Comparing ResMem and MemNet

2021-03-24T16:24:07-05:00

In my previous posts about Memorability (see the project link above), I've been talking about the performance of my models fairly matter-of-factly. I've been comparing their scores on things, reporting them in abstracts, and talking about how one model performs better than another, and why I think that is happening. Some questions arise though, for example, why did I get such a vastly different score with MemNet than what Khosla reported in his original paper? Why is it possible, as has occurred in some cases, for rank correlation to improve while loss worsens? If a given model is better than another, shouldn't both measures improve? Why are we using rank correlation to measure the effectiveness of the model in the first place, if they're being trained on MSE Loss?

Memorability and "Performance"

Memorability is a tricky thing. It fits somewhere between vision and memory. While it's fundamentally intertwined with human behavior, ability, and cognition, it's been shown to be an intrinsic attribute of an image. Memorability is measured as a score from 0 to 1, that describes roughly the probability that someone will correctly recall the image in laboratory conditions. The problem is, the number changes depending on those laboratory conditions. In the memorability test

that has become the standard (Isola et al. 2014) a image may receive a score of 0.79. A test that is designed differently would yield a different absolute score, for example, a test that flashes the images for a shorter period of time. We can also imagine a test that skews the results in the other direction. Thus, we need to consider what this score actually is.

We have two claims to reconcile, that the score can be manipulated by other experimental conditions, but it is also innate to the image. This tells us that the memorability score must be ordinal, and not cardinal. In other words, if we have a set of ten images, with memorability scores derived from the Isola test, and we run these images through another experiment, while we may get different numbers, they will be in roughly the same order. This makes regressing upon memorability a difficult task. If the absolute number isn't universally consistent, then how do we model it?

We also have an opposing problem; the notion of "Performance" in deep learning is not straightforward either. Any paper you read on a new deep learning architecture will report a series of numbers at you. The standard ones are MSE Loss, Categorical Cross-Entropy, Maximum Likelihood, and the like. For more complex tasks we might see GLUE (Wang et al. 2018), SuperGLUE (Wang et al. 2020) scores in the natural language space. We might see things like Average Log-Likelihood or the Inception Score for GANs. In a paper you see one score reported. That score is often an average of a number of runs of the best model that the researchers made, over the course of a year. ResMem took me somewhere on the order of 400 experiments over the course of four months. A standard deep learning model takes ~3000 experiments to perfect (Karras et al. 2020). With every experiment, there is some sort of evaluation stage, and the best thing for a researcher to do is to have that stage spit out a number (or a set of numbers) that describe performance. Of course, those performance

figures have uncertainty attached to them too, leading us into an epistemic quagmire. On top of that, you need to choose one performance metric to be the loss function, the thing that is being optimized. The loss function must also be differentiable, which many of these metrics are not. It helps if the loss function can take a single x, y pair as an input, but it isn't required.

So, for measuring performance on the memorability task, we are dealing with two concepts that are both amorphous. This poses a challenge. Memorability estimations in the past have used Spearman rank correlation as their measure of goodness-of-fit. This arises from the fact that Spearman rank correlation is used as a measure of consistency within a dataset. In this context, the number refers to the rank correlation between the memorability scores generated by two subset of the test population. In the context of a deep learning model, the number refers to the rank correlation between the scores generated by human subject trials and the scores generated by the model. In general it is thought that if a model's predictions have higher rank correlation with the data than the data has within itself, then the model has succeeded, and any further increases in rank correlation do not increase predictive power. I will argue that this is not necessarily true.

High Level Features

Performance of MemNet

Performance of ResMem

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What Makes a Photograph Memorable? IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7), 1469–1482. https://doi.org/10.1109/TPAMI.2013.200

Pieters, M., & Wiering, M. (2018). Comparing Generative Adversarial Network

Techniques for Image Creation and Modification. ArXiv:1803.09093 [Cs, Stat]. http://arxiv.org/abs/1803.09093

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 353–355. https://doi.org/10.18653/v1/W18-5446

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2020). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. ArXiv:1905.00537 [Cs]. http://arxiv.org/abs/1905.00537

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training Generative Adversarial Networks with Limited Data. ArXiv:2006.06676 [Cs, Stat]. http://arxiv.org/abs/2006.06676