Hierarchical Explanations for Video Action Recognition

Sadaf Gulshad, Teng Long, Nanne van Noord
University of Amsterdam
{s.gulshad, t.long, n.j.e.vannoord}@uva.nl

Abstract

To interpret deep neural networks, one main approach is to dissect the visual input and find the prototypical parts responsible for the classification. However, existing methods often ignore the hierarchical relationship between these prototypes, and thus can not explain semantic concepts at both higher level (e.g., water sports) and lower level (e.g., swimming). In this paper inspired by human cognition system, we leverage hierarchal information to deal with uncertainty. To this end, we propose HIerarchical Prototype Explainer (HIPE) to build hierarchical relations between prototypes and classes. The faithfulness of our method is verified by reducing accuracy-explainability trade-off on UCF-101 while providing multi-level explanations.

1. Introduction

When describing the world around us we may do so at different levels of granularity, depending on the information available or the level of detail we intend to convey. For instance, a video might open with a shot of a cheering crowd, allowing us to recognize it as a *a sports event*, as the camera then pans to the river we can deduce that it is *a water sports event*. However, only when the raft comes into the frame can we determine that it concerns *rafting*. Nonetheless, in our description of this video, we may still only refer to it as a sports or water sports event. Our reasoning and description processes build on the hierarchical relation between concepts, allowing for navigation between generic and specific. In this work, we implement this process for video action recognition by learning hierarchical concepts that we leverage for improved classification performance and explanations at multiple levels of granularity.

Prototype-based models, [6, 8, 21] focus on learning prototypes during training and make predictions based on the learned prototypes during inference. This enables *this look like that* explanations. However, previous case-based reasoning works are limited to 2D images and models. Moreover, they provide a single level of explanations and in case of uncertainty, the explanations can be as bad as arbitrary, as each explanation is considered equally apart. In this work, we focus on capturing the hierarchical relations between actions to provide multi-level explanations for videos.

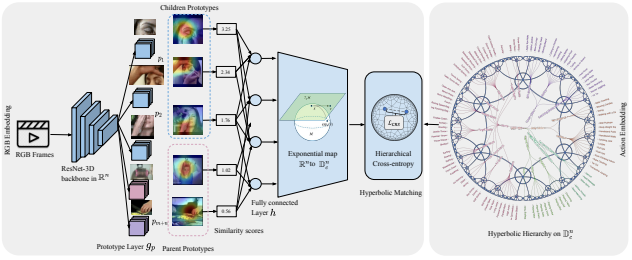

Figure 1. **Overview of the Hierarchical Prototype Explainer.** The Resnet-3D backbone extracts video features and the prototype layer learns prototypes for children and parents, these prototypes are then converted to a single similarity score through max pooling. Finally, scores are converted from $\mathbb{R}^n$ to $\mathbb{D}^n$ through a fully connected layer followed by an exponential map, to the shared hyperbolic space for hierarchical learning. Actions are mapped onto the shared hyperbolic space by learning a discriminative embedding on $\mathbb{D}^n$.

A challenge for explainable models is that it introduces an accuracy-explainability trade-off, where explainability comes at the cost of accuracy. With this paper, we aim to introduce a model with built-in explainability which is less affected by this trade-off. To achieve this goal we are inspired by recent works on learning hyperbolic embedding spaces, as opposed to euclidean, in natural language processing [7, 41] and computer vision tasks [1, 11, 25]. This belief is guided by the hierarchical cognition process of humans, that is likely to organize concepts from specific to general [28, 29, 46] and the representation of categories in the hyperbolic space.

Our contributions are: 1) We propose HIPE, a reasoning model for interpreting video action recognition. 2) We demonstrate that HIPE can provide meaningful explanations even in the case of uncertainty or lack of information by providing multi-level explanations i.e., at class, parent, or grandparent level. 3) We perform a benchmark and show that HIPE outperforms its non-hierarchical counterpart.

2. Related Work

2.1. Interpretations for Videos

Interpretations for neural networks can be broadly classified into two categories: 1) fitting explanations to the decisions of the network after it has been trained i.e. *posthoc* [12, 14, 19, 27, 34], 2) building explanation mechanism inherent in the network i.e. *built-in* explanations. [22, 23, 42, 43]. In this work, we focus on learning semantic representations which are used for classification during training rather than explaining a black box network posthoc.

A great deal of previous work has focused on video action recognition, detection, segmentation and more [3, 4, 13, 16, 33, 35, 36, 45], however, most of these works focus on designing black box models for specific tasks. They do not explain why a certain decision is made by the model. Moreover, most of the research in the domain of visual explanations focuses on images. Only a few works focus on the interpretation of these networks for videos [2, 15, 22, 38, 39], and it is not possible to directly apply image-based explanation methods to videos due to an extra time dimension in videos.

[15] and [2] focus on visualizing spatio-temporal attention in RNNs, CNNs are used only to extract features. Inspired by class activation maps (CAM) [48] for images [39] extended it for videos by finding both regions and frames responsible for classification. [22] utilized perturbations to extract the most informative parts of the inputs responsible for the outputs. Both [22, 39] are posthoc methods, which means they do not use explanations during prediction therefore they might not be faithful to what the network computes [8]. We enable multi-level built-in explanations for videos.

2.2. Case-based Reasoning Models

There are two main categories of case-based reasoning models: *concept bottleneck models* which introduce a bottleneck layer that learns human understandable concepts, and *prototype-based models* that learn prototypes that are closer to the samples in the training set. Concept bottleneck models provide posthoc explanations by replacing the final layer of the neural network with a layer that predicts human understandable concepts *e.g.* for a cardinal bird class the concepts will be red wings, red beak, black eye [18, 20, 26, 44]. These predicted concepts are then used to perform classification. However, concept bottleneck models require dense concept annotations for the model to learn them. [32, 47] focus on addressing these limitations by either incorporating concepts by transferring them from natural language descriptions or generating them from a GPT model. In contrast, our work focuses on providing built-in explanations by learning representative samples for each class and its (grand)parent class. Our method do not require heavy annotations but utilize either the hierarchy available with the datasets or it can be easily defined based on the relations between classes.

The idea behind prototype-based models, to provide built-in explanations with prototypes was first explored in [21], where the authors introduced a prototype layer in the network with an encoder-decoder architecture. The prototype layer stores weights which are close to encoded training samples, and a decoder is used to visualize them. [6] further improved it by learning prototypes for each class and visualizing them by tracing them back to the input images without a decoder. We get inspiration from [6] to provide built-in explanations, but where their work is limited to 2D images and provides only one-level explanations we extend it to multi-level explanations for videos.

Most closely related to our work is [30], they use a decision-tree with a pre-defined structure, where individual prototypes are learned at each decision. The prototypes are optimised to increase purity along the path through the tree. However, for PrototypeTrees the position in, and order of, the tree does not describe a hierarchy, that is closer to the root does not imply a more general semantic level. Moreover, as the number of prototypes depends upon the size of the tree, learning a ProtoTree becomes computationally expensive. Our proposed multi-level explanations follow a very clear semantic distribution , where (grand)parent prototypes are more generic and do not add any extra computational complexity.

2.3. Hyperbolic Embeddings

Hyperbolic embeddings have recently received increased attention as they enable continuous representations of hierarchical knowledge [5, 31]. This continuous nature makes it such that information of (grand)parent classes is implicitly included, allowing hyperbolic training to remain single-label. Their effectiveness has also been shown for textual [10, 41, 49] and visual data [1, 11, 17, 25]. Hyperbolic embeddings have also been used for zero-shot learning [9, 24] and for video action recognition [25, 40]. The hierarchical relationship between videos and the hierarchical way of explaining decisions for humans calls for the need of using hyperbolic spaces. Here, we utilize hyperbolic embeddings for learning hierarchical prototypes to provide human-like explanations for video action recognition.

3. Hierarchical Prototype Explainer

Figure 1 gives an overview of our proposed Hierarchical Prototype Explainer (HIPE) for video action recognition. HIPE consists of a 3D-CNN backbone f for extracting features from the video frames, and a hierarchical prototype layer g_p for learning prototypes for each frame. The prototype layer is followed by a fully connected layer h that combines the prototype similarity scores and maps them to

the shared hyperbolic space through exponential mapping. Prior knowledge about the relations between actions, in the form of the action hierarchy, are projected to the shared space through discriminative embeddings. Subsequently, we use hyperbolic learning to obtain hierarchical prototypes that enable multi-level explainability.

As the backbone architecture, we use the video action classification network 3D-Resnet [13]. For each input video $v \in \mathbb{R}^{W \times H \times T \times 3}$ with T frames it extracts video features $Z \in \mathbb{R}^{W_0 \times H_0 \times T_0 \times D}$ with the spatial resolution $W_0 \times H_0$, frames T_0 and channels D. A key aspect of this backbone is that $T_0 < T$ due to temporal pooling, as such the features Z are extracted for segments rather than individual frames. Because of the temporal pooling, the prototypes learned by HIPE are spatio-temporal thereby explaining which parts of the segment are indicative of the action in the video.

3.1. Hierarchical Prototype Layer

Given the features extracted from the 3D-Resnet Z, and the set of action classes $\mathcal{A} = \{1, 2, ..., |\mathcal{A}|\}$, in hierarchical action recognition we also consider their ancestor classes $\mathcal{H} = \{|\mathcal{A}| + 1, |\mathcal{A}| + 2, ..., |\mathcal{A}| + |\mathcal{H}|\}$, which allows us to construct a hierarchical tree with three levels, i.e., grandparent, parent, and child (see Figure 1 right). These hierarchies can be easily defined by considering relations between classes, and do not require annotation of individual instances. The process of embedding the hierarchies is performed once, offline, per dataset. However, this process can easily be repeated for alternative hierarchies.

For each child $\mathcal{A}$ and its parent $\mathcal{H}$ action, the network learns m and n prototypes respectively $P = \{p_j\}_{j=1}^{m+n}$, whose shape is $W_1 \times H_1 \times T_1 \times D$ with $W_1 \leq W_0$, $H_1 \leq H_0$ and $T_1 \leq T_0$. As such each prototype represents a spatio-temporal part of the video. Given the convolutional output $Z = f(v)$ and prototypes p a prototype layer g_p computes the distances between each prototype p_j and the patches from Z and converts them to the similarity scores using

$$g_p(p_j, Z) = \max_{z \in Z} \log \frac{(||z - p_j||_2^2 + 1)}{(||z - p_j||_2^2 + \epsilon)}, \epsilon > 0 \quad (1)$$

The distances between each prototype and the patch determine the extent to which a prototype is present in the input. We learn different prototypes for child, parent and its grandparent as in Figure 1. We then multiply similarity scores with the weights of a fully connected layer h to obtain embeddings to be projected in the hyperbolic joint space for learning hierarchical prototypes.

3.2. Hierarchical Video Embeddings.

The embeddings $\mathbf{h} = h(g_p(p, f(v)))$ obtained from the fully connected layer are in the Euclidean space, we use exponential mapping [10] to map video embeddings into the hyperbolic space.

3.3. Training

Our training process consists of a multi-step procedure: In the initial epochs we perform warm-up of the newly added layers. Following the warm-up, we train the entire network end-to-end. Every 10 epochs we update the prototype layer only, followed by a phase of fine-tuning the layers after the prototype layer.

Our goal is to optimize:

$$\mathcal{L}_{crs} + \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{sep} \quad (2)$$

Hierarchical Cross Entropy $\mathcal{L}_{crs}$.

$$\mathcal{L}_{crs} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log p(y = k|v) \quad (3)$$

The Softmax in the cross entropy is defined as:

$$p(y = k|v) = \frac{\exp(-d(\mathbf{h}_e), \mathbf{\Phi}_k)}{\sum_{k'} \exp(-d(\mathbf{h}_e), \mathbf{\Phi}_k))}, \quad (4)$$

where $\mathbf{h}_e = \exp_{\mathbf{0}}(\mathbf{h})$ is applying exponential map to the fully connected layer output $\mathbf{h}$.

Hierarchical Clustering $\mathcal{L}_{cls}$.

$$\mathcal{L}_{cls} = \frac{1}{N} \sum_{i=1}^{N} \min_{j: p_j \in P_{|\mathcal{A}|+|\mathcal{H}|}} \min_{z \in patches(f(v_i))} ||z - p_j||_2^2 \quad (5)$$

Hierarchical Separation $\mathcal{L}_{sep}$.

$$\mathcal{L}_{sep} = -\frac{1}{N} \sum_{i=1}^{N} \min_{j: p_j \notin P_{|\mathcal{A}|+|\mathcal{H}|}} \min_{z \in patches(f(v_i))} ||z - p_j||_2^2 \quad (6)$$

3.4. Updating and Visualizing Prototypes

We project prototypes onto the closest video features from the training videos. We do so for child, parent, and grandparent action categories. Mathematically, for the prototypes p_j from child, parent and grandparent class i.e. $p_j \in P_{|\mathcal{A}|+|\mathcal{H}|}$, we update the prototype layer as:

$$p_j \leftarrow \underset{z \in \mathcal{Z}_j}{\text{argmin}} ||z - p_j||_2 \quad (7)$$

where $\mathcal{Z}_j = \{\tilde{z} : \tilde{z} \in patches(f(v_i)) \, \forall i \, \text{s.t.} \, y_i = |\mathcal{A}| + |\mathcal{H}|\}$. Our prototype layer is updated not only with the prototypes belonging to the child class but also with the parent and grandparent classes enabling the learning of hierarchical relations between classes.

We visualize the patch which highly activates for the prototype by forwarding the input through the network and upsampling the activation map generated by the prototype layer both spatially and temporally (for videos).

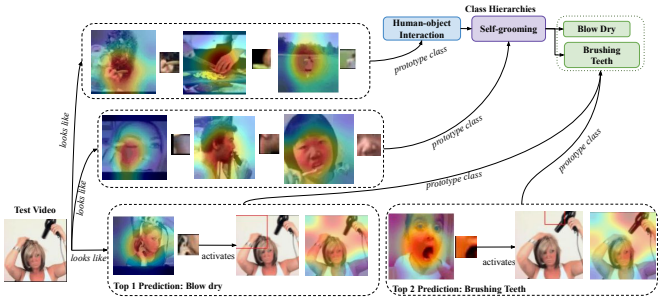

Figure 2. **Hierarchical Explanations.** This example shows the prototypes from grandparent *human-object interaction* class, parent *self-grooming* class and ground truth *blow dry* class, we also observe that the top 2 prediction for the model is its sibling *brushing teeth* class. This conforms that our model is learning hierarchical relations between classes.

Network	Accuracy	Sibling Accuracy	Cousin Accuracy
3D-Resnet [13]	83.34	89.73	93.62
Resnet-Hyperbolic [25]	82.64	89.99	93.28
ProtoPNet [6]	78.30	85.92	90.98
HIPE	**80.40**	**89.30**	**93.02**

Table 1. **Accuracy comparison for different models on UCF-101 videos.** We observe that HIPE recovers the drop due to accuracy-explainability trade-off significantly. The sibling accuracy is the rate of correct prediction 2-hops away, and the cousin accuracy as the 4-hops away from the ground truth.

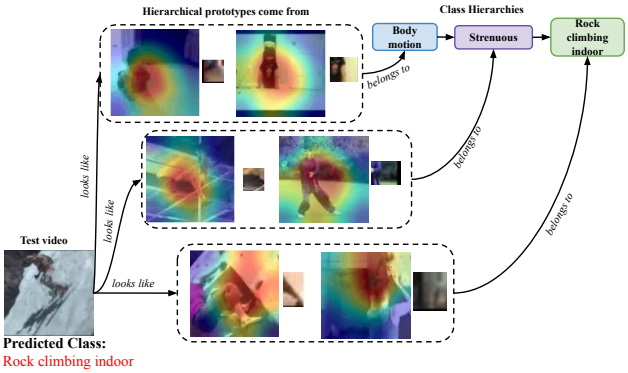

Figure 3. **Effectiveness in case of failure.** Our multi-level explanations provide useful information even in the case of misclassification through the prototypes learned for parent and grandparent classes.

4. Experiments

Datasets To evaluate HIPE for videos we conduct experiments on the UCF-101 [37] video dataset.

4.1. Visual Explanations

Hierarchical Explanations. Figure 2 shows the learned prototypes (only three out of ten prototypes shown for better presentation) from the grandparent class *human-object interaction*, parent class *self-grooming*, and the action class *blow dry* (only one prototype and its activation on the original video shown). We observe that the second most likely prediction is its sibling class *brushing teeth*. Thus, our model learns to represent the video clip features as hierarchical prototypes that belong to grandparent, parent and child classes.

Effectiveness of Hierarchical Explanations in case of Failure. In Figure 3 we show another scenario where the multi-level explanations are useful. We see that the original skiing video is misclassified into the *rock climbing indoor* class. However, for the more abstract explanations we can observe that its parent class *strenuous sports* and grandparent class *body motion* are correctly recognized. Hence our hierarchical explanations give us useful information even in the case of misclassification.

4.2. Accuracy-Explainability Trade Off

Table 1 contrasts the performance of two non-interpretable (top) with two interpretable (bottom) models on UCF-101. We can observe that HIPE is less affected by the accuracy-explainability trade-off whilst also providing multi-level explanations.

5. Conclusion

In this work, we proposed Hierarchical Prototype Explainer for video action recognition. By learning hierarchical prototypes we are able to provide explanations at multiple levels of granularity, not only explaining why it is classified as a certain class, but also what spatiotemporal parts contribute to it belonging to parent categories. Our results show that HIPE outperforms a prior non-hierarchical approach on UCF-101. Additionally, we demonstrate our multi-level explanations that make it possible to see which spatiotemporal parts contribute to grandparent, parent, and class-level classifications. Our hierarchical approach thereby provides richer explanations whilst compromising less performance to gain explainability.

References

[1] Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4453–4462, 2022. 1, 2

[2] Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff. Excitation backprop for rnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1440–1449, 2018. 2

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[5] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019. 2

[6] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 1, 2, 4

[7] Bhuwan Dhingra, Christopher J Shallue, Mohammad Norouzi, Andrew M Dai, and George E Dahl. Embedding text in hyperbolic spaces. *arXiv preprint arXiv:1806.04313*, 2018. 1

[8] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10265–10275, 2022. 1, 2

[9] Pengfei Fang, Mehrtash Harandi, and Lars Petersson. Kernel methods in hyperbolic spaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10665–10674, 2021. 2

[10] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655. PMLR, 2018. 2, 3

[11] Mina Ghadimi Atigh, Martin Keller-Ressel, and Pascal Mettes. Hyperbolic busemann learning with ideal prototypes. *Advances in Neural Information Processing Systems*, 34:103–115, 2021. 1, 2

[12] Sadaf Gulshad and Arnold Smeulders. Explaining with counter visual attributes and examples. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, 2020. 2

[13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 2, 3, 4

[14] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *ECCV*, September 2018. 2

[15] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015. 2

[16] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2

[17] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6418–6428, 2020. 2

[18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 2

[19] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *ECCV*, September 2018. 2

[20] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 2

[21] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 2

[22] Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. Towards visually explaining video understanding networks with perturbation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1120–1129, 2021. 2

[23] Xinmiao Lin, Wentao Bao, Matthew Wright, and Yu Kong. Gradient frequency modulation for visually explaining video understanding models. *arXiv preprint arXiv:2111.01215*, 2021. 2

[24] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9273–9281, 2020. 2

[25] Teng Long, Pascal Mettes, Heng Tao Shen, and Cees GM Snoek. Searching for actions on the hyperbole. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1141–1150, 2020. 1, 2, 4

[26] Max Losch, Mario Fritz, and Bernt Schiele. Interpretability beyond classification output: Semantic bottleneck networks. *arXiv preprint arXiv:1907.10882*, 2019. 2

[27] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 2

[28] James L McClelland and Timothy T Rogers. The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, 4(4):310–322, 2003. 1

[29] Marvin Minsky. Semantic information processing. 1982. 1

[30] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14933–14943, 2021. 2

[31] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 2

[32] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*. 2

[33] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In *European conference on computer vision*, pages 744–759. Springer, 2016. 2

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 2

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 2

[36] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3637–3646, 2017. 2

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[38] Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Ronald Poppe, and Remco Veltkamp. Class feature pyramids for video explanation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4255–4264. IEEE, 2019. 2

[39] Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Remco Veltkamp, and Ronald Poppe. Saliency tubes: Visual explanations for spatio-temporal convolutions. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1830–1834. IEEE, 2019. 2

[40] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12607–12617, 2021. 2

[41] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018. 1, 2

[42] Loc Trinh, Michael Tsang, Sirisha Rambhatla, and Yan Liu. Interpretable and trustworthy deepfake detection via dynamic prototypes. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1973–1983, 2021. 2

[43] Tomoki Uchiyama, Naoya Sogi, Koichiro Niinuma, and Kazuhiro Fukui. Visually explaining 3d-cnn predictions for video classification with an adaptive occlusion sensitivity analysis. *arXiv preprint arXiv:2207.12859*, 2022. 2

[44] Andong Wang, Wei-Ning Lee, and Xiaojuan Qi. Hint: Hierarchical neuron concept explainer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10254–10264, 2022. 2

[45] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 2

[46] Elizabeth K Warrington. The selective impairment of semantic memory. *The Quarterly journal of experimental psychology*, 27(4):635–657, 1975. 1

[47] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022. 2

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2

[49] Yudong Zhu, Di Zhou, Jinghui Xiao, Xin Jiang, Xiao Chen, and Qun Liu. Hypertext: Endowing fasttext with hyperbolic geometry. *arXiv preprint arXiv:2010.16143*, 2020. 2