Deployment and Explanation of Deep Models for Endoscopy Video Classification

(K Vikas Mahendar, Chandrashekar Lakshminarayanan, Arun Rajkumar)[†], Varun Seshadrinathan[+]
[†]Indian Institute of Technology, Madras
[+]MIMYK, IISc Bangalore

Abstract

Deep neural networks achieve state-of-the-art performance in several standard datasets across a variety of domains. While many pre-trained models on such standard datasets are available off-the-shelf, practitioners aiming to deploy deep learning based solutions to a specific real world application face two important challenges. Firstly, the standard models cannot be readily used and need significant reconfiguration and downstream training to suit the specific application. Secondly, in many critical applications involving humans, along with the model, it is also important to provide a mechanism to explain its decisions. In this paper, we address these challenges in the context of deploying deep models for endscopy video analysis. Our contribution is (i) a decoupled CNN-Transformer for classifying intubation procedures, and (ii) a mechanism that explains the model's decisions. The CNN-Transformer performs better than the baseline off-the-shelf-model with downstream training, and our explanations show that the CNN-Transformer model uses the right spatial and temporal features to arrive the final classification.

1 Introduction

Machine Learning models are supposed to be all-pervasive, to an extent that they should be deployable in high-risk domains such as self-driving cars [1], science [2] and medical-imaging [3]. To catalyze the rapid growth, researches have constantly contributed by developing several state-of-the-art models for a variety of tasks and data-representations. In spite of the advances, there still exits a gap between state-of-the-art models trained on standard benchmark datasets and those that need to be deployed in practical real-word applications. Firstly, models/architectures trained on benchmark dataset need to be extensively reconfigured and fine-tuned to the specific application settings in order to have satisfactory performance. In addition to this reconfiguration, it is highly pivotal that the model's decision making process is understood and trusted by humans. Especially, in high-risk domains such as medicine and healthcare, it is crucial to not only understand the decision-making capabilities of such deep models, but also allow humans to intervene at any stage of the pipeline to prevent such systems from making fatal decisions.

In this paper, we aim to build a deployable and explainable model in the domain of medical imaging. In particular, we look at the problem wherein a video of an endoscopic procedure is given and we need to classifying whether or not the endscopic probe was intubated correctly. Our specific contributions to this problem are

- We propose a CNN-Transformer model that encodes spatial and temporal information separately. We show via experiments that this model outperforms the baseline model.
- We use the learned attention weights of the temporal-transformer to visualise the important frames in which a key-action took place. This action could be performed correctly or incorrectly.
- Finally, we employ a Grad-CAM approach to provide spatial explanations so as to ground the salient spatial regions that the probe was attempting to navigate.

2 Related Work

2.1 Image Attribution Methods

One of the most straightforward approaches is identifying pixels or regions of the image input that influence the soft-max probability of the target label in an image classification task. These salient regions are represented as

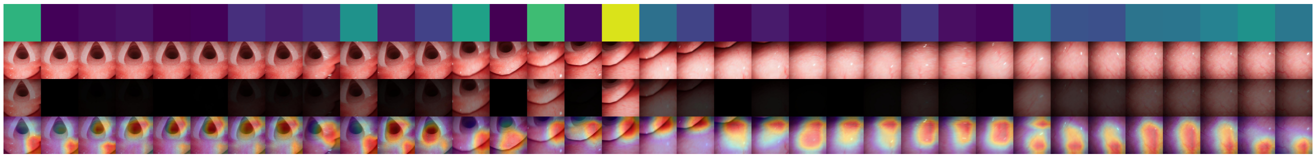


Figure 1: Visual Explanations of Video Transformer Model, First row indicates raw attention scores. Second row indicates the raw image frames. Third row superimposes the attention scores over the frames and the fourth row visualizes Grad-CAM explanations.

importance maps. Currently, there are many attribution procedures that are often applied to a model with fixed parameters to provide explanations.

Activation-Based Methods The activation maps that can be obtained from the internal layers of a convolutional neural network are linearly combined to generate the importance maps. The weights with which these activation maps are linearly combined differs across different methods. [4] proposed Grad-CAM in which the gradients from the activation outputs are average-pooled to obtain the weights. In Grad-CAM++ [5], the coefficients obtained from the second derivatives of the activation outputs are used as replacements instead of the average-pooling procedure in Grad-CAM. The major challenge associated with these methods is that the resolution of the intermediate maps is lower than the input image and hence the explanation is coarse-grained. When extended to the video-domain, such as for 3D-CNNs, same explanation results will be assigned to adjacent frames.

Backpropagation-Based Methods The underlying principle in these methods is that the importance of a pixel in determining the output is directly related to the gradient of the output with respect to that pixel. This is because gradients reflect the sensitivity of the output with respect to a change in input. Although several methods exist [6, 7], the resultant map generated is noisy. Despite elimination noise, methods like Integrated Gradients [8], do not generate class-discriminative explanations as they average-out the network properties. Certain works [9, 10] introduce a modified backpropagation method which inhibits user-friendliness/interpretability.

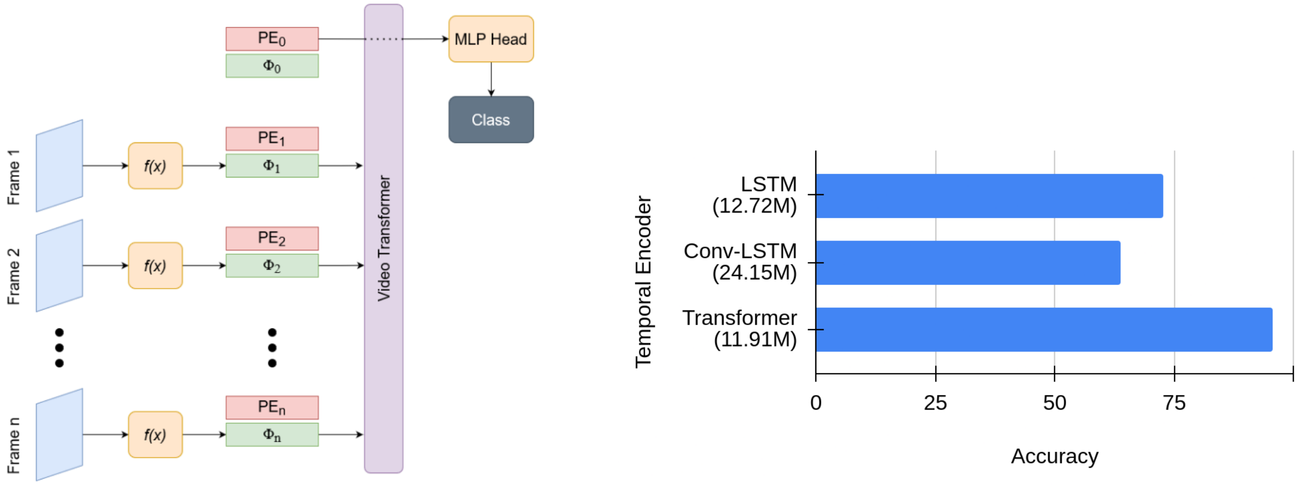
Perturbation-Based Methods These methods involve removing certain patches from the image input and observing the variation of the output with different input combinations. Although the reasoning seems intuitive, one has to traverse through all input elements and test all possible input combinations, which is computationally heavy. A major research focus in this space has been to reduce the computational overhead and obtain an optimal explanation. RISE [11] is a popular method which employs a sliding-window mechanism to obstruct a region of the input image with a grey patch. LIME [12] employs a simpler interpretable surrogate model (such as a linear model) to train on points sampled around the input example whose evaluations are used to generate explanations. Similar to LIME, [13] utilizes a second neural-network to predict a perturbation mask. [14] improvised the optimization process by introducing local gradients.

2.2 Video Attribution Methods

Video-representations have an additional dimension (time) than images. This means that an explainer has to look for regions in an inflated searching space and hence there is an increase in the computational overhead. [15] and [16] extended the gradient-based methods (like LRP) to identify the salient regions which are influential for a video-understanding network. However, these methods do not consider the temporal relationship between image frames and hence are not suitable. [17] and [18] adapt activation based methods for 3D CNNs to generate interpretations over time. Further, EB-R [19] (dubbed Excitation BackProp) is a popular method that utilizes a CNN-RNN structure to introduce time relationships. However these methods aren't intuitive by themselves especially in the video domain. While there are few works which focus on interpretability of video models, these approaches do not conduct extensive experimentation.

3 Intubation Classification

In this work, we focus on a binary-classification task which classifies video actions into a correct or faulty procedure. Specifically, we take a medical endoscopic intubation procedure, in which a probe (attached with a camera) is navigated from the mouth through the food pipe and to the gastroesophagl junction. Videos in which the performer correctly enters the food-pipe are considered positive-classes (label=1) and those in which the performer enters the trachea (wind-pipe) instead of the food-pipe are considered negative-classes (label=0).



(a)

(b)

Figure 2: (left) Our proposed architecture contains three modules - a feature extraction module followed by a temporal transformer that operates on ϕ_i (combined with positional encodings) and finally a MLP classification head that processes the output CLS token. (right) Performance improvement obtained by using a transformer for the temporal-encoder.

4 CNN-Transformer Model

The CNN-Transformer model (Figure 2(a)) is a hierarchical model composed of three core modules, namely, a 2D spatial encoder f(x), a temporal self-attention transformer composed of N layers, and a classification MLP head.

Spatial Backbone The spatial backbone is an embedding network that learns local spatial features for each frame, which leaves the following Transformer module to focus on capturing the temporal interactions. This embedding network can be any model that operates on 2D images and generates a d-dimensional vector for that frame. We make use of a ResNet50 [20] encoder, pretrained on ImageNet [21], as our spatial backbone. For each frame, the encoder generates a 512-dimensional feature vector which is then projected onto a 128-dimensional feature space (the latent representation dimension used in transformers).

Temporal Encoder We make use of the vanilla self-attention transformer [22] that enforces attention mechanisms to capture global dependencies in a sequential data-representation. Transformers take sequences of fixed length as input. After performing frame-level tokenization using the spatial backbone, transformers primarily model the temporal interactions between the frames. In our work, following [22], we extend a class token to learn global discriminative features. The frame-order information is enforced by the positional encoding step. The output state of the [CLS] token is passed onto the MLP classification head which is made up of two linear layers with RELU activation and dropout between them. In our model, Transformer module has 4 layers, 4 heads and 128-dimension hidden representation.

Design Rationale Our choice of transformers for the temporal module further makes the explanation pipeline inherently interpretable as opposed to post-hoc strategies (like attribution methods) currently available which are often not reliable, and can be misleading [23, 24].

Novelty In computer vision, a more sought-after direction for interpreting models is to generate visual explanations which represent the contribution of different segments in the input towards the final prediction. These methods (often termed as Attribution methods) in image domains, which have recently attracted attention of many researchers, often involve heatmap or saliency maps that localize the salient areas of the image for a model's decision [25, 26]. Albeit being successful, significant challenges arise when these methods are leveraged for inputs from the video-domain. A typical video-representation comprises image-frames stacked in a sequential fashion. Thus a target object/scene in such a representation would possess a spatio-temporal trajectory, and hence a model's outcome should be solely dependent on this target trajectory. Therefore, directly applying image interpretation methods in the video domain will not generate satisfactory explanations as they tend to disregard the temporal relationships.

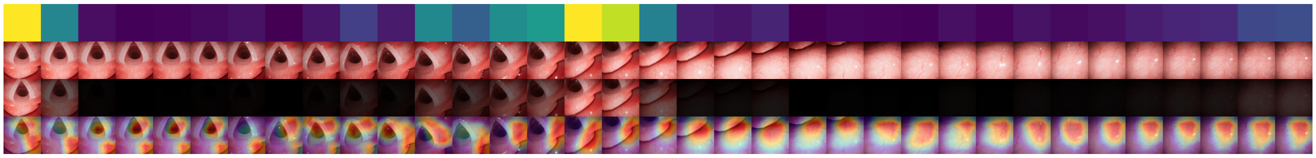


Figure 3: Visual Explanations generated by our approach. Note that whenever there is negligible motion flow, attention scores drop to zero (black frames) and hence neglects these frames for model's decision.

5 Experiments

Dataset Along with domain-experts, we collected 96 video samples, with each sample being 8-15 seconds long , captured at 60 frames-per-second. We use a stratified-split to generate 71 training and 25 validation samples. Further, we fix a maximum sequence length of 100 image frames, and videos are appropriately padded with null-frames (zero-vectors) or truncated accordingly based on its length.

Implementation Details We train the model end-to-end using the cross-entropy loss function. We randomly crop all image frames to 224 x 224 size and apply simple augmentations like horizontal flips and random rotation. During inference, we resize the images to 224 x 224. The experiments were conducted on a single GTX 1080 GPU machine, in which we used a batch size of 8. We optimize the model using an Adam optimizer. The learning rate is initialized to 1e-4 which is gradually decreasing by using a cosine annealing scheduler.

6 Results

Fig. 2(b) summarizes results obtained using different architectures. We first observe that our approach significantly outperforms the CNN-LSTM baseline by 22% (absolute). Our CNN-Transformer model has only 11.91 M parameters which further improves efficiency. During inference on a video sample, we extract the self-attention weights and compute a pairwise score to weigh the relation of each query frame with all other key tokens, which are then used to aggregate information from the values. Scores are averaged across attention heads, and the bottom 30% of the weights in magnitude are discarded, retaining the top 70%. Finally, the attention scores of the CLS token with respect to all other tokens are extracted and visualized to generate explanations.

Fig. 1 represents such scores for a validation sample. In the second row, darker colors (like violet) denote frames that have less of an impact on the model's decision, and a frame that is lighter in colour (like yellow) contains the essential action that can directly affect the output class. As can be seen from the figure, the model is correctly focusing on the influential frames when the surgeon enters the food pipe (17^{th} frame), as evidenced by the high attention score that is observed. The fourth row of images represents Grad-CAM explanations obtained from the spatial ResNet50 backbone. The probe's attempted entry points are accurately grounded (spatial regions) by the model. Fig. 3 visualizes another example of an interpretation.

7 Conclusion

Besides generating visual explanations, our paradigm can be used to unlock key-insights from the video actions. In other words, the attention scores can be primarily used as a tool to measure the learned agent's ability to perform an action as close to the ideal trajectory. For example, for a surgeon with little-to-no experience in performing endoscopy, our approach provides quantitative feedback that would help the surgeon to master the desired actions. Our approach further demystifies the black-box abstraction which allows us to intervene at any stage of the procedure and correct any faulty actions performed by the system. However, we have primarily conducted our experiments in a simulated environment and an agent which is trained on a simulated-environment would find hardship in performing actions in real-environment. In the future, given a trajectory of actions in one environment, we aim to explore ways that can be used to estimate the trajectory in a downstream environment.

References

[1] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art, 2021.

[2] Alexander Pritzel Tim Green Michael Figurnov Olaf Ronneberger Kathryn Tunyasuvunakool Russ Bates Augustin ˇZ ıdek Anna Potapenko Alex Bridglan Clemens Meyer Simon A. A. Kohl Andrew J. Ballard Andrew Cowie Bernardino Romera-Paredes Stanislav Nikolov Rishub Jain Jonas Adler Trevor Back Stig Petersen David Reiman Ellen Clancy Michal Zielinski Martin Steinegger Michalina Pacholska Tamas Berghammer Sebastian Bodenstein David Silver Oriol Vinyals Andrew W. Senior Koray Kavukcuoglu Pushmeet Kohli ohn Jumper, Richard Evans and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 2021.

[3] Zeynettin Akkus Bradley J Erickson, Panagiotis Korfiatis and Timothy L Kline. Machine learning for medical imaging. *Radiographics,*, 2017.

[4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019.

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2018.

[6] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Mueller. How to explain individual classification decisions, 2009.

[7] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[8] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

[9] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences, 2019.

[10] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, Stan Sclaroff, and Sarah Adel Bargal. Top-down neural attention by excitation backprop, 2016.

[11] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[13] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[14] Zhongang Qi, Saeed Khorram, and Li Fuxin. Visualizing deep networks by optimizing with integrated gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11890–11898, apr 2020.

[15] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alexander G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2568–2577. IEEE Computer Society, 2015.

[16] Christopher Anders, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Understanding patch-based learning by explaining predictions, 2018.

[17] 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)*. IEEE, sep 2019.

[18] Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Ronald Poppe, and Remco Veltkamp. Class feature pyramids for video explanation. Institute of Electrical and Electronics Engineers, 2020.

[19] Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff. Excitation backprop for rnns, 2018.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

[23] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, 2020.

[24] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, 2019.

[25] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks, 2019.

[26] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.