Drift experiment – activations

Motivation

After reading a recent paper on using monitoring of statistical change of activations to detect input drift on an untrained neural network - [1]. I wondered how activations monitoring would work on a trained neural network, finding no literature on this I've begun a small experiment. The Experiment seeks to answer the question: Does monitoring model activations in different levels of a neural network detect drift differently to input space monitoring? And: Do any of the methods more closely correlate with the reduction in model performance?

Method And Results

To test this I trained a simple ResNet-18 architected NN which has been trained on the CIFAR -10 (https://www.cs.toronto.edu/~kriz/cifar.html) to do object recognition on the 10 object types within the dataset and achieved 93% accuracy on the validation data set.

A pipeline to introduce simulated drift of 4 types were created: Blur, Noise, rotation, and brightness was created.

Test images were run through the model at increasing levels of drift and input space monitoring using MMD, A middle layer (layer 3) using MMD and the final fully connected layer (fc) MMD, all compared to the corresponding measure from the unaltered baseline image.

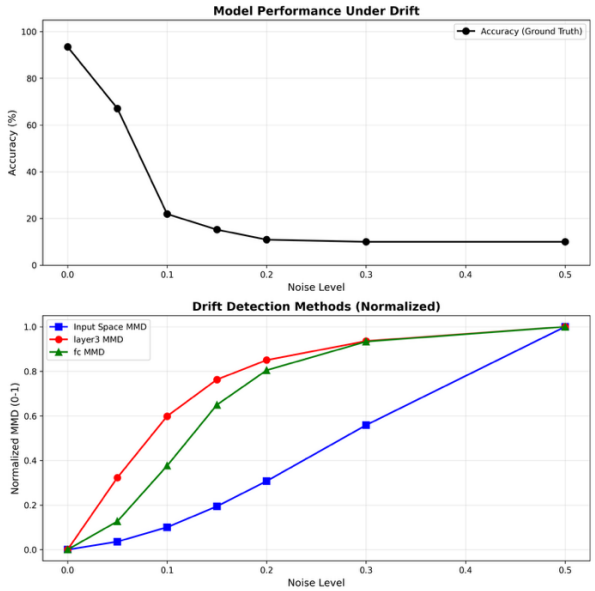

Fig 1. Model performance under increasing noise.

[1] Komorniczak and Ksieniewicz, *Unsupervised Concept Drift Detection Based on Parallel Activations of Neural Network*, vol. 15013.

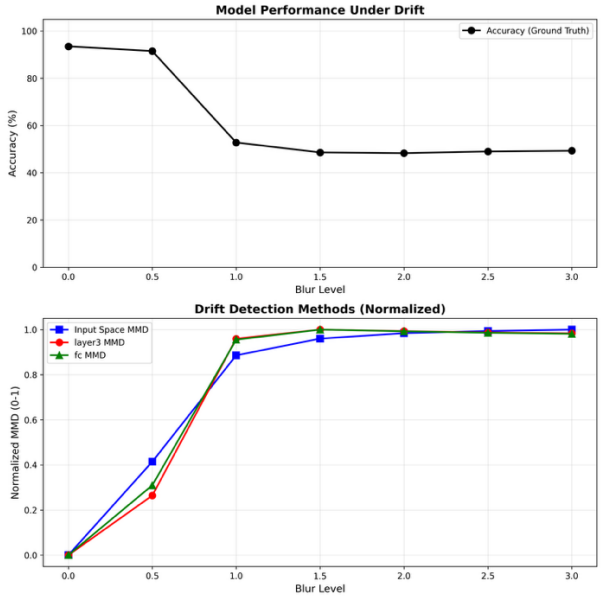

Fig.2 - Model performance under increasing blur.

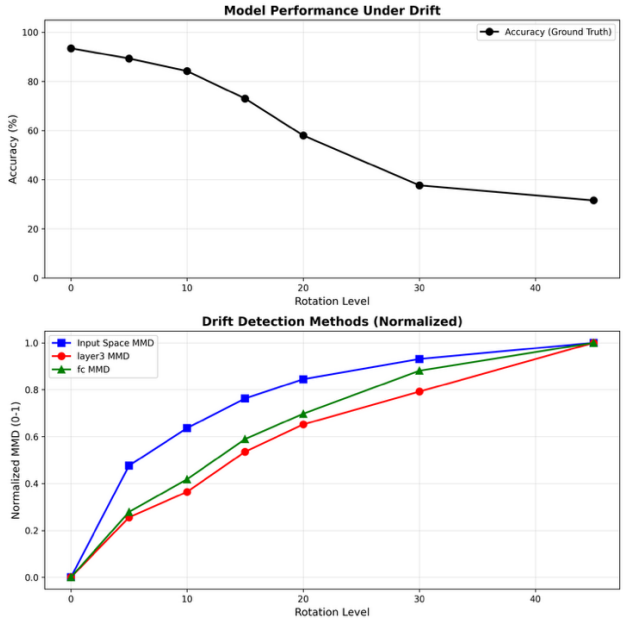

Fig 3. Model performance under increasing rotation.

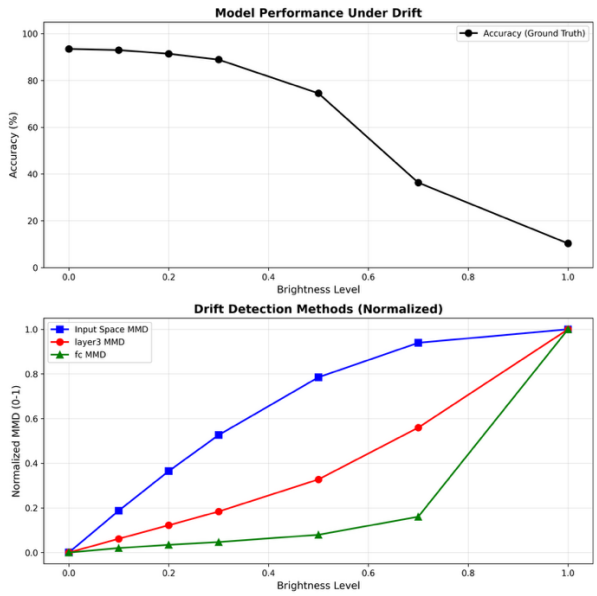

Fig 4. Model performance under increasing Brightness.

Discussion:

In Fig 1- Noise. In this figure we can see that monitoring input space for noise performs quite poorly, subtle noise changes in the image appear to be missed by the MMD statistical test, dramatic drops in performance are observed before MMD scores increase significantly. The model activation monitoring method performs better tracking model performance much more closely. The middle layer gives a more detectable score earlier, before the large drop in model accuracy, with the final layer tracking most closely.

Fig 2 -Blur. We can see that monitoring input space, a middle layer and the final layer all detect drift before accuracy is affected. They all then give an MMD score of close to one as soon as more blur is introduced and then it plateaus.

Fig 3 – Rotation. In this all the 3 methods track the accuracy line. Input space giving the largest clearest signal of drift, but it goes faster than the accuracy drops once again, model monitoring tracks more closely to actual measured model accuracy reduction.

Fig 4- Brightness. Once again, model activations provide a closer relationship to model performance. Final layer monitoring being aligned closely to the drop in performance. Though Input space giving a much earlier warning signal.

Conclusions:

In a situation where I need a drift monitoring system that only gives alarms when performance is being, or will soon be affected, then monitoring model internals and specifically the final layers provide the clearest signal of that. Input space monitoring is the most sensitive, detecting changes before model activations respond. Potentially combined system could be useful. Input space has detected drift (high sensitivity),

amber warning, model activations detect drift (high specificity), red warning review decisions made here model is or will be compromised soon.

Existing drift monitoring techniques are sensitive and can detect changes that don't affect actual model performance. This is useful, but for a radiology department, or cyber security tool being able to detect drift that you can be more confident is having an impact on your models, would also be valuable. Drift is common, a detector that detects, theoretical drift (not affecting model performance) and real drift (is affecting performance) are solving two different but related problems.

Finally in answer to our 2 research questions.

Does monitoring model activations in different levels of a neural network detect drift differently to input space monitoring? Yes.

Do any of the methods more closely correlate with the reduction in model performance? Yes, final layer signals particularly correlate closely with model performance drops.

Further work

To go beyond this toy experiment it would be interesting to test on real-world drift. Perhaps the same image acquisition protocol but between 2 hospitals using different equipment. It would then be interesting to see if the trend held true.

Looking at this the other way, could model activation monitoring be useful for difficult generalisation tasks? Could monitoring activation patterns as test set of images were passed through reveal what parts of the model are focussed on what sorts of features are processed by different areas. Or do certain areas change when certain features are masked or altered? This approach could be valuable in several scenarios: (1) diagnosing why models fail to generalize across domains by identifying which layers exhibit the most drift, (2) informing the design of targeted data augmentation strategies based on observed activation patterns under distribution shift, or (3) guiding selective layer fine-tuning by revealing which network components are most affected by domain changes.