ЖЧ. MLSS. 1) The Mythos of Model Interpretability. 2) ViM Out-Of-Distribution with Virtual-logit Matching. 3) Exemplary Natural Images Explain CNN Activations

Lipton, The Mythos of Model Interpretability. 2016

The paper clearly states why interpretability is important and what interpretability means to facilitate research and engineering goals in this area. The motivation behind interpretability is trust in the systems (give them control), causality (real world understandability), transferability (accuracy in the real world), informativeness (useful information), fairness. Interpretability can mean any of the seven listed terms, such as transparency, explanations and other.

Introduction. They note that many agree in that interpretability is important, but then interpretability is vaguely defined. One example is European Union laws propose the right of explanation to users for whom ML systems decide but what explanation means is undefined. This originates from that ML models are trained in specific environments (training and validation sets) while they operate in a real world. ML systems just minimize error (objective function) and their models don't catch up with the non-stationary world. Also, they raise the question that we do not understand ourselves still we want to understand a ML model we expect to be so intelligible.

2. Desiderata. Here they list why to research interpretability.

2.1. Trust. Here it is twofold understanding of trust. First it is that models should meat performance expectations in the real world. And another is that they should be wrong and right in similar cases as humans.

2.2. Causality. This is models should represent a real understanding of the world that is cause-and-effect relations that later can be used for research.

2.3. Transferability. Transferability means the expectation that models trained in their environments will perform well in the real world. This links to adversarial robustness which is the behavior of a system to specifically crafted examples that fool ML models.

2.4. Informativeness. ML systems should not only perform well but give meaningful information why they behave (decide or act) in a particular way.

2.5. Fairness. As ML models become pervasive, we expect them to make ethical decisions such as racial discrimination. So systems should give reasoning that can be tested and if they proved wrong, they should change accordingly.

3. Properties of Interpretable Models

Those fall into two categories. First one is observing how ML system works at the moment and another category is about after the work is finished.

3.1. Transparency.

They describe transparency from a high level to a low level:

- Simulatability is when given the state of the model we can calculate in reasonable time the outputs of the model.
- Decomposability. The property now is that a component or a part of a model (layer, a set of weights, etc.) should be intuitively understandable.
- Algorithmic transparency. This property states that algorithms used in ML systems should be understandable or provable.

3.2 Post-hoc Interpretability

This includes properties that enable us to understand models after their work done.

- Models should be able to give text explanations to their decisions.
- Visualization. It is when we visualize internal layers in CNN or generate images from a ready model by using the information on how different nodes activated.
- Local Explanations is a saliency maps which are images of what a model considered important for its output or what model focuses at.

- Explanation by example is when a model explains its decisions or output by a given analogy or another example similarly to humans justify their actions by behavior of others.

4. Discussion.

Here they list mythos and warnings regarding interpretability.

- Contrast to more or less common notion that one should take linear model instead of deep NN because they are more interpretable. This is true for transparency in algorithms because linear models has much simpler architecture compared, for example, to ConvNets or RNN deep NN. But linear models require heavy changed examples (one-shot, artificial categories, etc) while deep NN operator on raw data (they keep images sizes, raw text).
- We should avoid vague usage of interpretability in important context but use it in one or more senses as defined in the paper.
- There might be a trade-off between transparent and performant ML systems because of the human limitations.
- Visualization of layers in ConvNets, highlighting areas of focus in attention layers and other post-hoc interpretability might not be good in explaining how models decide, be deceptive.
- Still there is this gap between real world behavior of ML systems and at the training time which can be addressed by created better algorithms (uses better loss functions).

5. Contribution.

They conclude researches should avoid targeting their efforts on vaguely defined interpretability but use these specific notions of interpretability to aim the listed expectations from the ML systems (trust, causality, other).

Judge

Pros:

- The paper lists specific notions of interpretability and clearly defined motivation behind pursuing it.
- They pointed out why linear models are actually less interpretable in decomposability or simulatability which is surprising given the complexity of deep NN.
- They point out that current visualizations, text explanations and such might mislead which is important for safety.

Cons

- They did not address reinforcement learning paradigm.
- They leave the question of possibility of understanding the ML systems open without more clarifications for research or discussion.

ViM Out-Of-Distribution with Virtual-logit Matching

The paper presents an anomaly detection method (unusual data, out-of-distribution (OOD) examples, unknown unknowns, Black Swans) that incorporates feature, logit (input to sigmoid function) and probability spaces to assign an anomaly score that is based on an additional logit (additional class or virtual logit), i.e. Virtual-logit Matching (ViM). Also, they built a new benchmark for Image-Net-1K models specifically for OOD detection.

1. Introduction.

Anomaly detection is done by assigning anomaly scores to examples, this is the scoring function ϕ. They OOD examples are those the cross some threshold (0.95 for FPR95). Research found numerous ways to find those scores. Approaches without retraining the model are using the accuracy, using density, negative maximum prediction probability and other. So sources for the function are feature, probability, logit spaces. They propose a method to use all those spaces.

2. Related Work lists such methods in OOD, Network designs and exposure of OOD data. OOD methods they list are the maximum predicted softmax probability (MSP) (it uses the negated largest probability as an anomaly score), MaxLogit (similar to MSP but uses logits), and other. Also, there are methods that redesign their network to better output OOD scores with extra branches in network, with modifying loss function and other. Related work includes the methods that require retraining models, i.e. given them exposure to OOD examples.

3. Motivation. Here they prove it is not enough to use only feature, logit, or probability spaces for OOD detection.

4. ViM

Let P be the subspace of our weights. This space can be computed by finding principal subspace with such a method as Principal Component Analysis (optimization algorithm used to compress data to accelerate algorithms or for visualization; based on finding space closest to points). The orthogonal to this P is $P^{\perp}$. Then for an example x its projection is $proj_{P^{\perp}}(x)$. We add the length of this projection to our logits (input to softmax) (the logits then used to calculate probabilities for our classifier):

$$\alpha||proj_{P^{\perp}}(x)|| = l_0$$

Where α is a coef to match the scale of original max logit:

$$\alpha = \frac{\sum \max l_k^{(n)}}{\sum ||proj_{P^{\perp}}(x^{(n)})||}$$

Then the ViM score is:

$$\mathrm{ViM}(x) = \frac{e^{l_0}}{e^{l_0} + \sum_{k=1}^{C} e^{l_k}}$$

Intuition behind this is that it is like an additional virtual class, the anomaly score. Where C - number of original classes.

$\mathrm{ViM}(x)$ can be approximated to

$$l_0 - \max_{k \in \{1,\ldots,C\}} l_k$$

So intuition is that we take this virtual logit (length of projection to outer space) and subtract the maximum evidence (logit) that the classifier has.

5. OpenImage-O dataset

They for the in-distribution dataset ImageNet-1K they constructed OOD dataset called OpenImage-O. This has advantages over other benchmarks as it does not query other datasets by labels, and other.

6. Experiment. They conducted several experiments based on ImageNet-1K comparing ViM with other methods based only on feature, logit, probability spaces. They used more modern architectures such as ConvNet based networks (redisual networks), transformer-based (uses only self-attention mechanism) networks. Their ablation (removing some part of their method) tests indicated the effects of parameters (hyperparameter α, other).

7. They conclude they presented ViM, conveyed many extensive tests using different architectures and methods.

Judge

Pros

- They presented non-intuitive method for OOD detection that doesn't require retraining.
- Their method surpasses other methods that are based only on one source of the score OOD function.

Cons

- No comparison of ViM with the methods that require training a new model.
- No AUPR measurement which incorporates precision metric. Hence unclear if ViM has good precision which is how accurate a detector is. Consider if a test always gives positive result then its precision is low but its sensitivity is perfect.

Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization (2020) Borowski et al.,

Summary. This paper on the post-hoc transparency (monitoring ML) via feature maps visualizations of models aims to answer how good current images (synthesized or natural) explain neural nets (NN). It proves natural images are better than synthesized ones by ~10% (92% absolute) in accuracy of predicting if a chosen picture will activate a hidden CNN layer.

1.

In recent years researches found several ways to visualize internal representations of NNs, but it is unclear how those visualizations explain the inner work of NNs. Those visualizations try to represent what a particular part of trained NN (unit, feature map) responsible for by indicating what input activates it. Still, it is controversial. On the one hand, it seems those pictures show meaningful information. On the other hand, it is said that results are not representative (pictures were selected for papers), that actual activation images imperceptible (as adversarial examples show), and other objections. To resolve this disagreement, this study conducted a psychological experiment that found that those images allow to explain what activates part of NN. Also, they found natural images are better than synthesized.

2. In related work makes

First feature visualization idea was introduced by Erhan et al. (2009). It become active focus of research only after 2014 year. Important concepts here are saliency map (highlights the portion of an input that makes the most difference in the output), concept activation vectors (they learn how NN responses to concepts such as zebra), other.

Those works were criticized with that pictures were hand-picked, no falsifiable hypothesis, etc. Also, some argue that NNs should be studied as natural phenomena (like black holes, particles, etc.) with falsifiable hypothesis and also to incorporate philosophical findings.

No previous psychological studies were done to investigate inner work of NNs (only the final results). This work investigates inner work (feature layers activation).

3. Methods

The experiment is basically that they show reference pictures that minimally or maximally activate the feature map and they make a participant to choose between two other pictures that they think will activate the part of NN maximally. They conduced two types of this experiment, one type to get a maximum accuracy of predictions and another is to get general understanding of how good explanations work. Therefore, for the first experiment with expert participants, they selected those reference pictures that activate NNs to the full. For the second type of experiment with experts and non-experts, they did not select maximum ones but. Inside both experiments showed a various number of reference pictures. Also, they recorded confidence, response time.

4. Results.

4.1 Natural images are more helpful than synthesized ones

Participants responded with more confidence and faster with natural reference images. Results are 82±4% for synthesized and 92±2% for natural. When no reference images were presented (random choice), the results were below 50%.

4.2 Images allow to explain all layers in CNN (they used ImageNet) where lower-level layers handle textures, colors, etc. and higher levels are for high concepts (particular animal). Again, natural images are better.

4.3 Both experts and non-experts show the similar results.

This is important because this method of explanation of NNs aimed to help everyone (EU even states the right of explanation).

4.4 Same results for carefully chosen feature maps

As per previous critiques, carefully chosen layers give better results. This work also selected some feature maps, but the results were the same.

4.5 Additional reference pictures (min + max, only max, other) gives higher accuracy of predictions

As expected, the more information is given about how a feature map activates (minimally or maximally), the better predictions.

4.6 Here, they aggregate what participants said and felt about how intuitive the images are

Results are participants found those images neither intuitive nor non-intuitive. Intuition is slightly better to the end of experiments (participants learn how to better choose).

5 Discussion & Conclusion

Conclusion:

- Synthetic feature visualization is informative as it gives results higher baseline (random choice). Natural images (expected to be a baseline for synthetic images) outperformed synthetic ones in explaining feature maps.

- Experts (in CNNs) did not show better results compared to non-experts when they used synthetic or natural images.
- Experiments with hand-picked images or feature maps show the same results as with not selected ones, i.e. they demonstrate a general tendency.
- More information, esp. minimally activating pictures, improves explanation.
- People don't trust those pictures (there were no high score for intuition, it was in the middle).

Caveats & future work:

- They used max activating reference images. It is unclear what results will be with fewer activating images.
- There are other visualization methods not tested as they did.
- Still other things to be tested like hyperparameters.
- Also, if we add more reference pictures, if fewer activating images, if other NN architecture, other.

A Appendix

Here, they provide details of the experiment. How many participants (10+23), who they were presented images (in one session differently), how long, what hardware used, what software showed, what model used (CNN), what dataset (ImageNet), how synthesized images selected, etc. They also show examples of their screens that participants were shown.

Judge

Pros

- This is the first work that conducts psychological test on hidden feature maps representation.
- In the end they achieved ±92% accuracy for activation images, which is impressive. This suggests that we can explain inner feature maps with highly activating images.
- They collected information not only about correctness but also about intuitiveness, response time which allows to infer how people trust such images.

Cons

- Still unclear if true activation images are imperceptible to humans as adversarial examples show.
- The trial only with 33 people and 10 of them are experts in ML. The min final difference between natural and synthetic images result is only 6%. Thus it is unclear if the results will generalize to more people, i.e. will natural images be more explanatory than synthesized.
- Still, the question of trust is open. They used highly activating images and there might not always be such images for all layers. Also not clear if we can explain feature maps, does it allow us to understand the whole neural net?