Automating Meme Generation Through Deep Learning

Mario Garrido
Universidad de Chile
Santiago, Chile
mario.garrido@ing.uchile.cl

ABSTRACT

This work explores and builds upon the idea of automatically generating memes that are competitive with human-produced material in terms of hilarity or entertainment. In order to explore whether this is possible, it is proposed to construct a data set to train a deep neural network architecture capable of generating new captions for one meme template.

CCS CONCEPTS

• **Natural language processing → Natural language generation**; • **Computing methodologies → Neural networks**;

KEYWORDS

Natural language processing, Deep Learning, Natural language generation

1 INTRODUCTION

The last 10 years have seen hardware advancements which have enabled the application of powerful, yet resource intensive, techniques in the field of machine learning, particularly the area branded as Deep Learning[8]. These techniques are powerful enough to allow the replication of elements of increasing complexity, a surprising example being human face generation by generative adversarial networks. In order to perform these feats, a model must be generated and trained over a very large data set comprised of samples similar to what we wish to classify or emulate. We say that, through this training, the model learns an abstract probability density function, from which to sample items from, that could be regarded as similar to the original samples'. How well this p.d.f. approximates the actual p.d.f. of the samples, something that depends on the quality and size of the training data set and also on the model's capacity, will determine the quality of the generated elements (or the quality of the classification, in other applications).

Several successful architectures have been designed for the task of machine translation and also automatic image caption generation, but the models require substantial amount of training samples to properly approximate the subjacent p.d.f. Luckily, current global network infrastructure and usage practices have yielded a reality where a vast amount of data sets can be collected from the internet, through scraping, so, even though the quality of the collected samples could be brought up to question, for a substantial amount of tasks there is most definitely an easy to acquire start-point data set waiting to be collected and curated.

Under these conditions of powerful consumer hardware, capable of running state of the art Deep Learning architectures, and readily available samples waiting to be collected from the internet, the following idea is explored: Can the machine be trained to automatically generate new multimedia elements on par, in terms of quality, to those generated by humans, in a given field? In order to explore the possibilities brought forth by this line of reasoning, an idea is developed to find out whether the machine could be tasked with the automatic generation of memes, which are, nowadays, ubiquitous on the internet.

Considering a meme to be the union of a template (background image) and an accompanying text (caption), it should be noted that this problem could be potentially tackled from, at least, 2 angles: Directly generate the composition of both the template and the text as a finalized image or consider both elements as separate entities, thus only needing to generate the text sequences that go along with each template. In the following background section some works that could enable the pursuit of these 2 angles are lightly discussed, but ultimately, due to results found in related literature, the last angle is prominently considered.

2 BACKGROUND

In this section some related works will be discussed lightly, and mainly from the perspective of their value to the idea of automatic meme caption generation.

2.1 GANs

This work[2] introduces us to the idea of using a model comprised of 2 parts: a generator part G and a discriminator part D. The generator is fed the training samples and must produce as output counterfeits that resemble the originals, while the discriminator must learn how to detect if a given sample corresponds to an original or a new item generated by the generator. As the 2 parts compete against each other and improve, the end result is that the generator has, not only, approximated the p.d.f. of the original samples, but can now be fed random Gaussian noise to produce new elements which are not identical to the originals but are sampled from the approximated p.d.f., so they are, in some sense, the same kind of items.

This is relevant because not only does it illustrate a model capable of generating new elements, but also does so with a sufficiently complex example as shown in the image[1].

2.2 Progressive Growing of GANs

In this work[4] we see a refinement of the GANs technique, yielding astonishing results over the same human face sample generation task. To achieve this, the authors train the Discriminator and Generator pair on a set of progressively better defined images, starting from samples with very low resolution, among other minor tweaks. It is fairly evident from the image[2] that automatic content generation has advanced dramatically the last 3 years.

Figure 1: Example of generated faces. Yellow column corresponds to closest face in training data set for the generated elements in the row.

Figure 2: Example of generated faces using GANs in a more sophisticated training configuration.

2.3 MaskGAN

This work[1] showcases an architecture based on GANs, but for the task of sequence to sequence text generation, which could be employed for the caption generation angle. However, this work is notable for being a rather convoluted effort of trying to fit GANs for a task which doesn't seem to favor the architecture too much. It seems that GANs would be well suited for direct meme image generation, but it'd be convenient to look elsewhere for standalone caption generation.

2.4 LSTM

It was the year 1997 when this architecture[3] was proposed, but it wouldn't be until 10 years later that it would experience a resurrection. This gated recurrent neural network is capable of consuming and outputting sequences, so it is well suited for text processing. The standalone model, however, falls rather short of what's necessary for today's applications, so it is of paramount importance to complement it with other tools in order to increase the capacity to an extent that allows us to capture the underlying logic of the studied meme.

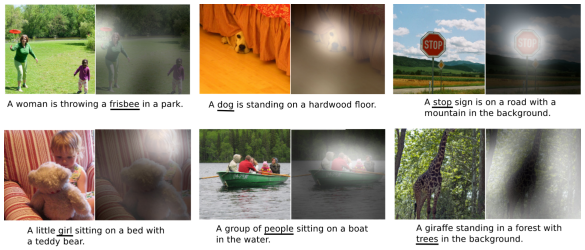

Figure 3: Example of visual attention in action. The highlighted elements are emphasized when the underlined word must be produced in the sequence.

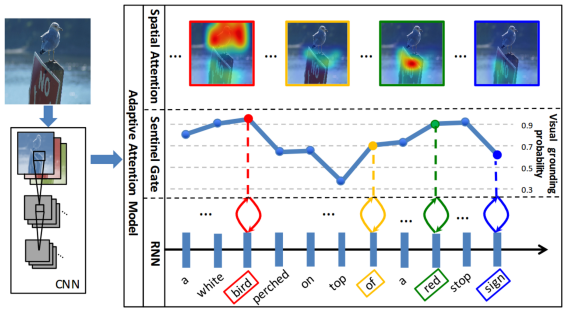

Figure 4: Example of visual attention in action. The highlighted elements are emphasized when the underlined word must be produced in the sequence.

2.5 Neural Image Caption Generation with Visual Attention

One of the required tools[7] is the so called Visual Attention. Attention is a mechanism that has been well developed in the field of recurrent neural networks, but usually in the word sequence to word sequence text generation setting. This work, however, proposes an attention mechanism which can be used over the features generated by a convolutional neural network over input images. In layman terms it tells the LSTM network what details of the image are important at each step of text output generation. Noting that the LSTM network outputs the elements of the text sequence one by one we see how this could be possible[3].

2.6 Adaptive Attention via A Visual Sentinel for Image Captioning

The last tool in the shed[5] complements the mechanism of visual attention. This proposed sentinel gate fills in the gap in the attention logic: What if the next element to be produced is better served by ignoring the contents of the image? To see why, we must think of language structure, many words are used to bind and connect others, so their presence in a text is better determined by an underlying understanding of the syntactic rules of a language, rather than the contents being presented in an image. The model filters[4] the results fed by the spatial attention, through a probability gate.

2.7 Dank Learning

This is the most related work[6], since the authors set out to accomplish the same goal: Automatically generate memes. Here we can see that a caption text generation approach using LSTM networks with visual attention[7] produces results which could be considered passable. However, these results are not completely satisfactory, as the quality of the generated memes is, at most, lacking. Also, the authors seem to have used few caption samples per meme template and they did not test the sentinel gate mechanism[5].

3 GOALS

Assuming that the underlying complexity of one meme can be captured in a Deep Learning architecture, with sufficient capacity, through language, then the chief goal is to manufacture a system that can automatically generate text captions for one meme template. Since humor pokes at the holes of our very understanding of reality, defying and reinforcing the elements that, according to our model of the world, should not happen or are, under that context, illogical or unexpected, it is a tall task to expect that a learning algorithm that has no underlying model to understand reality could generate hilarious content. However, there are many types of humor, and while some of those types rely heavily on context, other types tend to rely more on wordplay or themes that are easy to replicate with a basic understanding of a language. Can we construct, with modern technology and with available meme data, a system that captures the logic of more simplistic humor? Is it possible for a machine to have an operational understanding of the reality in which we operate simply through language? Is a good understanding of a language model sufficient to integrate with reality? Or, in other words, if we can make the machine construct an operational understanding of language would it be, then, easier to integrate to that model other glimpses of reality, such as images, sounds, etc? These questions are rather heavy for the current task of automatic meme generation, but they are tightly bound to it.

4 PROPOSED METHODOLOGY

In order to construct the seeked system, the following steps are proposed:

- Scrape the websites 9gag.com and memegenerator.com for all the available images of the meme *Socially awesome awkward penguin* and all its variants. The estimates are around 12.000 samples, which are way above the approximately 100 to 200 samples used in[6], by orders of magnitude. This should take around 3 weeks, due to exponential backoff to avoid overloading the servers with requests.
- Curate the data set, obtaining the captions for each of the images and defining the proper templates. This should take around 2 weeks, by training an OCR system and running it through the collected samples.
- Construct a Deep Learning architecture which uses convolutional neural networks to extract features and a visual attention[7] and sentinel gate[5] mechanism to feed an LSTM[3] cell with 4 layers of depth. Finding the proper hyper-parameters for the problem and correctly defining the main components of the architecture should take around 2 months.
- Train the model with the scraped data set and produce new memes. Since the quality of the results obtained here could force us to go back to tuning the model, this should take around 2 months.
- Define a way to test the memes. A simple solution would be to upload them to 9gag.com and see how they fare.

5 POSSIBLE DRAWBACKS

It should be noted that the most promising architecture, which is the one proposed, could fail to capture the underlying logic of the chosen meme, due to being fed bad features by the convolutional neural network, in which case it should be replaced by another, more suited, convolutional neural network, or downright abandoned in favor of a text sequence to text sequence architecture of LSTM cell, with attention. It could fail, also, due to having a small data set, in which case it would be necessary to augment the existing data set with more online samples. The estimates are that, by using other meme web pages, an additional 2000 samples could be collected from the internet, after filtering for overlaps with the original data set. If the main problem is a serious lack of training samples, then another meme with more samples can be chosen. What meme is modeled is not relevant, since doing one constitutes a sufficiently good proof of concept.

If the above idea would irredeemably fail, then we can consider the direct image generation approach, through the usage of GANs[4], in which case we need only modify the architecture being used, since the collected data set corresponds, originally, to images themselves. As a last ditch effort we could consider a third architecture option: MaskGan[1].

6 EXPECTED RESULTS

It is expected to train a system that can generate an endless stream of new samples for one meme. Once a meme that admits modelling has been found, it would be interesting to investigate the qualities that make it possible to approximate. Is it because the humor it employs requires less understanding of reality and is sufficiently served by an appropriate understanding of syntax? Is it because it uses a rather constrained subset of words from the language?

It should be noted that the samples are all based on the English language, it could be interesting to consider a language agnostic version of the automatic meme generator, although that would task the model with solving, not only, the meme problem but the machine translation problem.

Another interesting aspect of a working model for this problem would be to try to break it apart in order to understand what constitutes, according to the machine, the core aspects of this particular meme's structure. It could be possible for it to be different from the human understanding of it, but if it elicits laughter then it would push us to question if our understanding of reality is as good as we think it is.

REFERENCES

[1] William Fedus, Ian J. Goodfellow, and Andrew M. Dai. 2018. MaskGAN: Better Text Generation via Filling in the _____. *CoRR* abs/1801.07736 (2018). arXiv:1801.07736 http://arxiv.org/abs/1801.07736

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets

[3] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *CoRR* abs/1710.10196 (2017). arXiv:1710.10196 http://arxiv.org/abs/1710.10196

[5] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 3242–3250. https://doi.org/10.1109/CVPR.2017.345

[6] Abel L. Peirson V and E. Meltem Tolunay. 2018. Dank Learning: Generating Memes Using Deep Neural Networks. *CoRR* abs/1806.04510 (2018). arXiv:1806.04510 http://arxiv.org/abs/1806.04510

[7] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. 2048–2057. http://jmlr.org/proceedings/papers/v37/xuc15.html

[8] Yoshua Bengio Yann LeCun and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (May 2015), 436–444. https://doi.org/10.1038/nature14539