Scientific Report for STSM at the University of Trento

Alexander Kuhnle
University of Cambridge (United Kingdom)
aok25@cam.ac.uk

13th March 2017

1 Purpose of the STSM

During the ESSLLI 2016 summer school in August last year, I met Raffaella Bernardi and her PhD student Sandro Pezzelle to talk about their recent paper on training multimodal neural networks to quantify (Sorodoc et al., 2016). My own PhD research focuses on evaluating multimodal deep learning systems with respect to linguistic understanding abilities, one being the ability to understand statements involving quantifiers. This is why we decided that it would be helpful for both sides to combine our insights and collaborate on a joint project to investigate multimodal quantifier learning. Besides Sandro, Raffaella and me, Aurelie Herbelot and Ionut Sorodoc (both currently in Trento) are also involved in this project.

Our plan for my visit at the University of Trento was to start off the project, to analyze available data, discuss the details of the task and experimental setup, and to set out further work aiming for a joint publication. In conversations before, we decided to focus on the *Quantified McRae Feature Norms* (QMcRae) dataset (Herbelot and Vecchi, 2015). McRae's feature norms (McRae et al., 2005) contain 541 concepts with, on average, around 11-12 typical attributes for each (2201 overall), which were obtained from an association task with humans. The quantified version of the dataset added human frequency judgements in terms of the five quantifiers *no*, *few*, *some*, *most* and *all*. Underlying our joint project is the idea that a representation for a concept, or instances of it, should contain information about these typical attributes and their frequency, to some degree at least.

2 Project Progress

Prior to the STSM, we started working on preparing the data we intended to use. We considered two datasets with images of scenes and detailed annotations of objects, relations, bounding boxes, etc. Sandro looked at *MSCOCO Attributes* (Patterson and Hays, 2016) and I looked at *VisualGenome* (Krishna et al., 2016).

2.1 Investigation of available datasets and annotations

We spend the first days of the STSM investigating the contents of these datasets with respect to the QMcRae dataset. In particular, we were interested in the number of images we could find for each concept and, even more important, the number of images per concept for each attribute. MSCOCO Attributes turns out to contain only 29 object categories and 196 attributes, of which 22 overlap with the QMcRae concepts and 39 with the attributes. However, each of these few concepts comes with many images (100+). VisualGenome does not adhere to a small fixed set of labels and

hence showed a promising overlap of around 73% (394 of 541) for concepts and around 36% (802 of 2201) for attributes (using a simple word overlap matching strategy). However, a closer look at the distribution for each concept and attribute showed that most combinations only had only very few (1-20) images and some combinations heavily dominated the overall distribution (500+).

While not optimal with regards to a balanced dataset, in a way this situation resembles the real-world experience of humans. Some objects dominate the visual experience of an individual person, while the data for many other objects is relatively sparse. Nevertheless, learning concepts and attributes as well as a fairly good estimate of their relative frequency seems possible. Hence we decided to use the VisualGenome data with its broader, but often more sparse and skewed coverage distribution.

2.2 Task, experimental setup and open problems

The next step was to decide about the details of how to set up an appropriate task that allows investigating quantifier learning in deep neural networks:

- What exactly will be the structure of the training data?
- How do we obtain data of this format from VisualGenome?
- What role do the QMcRae quantifier annotations play for the task?
- How do we handle the imbalanced distribution, particularly when it is different from the QMcRae annotations?

Early on in our conversations, we decided to emulate the sequential experience of concepts and attributes, from which humans learn frequency estimates. Accordingly, the training data would consist of sequences of images, with some images showing the concept-attribute-combination in question, while others show examples of either only the concept or the attribute, and still others present entirely unrelated scenes. A system would then be trained on identifying the correct quantification of the target concept-attribute-combination given such a noisy sequence of visual "experiences".

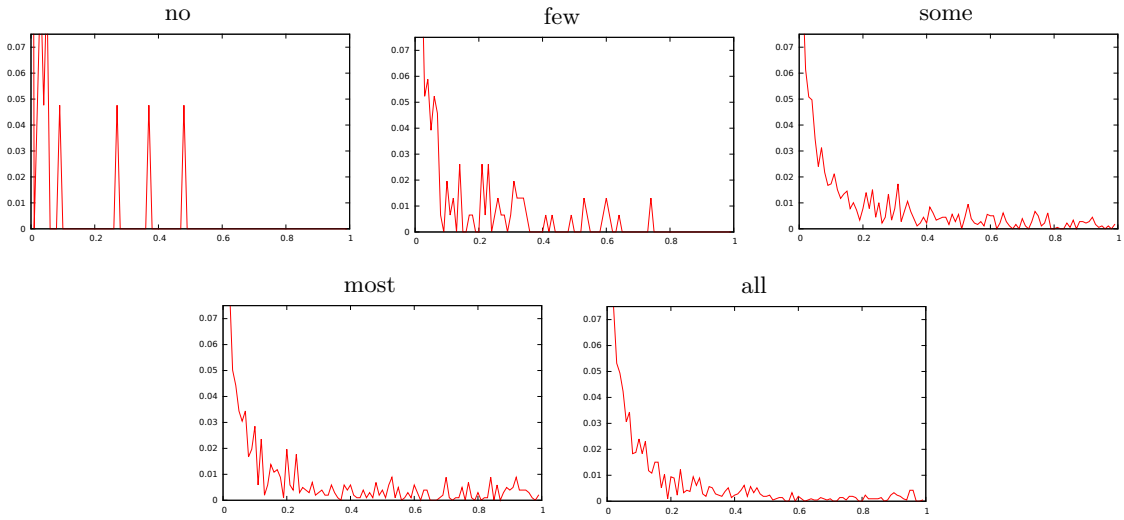
However, we realized that there are two problems with this setup. On the one hand, the task is very similar to the task of visual question answering. Datasets for this task usually contain count/quantification questions and, what is more, recently proposed architectures split up the image into smaller image parts, which then are processed sequentially. Such a behavior is essentially not too different from our idea of experience as sequence of images. In fact, the group in Trento was already working on this type of visual question answering for another project.

On the other hand, the problem of how to exactly extract these sequences of images remained. One solution would be to randomly extract images, but that would make a sequence with an interesting quantification pattern very unlikely, since most of the sequences would contain only one or two images per concept or combination. A second solution attempt was to control the frequencies of concepts and attributes within a sequence and consequently force more interesting situations. This would lead to rather artificial data without any relation to the QMcRae quantifier annotations. However, resembling human quantifier learning and ideally reproducing the QMcRae frequency estimates for typical concept attributes is a main motivation for our project. This is why, eventually, we abandoned our initial idea of presenting the quantifier learning task as sequences of past-experience images.

2.3 Insufficient VisualGenome annotations

While investigating the VisualGenome data and comparing it to QMcRae, we found that the VisualGenome annotations are in fact far from sufficiently detailed. For instance, the attribute "has_legs" appears only 88 times in over 1348 dog images. On the one hand, this shows that the attribute-per-concept distribution is different from what the QMcRae quantifier annotations suggest. On the other hand, this particular example suggests that many attributes shown in an image are not annotated, since presumably (almost) all dogs in VisualGenome have legs, which most often are also visible in the image.

The following figures show the frequency distribution of concept-attribute-combinations associated with a certain quantifier in QMcRae according to the VisualGenome annotations. These curves do not correspond to what we would expect as distribution for the respective quantifiers. In particular note the almost equal distributions for *some*, *most* and *all*, and the dominance of low frequencies in each of them. Hence we concluded that basing frequency estimates on attribute annotations in VisualGenome will definitely not yield data of good and reliable quality.



2.4 A new task addressing these problems

This observation led us to the question whether image embeddings in general contain sufficiently detailed attribute information, as opposed to dataset annotations which clearly do not. They definitely contain concept information since they are successfully trained on object classification, but to solve this task efficiently and be able to generalize, they presumably use attribute information to some degree. If it is possible to extract this information, we would expect to obtain attribute frequencies closer to what the QMcRae annotations suggest, and hence also better quantifier distributions than the ones shown above.

To investigate attribute information in image embeddings, the annotated examples of concepts and attributes from VisualGenome can be used as positive data. A simple approach is then to just take the average of the embeddings as the respective concept or concept-with-attribute representation. More sophisticated approaches could be trained, for instance, involving randomly sampled negative instances to improve embeddings in a word2vec-style fashion. We hypothesize that concept-attribute-combinations annotated with a quantifier suggesting lower frequency will be more distant from their parent concept representation than more "typical" characteristics.

3 Results and future work

In the two weeks of the STSM, we analyzed two available datasets with attribute annotations in detail, and uncovered various shortcomings with respect to our intended application. Moreover, when specifying the task setup in detail, we realized a close connection to recent experimental practice for the task of visual question answering. We eventually decided to shift the focus of our project, from evaluating quantifier learning based on image sequences to the question whether deep neural networks learn the "typicality" of visual attributes for concepts, expressed in terms of the QMcRae quantifier annotations. We realized that existing image annotations only very poorly reflect this, obstracting their use for more in-depth quantifier learning investigations as we were initially planning.

We will continue our work on this project and have regular Skype meetings. In the next weeks, we plan to implement and evaluate several systems for obtaining concept embeddings and measures for concept proximity of an image. We are confident that we can greatly improve on the poor correspondence between dataset annotation frequency and human-estimated frequency in the Quantified McRae Feature Norms. Our plan is to eventually report our results in a joint publication.

References

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Lisbon, Portugal. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris Mcnorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Genevieve Patterson and James Hays. 2016. COCO Attributes: Attributes for people, animals, and objects. *European Conference on Computer Vision (ECCV)*.

Ionut Sorodoc, Angeliki Lazaridou, Gemma Boleda, Aurélie Herbelot, Sandro Pezzelle, and Raffaella Bernardi. 2016. "Look, some green circles!": Learning to quantify from images. In *Proceedings of the 5th Workshop on Vision and Language*, Berlin, Germany.