Understanding the Meaning of Images

Daniela Mihai
Electronics and Computer Science
University of Southampton, UK
adm1g15@ecs.soton.ac.uk

Jonathon Hare
Electronics and Computer Science
University of Southampton, UK
jsh2@ecs.soton.ac.uk

ABSTRACT

The ability to process visual details is one of the most interesting parts of the nervous system. Humans have the ability to quickly analyse and understand the meaning behind a situation depicted in an image or video capture. This capacity is essential because it allows us to carry out different tasks in our every day lives. Training machines to recognise and tell the story behind an image can greatly benefit our lives. The importance of this research topic can be observed in areas such as surveillance, driving assistance and human-machine interaction.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**; **Learning latent representations**; **Interest point and salient region detections**.

KEYWORDS

Deep Learning, Image Understanding, Image Description, Factorised Latent Representations, Computer Vision

ACM Reference Format:
Daniela Mihai and Jonathon Hare. 2019. Understanding the Meaning of Images. In *Proceedings of ACM Celebration of Women in Computing womENcourage 2019, September 16–18, 2019, Rome, Italy.* 1 page.

1 INTRODUCTION

Developing machines that can perceive and understand the surrounding visual world and can communicate with us about it in natural language has been a long-standing goal [3]. Connecting images and natural language is a difficult task because it requires both the content of an image being understood and translated to its meaning, i.e. put into words. The main challenge in previous approaches is that the internal representation of an image is not directly intelligible and in order to effectively communicate its message, only the most important features must be transmitted. In this research, a new method is proposed which better resembles the way humans reason about and describe visual scenes.

2 PROPOSED APPROACH

2.1 Image Description Generation

Existing approaches for generating image description involve two main steps: feature extraction and description generation. So far our research has been focused on the first part of the problem. Unlike previous approaches where deep neural networks have been used to extract image features which cannot be interpreted, we aim to find a disentangled latent representation which captures interpretable meaningful features. After that, a language model can use these to generate the sequence of words which best describes the scene.

2.2 Disentangled Representation Learning

The desire for disentangled latent representation of images is not unique. For example, β-VAE [2] and Info-GAN [1] are two scalable unsupervised approaches for disentangled factor learning. Our approach is different from the known examples in that we aim to learn a factorised representation by squeezing information about the image as much as possible while still preserving its semantics.

2.3 First Steps

In the first stage of this research, a parametric map from the latent representations of two autoencoders to a shared factorised representation was learnt. The autoencoders had previously been trained on different subsets of the same dataset, either CIFAR10 or MNIST. This framework simulates two autoencoders sharing knowledge about the world they have been exposed to. By learning an optimal shared latent space, i.e. one which allows 'messages' from one autoencoder to the other to preserve the meaning of the image, we are able to get a sparse and interpretable latent representation of the depicted visual scene. The mapping between the two internal representations was learnt using a distance criterion between the encoding of the first autoencoder and its reconstruction in the second autoencoder's latent space. The weights of our ensemble were further optimised using a reconstruction loss between the input image given to the first autoencoder and the reconstruction produced by the second autoencoder given the information transmitted through the common latent space, and vice versa. Nevertheless, this framework could be improved by replacing the reconstruction loss with a criterion which only preserves the semantics of the image and does not enforce perfect reconstructions.

3 FUTURE PERSPECTIVE

The focus of this research is on developing more human-like learning and reasoning about image content. A collaborative framework is proposed for extracting the most meaningful features which convey the main idea of an image. This approach alleviates the issues of an unintelligible internal representation. As future work, we intend to include an attention mechanism which will enable the system to attend to features in the order of their importance.

REFERENCES

[1] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems.* 2172–2180.

[2] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations.*

[3] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3128–3137.