Skylar Wurster

CSE 5544

Professor Chen

Jan 31st, 2019

Assignment 2

[DeepCompare: Visual and Interactive Comparison of Deep Learning Model Performance](http://sugeerth.cs.ucdavis.edu/Papers/deepcompare.pdf); Sugeerth Murugesan, Sana Malik, Fan Du, Eunyee koh, Tuan Manh Lai; VDS 2018

DeepCompare aims to improve on traditional methods of comparing deep learning models by aiming to show qualitative characteristics of why a model might outperform another model for some subset of the input domain. Sugeerth Murugestan from UC Davis, Sana Malik, Fan Du, and Eunyee koh from Adobe Research, and Tuan Manh Lai from Purdue University offer a visual analytic approach called DeepCompare to assess these tradeoffs.

The team began their work with the intention of illuminating to humans how very complex models are currently working in comparison to each other. They note that the advent of deep learning and the usefulness of extremely large models, it has become significantly more difficult to interpret what they are doing, and how they can be improved. Other methods exist to dissect a single model, but none to compare multiple. This is especially useful for comparing networks with different architectures altogether.

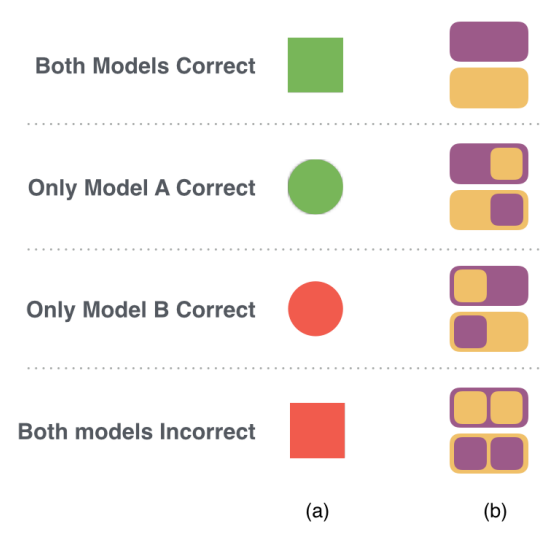
It is worth noting that these researchers interview five machine learning practitioners to find out what kind of tool would be useful for them. Since I am not a machine learning practitioner myself, I will not make any suggestions for improvement upon the three tasks and sex challenges the writers design from their interview with the researchers.



This visualization is comparing two models, a CNN and an LSTM, both trained on the samem data and with accuracies within 0.2%. Panel (a) is the Neruon Weights Detail panel, which shows the weights of one layer of the model as a heatmap, where green is higher activation, and red is lower activation. We can see from this image that the CNN is set to layer 5 and the LSTM is set to layer 3. If a single node is selected from panel (a), panel (b) populates with data about that specific neuron. The data here shows the neuron activation distribution through all the times the trained model was run. Panel (c), named the Test Result Detail panel, lists each of the inputs in order of which the selected neuron fired the most. In this panel we also see a comparison of the two models for that input. Under “result”, we see a background color (yellow or purple), and up to two squares on either the left or ride side of the background. The background color indicates the ground truth, the left square indicates the CNN’s result, and the right square indicates the LSTM’s result. These inputs can also be filtered to show items that passed or failed in certain circumstances. Lastly, panel (d) labeled Test Result Summary shows an overview for both models. It breaks the input dataset into negative (yellow) and positive (purple). From there, area is used to show correctness for both models within the negative and positive sentiments.

Overall, we see good use of Schneiderman’s “Overview first, zoom and filter, then details on demand”. Panel (d) offers a high-level overview of the accuracy of each model on the test set used. Color is used to distinguish between the negative and positive ground truths, accompanied by a legend above. The panel uses area to depict a proportion, which isn’t a good choice. We see that it is difficult to see whether the CNN correct in the negative sentiment is the smaller or larger than “both incorrect” in the positive sentiment side. A bar chart, which utilizes length, would better convey this piece of quantitative data. Two bar charts could be made, one with purple bars for the positive sentiments, and one with yellow bars for the negative sentiments.

Moving down and filtering, we come to panel (c). Since the actual input sentence fragments are more understandable than individual neurons, filtering by the fragments might be the next step a researcher would want to look at. The visualization gives us filter options to find instances of one model being correct and the other one being incorrect, vice versa, or where they’re both incorrect, or both correct. This coloring system went through iterations, and the researchers decided to scrap their previous glyph system which was misleading due to adding shape when unnecessary. Below shows the old glyph system on the left, and the new coloring on the right.



When clicking on a sentence fragment in (c), the neurons in panel (a) align from most activated to least activated for that given input. This intractability allows for the users to find trends in the data that is difficult for an algorithm to point out.

Following that, panel (a) works in much the same way as panel (c). Clicking on a neuron rearranges panel (c) to show which inputs cause that neuron to activate the most, giving a great deal of information to the user. For instance, one node might respond heavily to Spanish fragments. This is exactly “details on demand”.

Panel (a) is the most cluttered with numbers that often don’t have labels. Initially, it must be explicitly told to the user or inferred that the squares represent neurons. Further, the positioning of these squares is confusing. The squares are left+top aligned, so smaller squares appear in the top left of the character box. This causes awkward spacing between columns that might be interpreted as something significant from a user. If these squares were center aligned, this issue would disappear. We also see a lack of a scale for size of each square. A color gradient exists for activation power, but no legend exists to describe what the size means. Additionally, since these depict quantitative data, it is curious that they have chosen to use both color and area, which are the two worst attributes for quantitative data, and not length or angle/slope. We also have no visual indication as to what the two rows are corresponding to in this image.

Overall, the conclusion of the paper shows useful results from the machine learning practitioners interviewed for this software. Improvements can be made specifically to panel (a) and (d). Panel (a) takes up most of the room but doesn’t do much with the space. Also, the order of the panels isn’t confusing, but doesn’t lend itself to the overview, filter, details process. Colors are used intelligently, and the overall aesthetic is nice.

[Interactive Analysis of Word Vector Embeddings](https://graphics.cs.wisc.edu/Papers/2018/HG18/embeddings_preprint.pdf); F. Heimerl and M. Gleicher; (EuroVis) 2018