- 1. Paper found from Dipkamal Bhusal's Hierarchical Explanations Class lecture slide

  Paper Link : https://arxiv.org/pdf/1806.05337

  Code GitHub link: https://github.com/csinva/hierarchical-dnn-interpretations

  Documentation for code: https://csinva.io/hierarchical-dnn-interpretations/

Short Summary

The paper introduces **Agglomerative Contextual Decomposition (ACD)**, a method for explaining deep neural network (DNN) predictions through hierarchical interpretations.

ACD provides a hierarchical clustering of input features and their contributions to the model's predictions, addressing the black-box nature of DNNs.

Key Contributions:

1. **Hierarchical Interpretations**:
   - ACD builds a hierarchy of input features using importance scores based on Contextual Decomposition (CD), extended to general DNN architectures.
   - This allows visualization of non-linear feature interactions.

Results:

- ACD visualizations reveal meaningful feature groupings that align with model predictions, enhancing interpretability.

- 2. Paper: Hierarchical Explanations for Video Action Recognition

  Paper Link: https://arxiv.org/pdf/2301.00436

  Github code link : https://github.com/sadafgulshad1/HIPE

Short Summary

❖ The paper introduces Hierarchical Prototype Explainer (HIPE), a novel method for video action recognition that interprets deep neural network decisions by leveraging hierarchical class relationships.

❖ Unlike conventional explainability methods that offer single-level explanations, HIPE provides multi-level insights by explaining predictions at the class, parent, and grandparent levels.

Key Contributions:

1. **Hierarchical Explanations**:
   - Incorporates human-like hierarchical reasoning into video action recognition, enabling multi-level explanations.
   - Example: A misclassified video can still provide useful insights through its correct parent and grandparent classifications (e.g., "water sports" and "sports" for a specific misclassified action).

2. HIPE Framework:

- Combines a 3D-CNN backbone (ResNet-3D) with a hierarchical prototype layer and hyperbolic embedding spaces for action classification.
- The prototypes are learned for child, parent, and grandparent classes, mapping spatiotemporal video segments to hierarchical explanations.

3. Experiments:

- Conducted on the **UCF-101** and **ActivityNet** datasets.
- HIPE improves hierarchical accuracy metrics (e.g., sibling and cousin accuracies) compared to baseline models like ResNet and ProtoPNet.
- Demonstrated robustness in scenarios of misclassification by providing meaningful hierarchical insights

4. Results:

- HIPE recovers performance drops due to explainability compared to non-interpretable models while offering richer, multi-level explanations.
- For UCF-101, HIPE achieves better clip-level and video-level accuracies compared to regular ProtoPNet.