 8.4 object instances per image. Over 400 of these object classes are unseen in the COCO dataset. These images are captioned using Amazon Mechanical Turk in a similar fashion to COCO to create a test set containing novel objects.

4 Methodology

While most previous works on this task focus on the "zero-shot" task, I choose to focus on a few-shot setting. This means that rather than only training a model once and then releasing on novel images, it's possible to leverage a few images that contain novel objects to adapt and perform better on new unseen images.

4.1 UpDown

As a baseline model, I use the Bottom Up and Top Down Attention (UpDown) model for image captioning [3]. This model uses Faster RCNN [9] for object detection and featurization of objects in

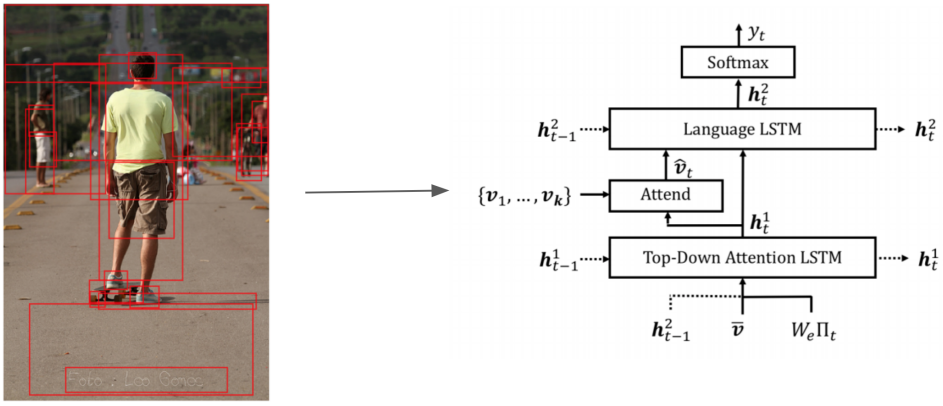

Figure 1: UpDown Image Captioning Model

different areas of the image. Language generation is done using an LSTM [5]. At each generation step, a soft attention is applied over the different features from the object detection and featurization stage, and this combination feeds visual information into the LSTM cells. This is shown below in Figure 1. Notice as well that attention is computed over detection regions that are not all the same size. This is a big addition that UpDown brings over previous attention-based visual feature methods, since images have objects of many sizes.

In addition, I use Constrained Beam Search (CBS), which is detailed further in [**?**]. Though I won't go further into detail, it is a beam-search to improve generation, but doesn't affect other parts of the model or training process.

4.2 MAML

In order to do this, I use a MAML training regime [4]. MAML learns a model that can be easily transferred to new tasks after only a few training examples from that task. This means learning optimal model parameters for adaption to new tasks. This is done by first adapting model parameters θ to new tasks L_i by calculating their losses. I then train on a few examples of a task and update θ to a new θ' that is fit for the task, and evaluate on validation examples of this new task. This is done as

$$\theta_i' \ = \ \theta - \alpha \nabla_\theta L_i(\theta)$$

where L_i is the loss for task i, θ_i' are the parameters learned for the task i, and α is the learning rate. MAML hopes to minimize L_i using θ_i' with respect to the original θ.

$$\min_\theta \sum_i L_i(\theta_i') \ = \ \min_\theta \sum_i L_i(\theta - \alpha \nabla_\theta L_i(\theta))$$

This gives the update rule:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_i L_i(\theta_i')$$

Again, β is a separate step size. Notice that this is optimizing over $theta$, and thus I am taking second order gradients. For this, I use a first-order approximation to speed up computation [4].

An illustration of MAML is given in Figure 2.

4.3 MAML for Image Captioning

MAML is often used for classification tasks, where each task is framed as an n-way classification task, and the n classes are chosen at random for each task example. Then finding θ_i' is done over support examples from these classes, and calculation of the loss is done over query examples from these classes.

It's fairly unusual to train generative models in a MAML fashion. For this, I define classes to be images that contain an object. This means that there are as many classes as there are object classes.

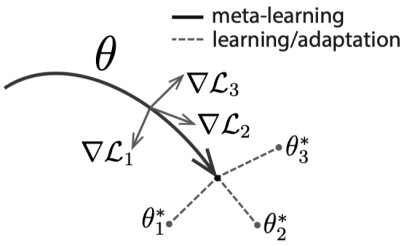

Figure 2: Model Agnostic Meta Learning Illustrated

Notice that this means many objects will belong to several classes. This shouldn't be a problem, since I still split the data in a way that novel objects in testing data are not seen before in training data. This allows the adaptions to still be useful, as the training images are not in the testing classes.

5 Experiments

For my experiments, I use the 5-shot setting for new images, meaning that there are 5 examples of images with each new object to leverage. In addition, I assume that inclusion of new objects will not happen one at a time, and often one will need to shift to a domain where several objects need to be learned by the model. To simulate this, I perform meta training over images with 10 different classes of objects.

I regret not having more time to properly evaluate this new approach, but instead can only show preliminary results, as evaluation data is stored on a server. This means that computation of metrics on evaluation data is not possible. Instead I simulate this by splitting the training data as described, and evaluating on a separate split. Because I'm unable to evaluate the common metrics, I instead only evaluate loss. While loss is scaled separately, and it usually is entirely not useful to only present loss, I do compare the per-image loss with a strong UpDown model [3] that is near the top of the nocaps leaderboard. I trained this model in the proposed fashion to reproduce the leaderboard results nearly exactly, then compare per-image loss to the meta learning model.

5.1 UpDown With Original MAML

For the first experiment, I trained the UpDown model with a MAML regime as described above. I show the loss in Figure 3 to demonstrate that it is indeed possible to train this model to improve and learn to caption images with this training setup. It's very difficult to train large generative models with MAML-style training, and thus I feel that this is significant, despite the lack of testing metrics to compare performance. In addition, I show that over time, the model learns to improve with adaption given a few training examples. This is shown in Figure 4

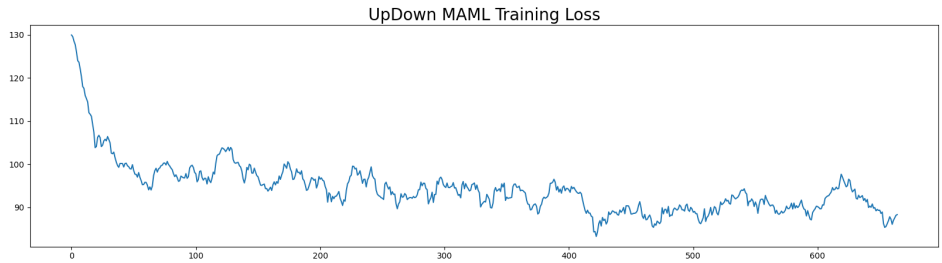

Figure 3: UpDown with original MAML

5.2 Pretrained UpDown With MAML

In the second experiment, I pretrained an UpDown model using the original method, and reproducing the metrics shown in [1]. This is shown in Figure 5. I then ran MAML with this initialization to

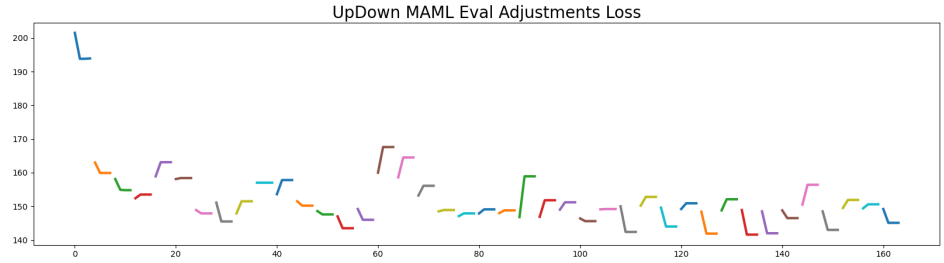

Figure 4: UpDown with original MAML Adaption

speed up training and see whether or not there was room to gain performance by adapting using a few examples. This showed very promising results, as the loss increased further, even with the strong initialization. This shows that there is significant performance to be gained from meta-learning style inference as well, even if a majority of the training is done in the typical supervised fashion.

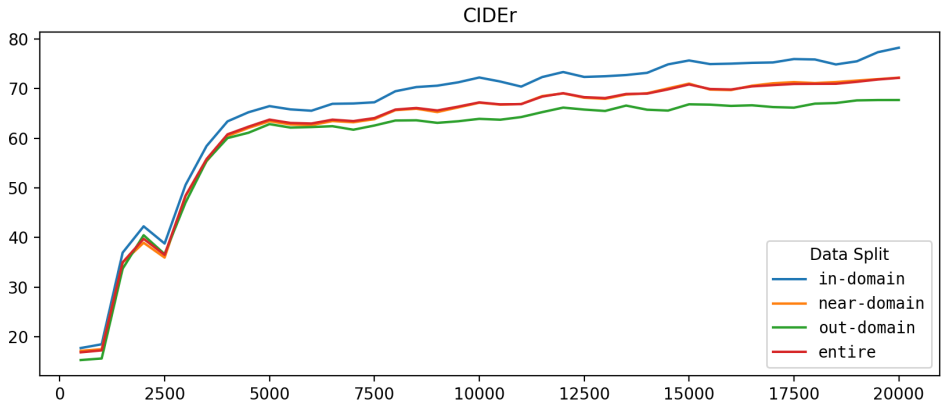

Figure 5: UpDown Classic Training CIDEr

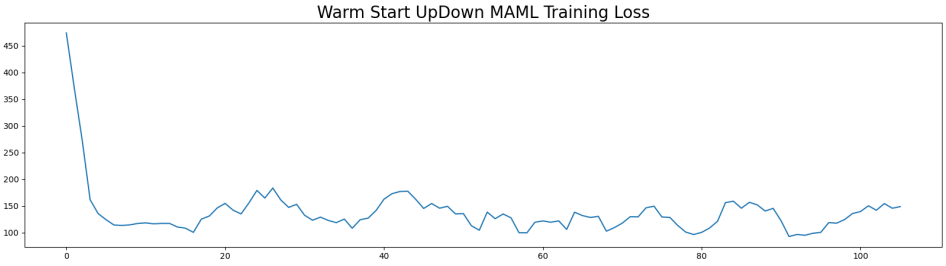

Figure 6: Warm UpDown With MAML

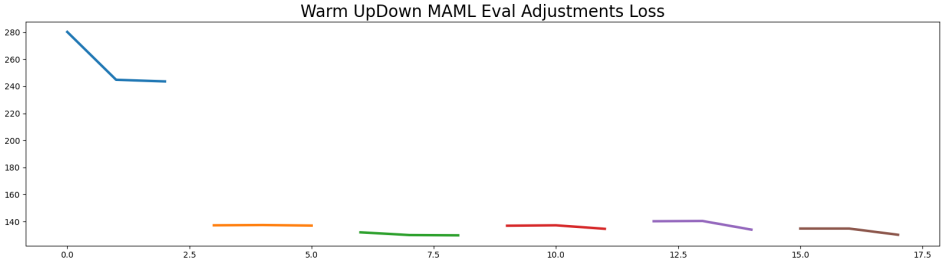

Figure 7: Warm UpDown With MAML Adaption

6 Discussion

While it's difficult to analyze performance of models only by loss rather than standard benchmarks, there are a few conclusions one can draw from these experiments.

Firstly, it's exciting that the large multi-stage UpDown model learns to improve and leverage examples through meta learning. This is often difficult to train, and any small improvement is showing progress.

In addition, it's very encouraging that one can take UpDown, an already performant model, and further improve the loss by using this MAML training regime. I was skeptical that there might not be much to gain, or that the pretrained UpDown would already be at a near-optimal point for adaption.

Despite this, there is still much more work to be done in qualitative analysis. It would be interesting to see how much this method improves recognition of novel objects. It would also be very important for this work to be tested and evaluated more thoroughly on standard benchmarks.

7 Conclusion

In this paper, I approached novel image captioning with a Model Agnostic Meta Learning training style. I compare this to the UpDown image captioning benchmark, and show that the pretrained UpDown model can also benefit from additional MAML-style training. This can be used to boost performance on out-of-domain examples in future applications when a small number of novel object images are present for adaption.

References

[1] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, 2019.

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Guided open vocabulary image captioning with constrained beam search, 2017.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[4] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

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

[6] A. Kuznetsova, H. Rom, N. Alldrin, J. R. R. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982, 2018. URL `http://arxiv.org/abs/1811.00982`.

[7] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.

[8] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015. doi: 10.1109/ICCV.2015.303.

[9] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[10] R. K. Sinha, R. Pandey, and R. Pattnaik. Deep learning for computer vision tasks: A review, 2018.

[11] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. Efficient transformers: A survey, 2020.