50.021 – AI

Alex

Week 06: Something on interpretability

[The following notes are compiled from various sources such as textbooks, lecture materials, Web resources and are shared for academic purposes only, intended for use by students registered for a specific course. In the interest of brevity, every source is not cited. The compiler of these notes gratefully acknowledges all such sources.]

Due: ?

1 running LRP explanations of imagenet-pretrained neural nets in keras (easy!)

- install in a virtual environment innvestigate `https://github.com/albermax/innvestigate` and tensorflow according to the instructions
- Take `lrp_analyze2_s.py`
- take a few images, run LRP on them for at least 2 neural networks. save the resized image and the heatmap outputs
- add according to the innvestigate doc analyzers for guided backprop and for gradient
- again take the same few images, run LRP-Preset-A, guided backprop and gradient on those images for the same 2 neural networks. save the resized image and the heatmap outputs for all three methods.
- find one image where gradient-based explanation is very much spread out across large parts of the whole image, and show the result of all three analyzers on it
- submit a pdf with resized images and heatmap results for all three methods (LRP-Preset-A, guided backprop and gradient).
- optional, for thinking: note that this code by default analyzes an image for the top predicted class. You can change that for custom outputs neurons (e.g. ground truth class of the image).

2 running t-sne on kaggle cats vs dogs

- take kaggle cats vs dogs (will be provided on a new image)
- take `tsne_codepieces.py` and complete the feature precomputation for a pretrained neural net and the t-sne part. for t-sne you can use the scikit-learn implementation
- optional: do the same for a neural net finetuned on cats vs dogs. that will give a different result.
- submit a pdf with plotted t-sne visualizations