Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Thesis author	Shubham Shubham
Thesis title	Image popularity prediction
Submitted	2023
Program	Computer Science **Specialization** Artificial Intelligence

Review author	Jan Hajič **Role** reviewer
Position	Institute of Formal and Applied Linguistics

Review text:

The thesis presents experiments with image popularity prediction as measured by feedback on social media, namely Reddit. The author collected two datasets sets of 2000 images and their popularities, reviewed relevant literature, and designed a deep learning pipeline that (1) replicated the state-of-the-art model as a baseline, (2) improved upon these results by adding a pre-trained language model for processing the image caption on top of the image itself and metadata derived from the social media post containing the image. The thesis meets all its goals.

My impression of this thesis is very good. This is what a good machine learning experiment should look like: well-defined and motivated (ch. 1), with a sufficient literature review (ch. 2), with a well-designed dataset (ch. 3), using a strong baseline derived from state of the art (sec. 4.1), with a good balance of re-using large pre-trained models and adding features and training task-specific ones (sec. 4.2) and and using multiple evaluation methods (ch. 4.3) to draw conclusions from the results (ch. 5).

The strongest point of the thesis is, in my view, that the author understood how their social media of choice – Reddit – differs in its patterns of user engagement from others that were used for previous experiments, and was thus able to filter out confounding factors (such as comment counts) to arrive The dataset analysis performed in section 3 is well done and provides a good understanding of the material. I also appreciate the report on various dataset scraping methods – knowing which approaches do not work is just as valuable as knowing which do.

The experiments are quite extensive. Instead of just using one pre-trained language model to try integrating the text modality via the image captions, the author tries a selection of pre-trained models both for the visual and the text modality. This dilligence pays off – there are rather significant differences among the models on the pics and earth datasets (in fact, I would have expected smaller differences).

If there is a weakest point worth mentioning, it is that at the end of the literature review, I

would have expected a short discussion of how the proposed method relates to the work reviewed in the chapter – what is adopted, what is new.

I do have several questions.

1. After PCA is applied (sec. 4.1.1, Feature integration and model training, p. 22) to obtain 20 dimensions from the 4711 visual features, further 1-D convolutional layers are added, with filter size 3 (according to tab. 4.1). What is the reason to use convolutions, with their spatial awareness, rather than fully connected layers? After PCA, what do the local field of view of convolutional filters achieve? If the PCA projection has dimensions in order of significance, then this locality is not random, so I am *not* saying that convolution makes no sense here – I am just curious about what the implications of using convolutions rather than fully connected layers is. The resulting feature combinations (a FOV of 5 PCA output dimensions in the 3rd visual network 1D-conv layer) are then fed through a dense layer anyway, including inputs from the social network: why not just feed the 20 PCA output dimensions into the merge layer directly?

2. How (and when) is PCA projection computed? From the representations of training data?

3. I am not sure about the interpretation of least accurately predicted images from the earth dataset (figs. 4.6, 4.12). What stands out to me is how blue the 2nd and 3rd images are. Is it possible that this is a relatively niche input for pre-trained image model (and other image features), which would presumably expect much warmer colors from human skin tones (which contain very little blue at all) and sunlight? Also, the second image in 4.6 and the first and second images in 4.12 exhibit lateral symmetry, which is a composition that is usually a mistake unless the photographer knows very well what they are doing (like in these images), and rarely happens without extra photographic effort to align the image properly, so again not a lot of training data will look like this. Do you think these are plausible reasons, or not? Why?

4. Do I understand correctly that in 4.3.3 "Accuracy per Score Difference Bin", the "accuracy" is the accuracy of the binary task of ordering a given pair of images according to their true score, with the bins corresponding to the score difference, so that the higher we go in the score difference, the easier it should be for the model to order the image pair according to the score correctly? If yes, this is indeed a very well-designed metric that shows a fundamental difference between the pics and earth datasets. This would have merited a bit more discussion in the Conclusions, perhaps, so let me ask here instead: what insights for predicting image popularity on social media do you derive from this phenomenon?

5. More broadly: would you expect fundamentally better results with (1) newer image representations (e.g. Stable Diffusion), (2) LLMs?

The text is written clearly and isn't unnecessarily complicated. The level of English is great, there are only minor mistakes. I have just a few comments.

- A full-page diagram of the LM-enhanced model that would include the dimensionalities at the various interfaces between model components would make the model easier to understand – there are many different kinds of features used (image, text, "social", dimensionality reduction is involved).
- It would be worth re-stating what exactly the "score" (the target value) is in the results section, just in order to help the reader remember the experiment design logic. This was the only thing I had to really go back and re-read a chapter for, when determining whether the results were being computed and interpreted correctly.
- In 4.3.3, it would have been helpful to explicitly define the "accuracy" per score difference bin – this was harder to understand from the text than it needed to be.

Overall, this is an excellent thesis and I wholeheartedly recommend attempting to publish the results as soon as possible.

I recommend the thesis for defense.

I suggest to consider the thesis for the annual award.

This thesis is a great example of a complete machine learning experiment done right, including data collection and analysis in a tricky setting – social media. The idea of adding a model for text modality content is a good one, it is executed well, and leads to a clear improvement over the state of the art. What more can one ask for in a Master thesis?

August 25th, 2023

Signature: