Project Report: Explaining video classification and regression models, with an application to Echocardiograms

J. Weston Hughes *
Stanford University
jwhughes@stanford.edu

1 Introduction

When machine learning methods in the real world make predictions, it is often required that they offer some explanation for those predictions. This is especially the case in medical fields, where doctors need to have high degrees of confidence in predictions if they are to be used in clinical decision making. Additionally, if models discover patterns previously unknown to medical science, it is hugely beneficial for models to be able to reveal that knowledge to physicians [6].

Applications of deep learning to video data have recently shown results similarly impressive to those in 2D vision, both in general and on medical problems [3, 5]. However, there has been surprisingly little work in applying explainability methods to video models. While many vision explainability models can be easily ported to video models, there is no review in the literature of how these methods work on video data, nor any consideration of how they should be updated for video. One particular area where we expect it will be possible to improve over single frame methods is in directly using optical flow information, which often may be the primary reason for a classification.

2 Methods

Previous applications of gradient-based attribution methods on images have treated individual pixels as features of the image, placing an importance score on each pixel. This makes sense, as the only prior on the correlation structure of these features is that adjacent features are likely to be similar. Video data, on the other hand, has the additional structure that pixels in different frames will depict the same object. While the positional relationship between objects in a single image may occasionally be important in classifying an image, the relationship between the location of an object in multiple frames of a video may be much more important. In fact, intuitively it seems there are two types of information that might be used to classify a video: the objects contained in the frames (along with their shapes, textures, etc.), and the motion connecting the frames. In light of this observation, we propose to decompose a video into features not based purely on pixels, but rather on a combination of pixel and movement information, in the form of approximate optical flow.

2.1 Optical Flow Decomposition

The optical flow between two frames is the apparent movement of the pixels in the two frames, determined by the intensity of each pixel. That is, for two images $I_0(x, y), I_1(x, y)$ it is a vector field (dx, dy) such that

$$I_0(x, y) \approx I_1(x + dx, y + dy)$$

*David Ouyang MD and James Zou will advise this project and will share authorship on any publication of this work. David will serve mainly to provide medical advice; James is the PI.

This problem is clearly under-specified, and a number of regularized and learned methods have been developed to approximate flow. In this work we use Farneback's method [2] to estimate flow, but any other method could easily be substituted. Additionally, previous work has considered how to differentiate through a vector field deformation of an image with respect to both the base image and the vector field [4], and their method is implemented in the grid_sample method in pytorch.

Our main contribution is to decompose a video $[I_0, \ldots, I_T]$ into three elements:

- The starting frame I_0, containing a large portion of the texture and object information in the video.
- The estimated optical flow f_i between each adjacent pair of frames I_{i-1}, I_i, containing the motion information in the video.
- The error E_i satisfying $I_i = g(I_{i-1}, f_i) + E_i$, where $g(I, f)$ is the application of the vector field f to the image I, containing additional pixel information obscured by the imperfect approximation of the optical flow, and the introduction of new objects into the frame.

and we re-parameterize the video as

$$\hat{I}_0 = I_0$$
$$\hat{I}_i = g(\hat{I}_{i-1}, f_i) + E_i$$

Note that indeed $I_i = \hat{I}_i$, but $\frac{\partial \hat{I}_i}{\partial f_i}$ is a meaningful expression, whereas $\frac{\partial I_i}{\partial f_i} = 0$.

Frames	$[I_0, I_1, \ldots, I_T]$
Optical flows	$f_i = \text{flow}(I_{i-1}, I_i)$
Errors	$E_i = I_i - g(I_{i-1}, f_i)$

2.2 Gradient-based attribution

Now consider a video classification model $\hat{y} = f(I)$, where I is a sequence of frames. Clearly we can compute $\frac{\partial \hat{y}}{\partial I}$, as in a typical naive gradient attribution method. Additionally, we can compute

$$\frac{\partial \hat{y}}{\partial I_0} = \frac{\partial \hat{y}}{\partial \hat{I}} \frac{\partial \hat{I}}{\partial I_0}$$
$$\frac{\partial \hat{y}}{\partial f_i} = \frac{\partial \hat{y}}{\partial \hat{I}} \frac{\partial \hat{I}}{\partial f_i}$$
$$\frac{\partial \hat{y}}{\partial E_i} = \frac{\partial \hat{y}}{\partial \hat{I}} \frac{\partial \hat{I}}{\partial E_i}$$

The gradients $\frac{\partial \hat{y}}{\partial I_0}$, $\frac{\partial \hat{y}}{\partial E_i}$ can be visualized as any typical gradient map would be. Visualizing importance in a vector field is more complicated; one option is two plot the top n vectors with highest gradient over the video.

2.3 Trajectory-based explanations

Additionally, we consider optical flow-based trajectorys through space and the saliency along these trajectories. It is largely straight-forward to determine point trajectories from optical flow data, and these trajectories have been considered as features for video classification in the past [8]. In the simplest terms, a trajectory P starting at time t might be defined

$$P_t = (i, j)$$
$$P_{t+1} = P_t + f_{t+1}(P_t)$$
$$P_{t+2} = P_{t+1} + f_{t+2}(P_{t+1})$$
$$\ldots$$

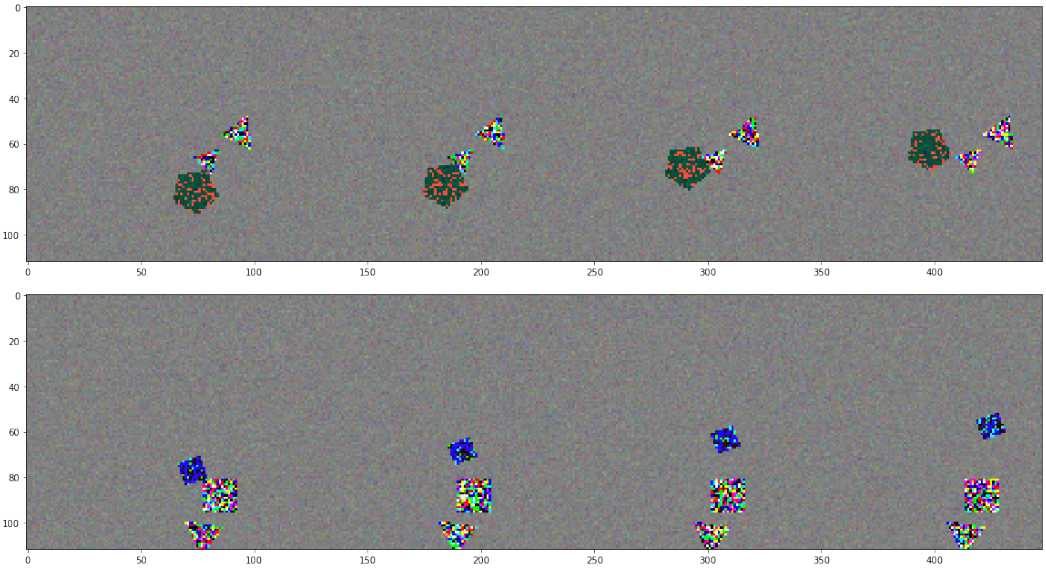

Figure 1: We developed the moving shapes dataset. For each video, the model must predict the number of sides and the angle of movement of the colored shape, in the presence of Gaussian noise and distractor shapes.

that is, starting from some point (i, j) at time t the appropriate optical flows are applied to "push" the point such that it follows a point of intensity in the video. Hopefully this trajectory will thus follow a moving point or object in the video. We can then choose subtrajectories along this trajectory such that the saliency along the trajectory is highest, and select the most salient sub-trajectories accross all sub-trajectories. The saliency of a sub-trajectory might be written:

$$S(P, k) = \sum_{k \leq t \leq k+10} \frac{\partial \hat{y}}{\partial x_{t,P_t}}$$

This method could also be jointly applied with the optical flow decomposition, giving a breakdown of whether the trajectory contains important motion or pixel information.

3 Datasets

We evaluate our methods on a toy dataset, an action recognition dataset, and on a specific medical task.

First, we designed the Moving Shapes dataset motivated by the desire to disentangle spatial and temporal information in videos. We generate 32 frame, 112 by 112 videos of a single shape of random color with 3, 4, or 5 sides moving at a random angle between 0 and 270 degrees across the frame. Optionally, we also add Gaussian noise, non-moving or white distractor shapes, and/or a pause in the shape's movement. There are two tasks on this dataset: to classify the number of sides the shape has, and to regress the angle the shape is moving at. This dataset is potentially useful for two reasons. First, we have some degree of "ground truth": while there are still many acceptable explanations, we can compare explanations in optical flow space against the correct angle, and in pixel space against the true location of the shape. Second, the sides task can be completed based solely on the first frame, while the angle task can be completed solely based on the optical flow. Thus we might expect that a good explanation method makes use of mainly the important part of the data for each task, and assigns little importance to the other part.

Second, we evaluated our method on Kinetics, the most popular large video action recognition dataset, sometimes referred to as the "Imagenet of video." This dataset is a good candidate for two reasons: first, there are many models developed specifically with the dataset in mind, and second, it is easy for someone with no training to determine if an explanation "makes sense" or not. It's also a

relatively noisy dataset with some classes which one might expect to be heavily motion-based, while other classes should be mostly classifiable based on a single frame.

Third, we will apply our method to an extended version of the EchoNet-Dynamic dataset [5], similar to the published dataset but with more videos and importantly many more classification labels. Important candidate tasks currently being explored include classifying heart attacks, heart failure, and diabetes, and regressing various clinically relevant pressure values. In this setting, explanations are important to increase physician confidence in models and explain new underlying mechanisms potentially discovered by the model (as would be the case in the diabetes classification problem).

4 Evaluation

Evaluation of explainability systems is challenging, since there are no metrics for the quality of an explanation outside of specific clinical workflows where physician improvement can be measured. Many explanation frameworks are presented without any quantitative evaluation, including highly cited works like SmoothGrad [7].

We've completed one set of quantitative experiments, and plan several more. First we evaluated various methods on our Moving Shapes dataset under different conditions, and consider how accurate the highest importance locations are with respect to the true location of the shape, as well as how accurate the most important angles are with respect to the true angle of the moving shape. Robustness to noise and distractors will be important comparisons. We also evaluated the methods using sanity checks for both data and model dependency. We compare both our own decomposition as well as explanations in pixel space.

Second, we plan to attempt to quantify the importance of flow versus pixel data in different classes in action recognition. For instance, in the UCF101 dataset there are distinct classes of "Handstand" and "Handstand pushup" which require flow information to distinguish between, while others like "Playing guitar" and "Playing sitar" have similar motion but different mostly static objects. An explanation system which disentangles motion and pixel information should be able to quantify this tradeoff. Additionally, if an explanation method gives importance to different frames, we can ablate the input and see if the important frames really are the most important in making a prediction.

5 Results

We evaluate our decomposition method on our toy dataset using both plain gradients and smoothgrad, and compare to plain gradients (See figure 2). First, we compute saliency maps for 100 examples, using the base model, a model with randomized logits, and a completely randomized model. We then perform "sanity checks," calculating the average Spearman correlation between the maps using the complete model and the two randomized models, as in [1]. We find that all considered methods have low Spearman correlations, indicating that they are heavily dependent on model parameters. Then, we calculate the average rank of points in and around target shape, as well as the points in all shapes (target and distractors) and of all points. We report the average target rank over average shape rank, and average target rank over average overall rank. We repeat the process for the fully randomized model, to ensure success here is model dependent. We find that the error saliency maps produced by our method are much more accurate than the baseline, and that the accuracy is lost when the model parameters are randomized. Thus, this method improves the quality of pixel-wise saliency maps. Confusingly, both optical flow explanations place more importance on the distractor shapes than the target shape. One hypothesis for why this might be the case is that because of the color of the distractor shapes, their optical flow is easier to predict, which biases the method. Although this would be easy to superficially fix, it seems to point to a larger issue with the method that should be addressed.

		Sanity Check		Accuracy		Randomized accuarcy	
Gradient Method	Input	Logits	All weights	Target over union	Target over all	Target over union	Target over all
vanilla	x	0.24	0.00	0.92	2.51	0.90	0.80
flow	error	0.20	0.01	**1.61**	**5.65**	1.27	0.26
smoothflow	error	0.56	0.14	2.42	0.02	1.82	0.03
flow	flow	0.26	0.00	1.43	6.83	1.29	6.08
smoothflow	flow	0.28	0.02	1.12	3.47	1.07	4.88

Figure 2: Results.

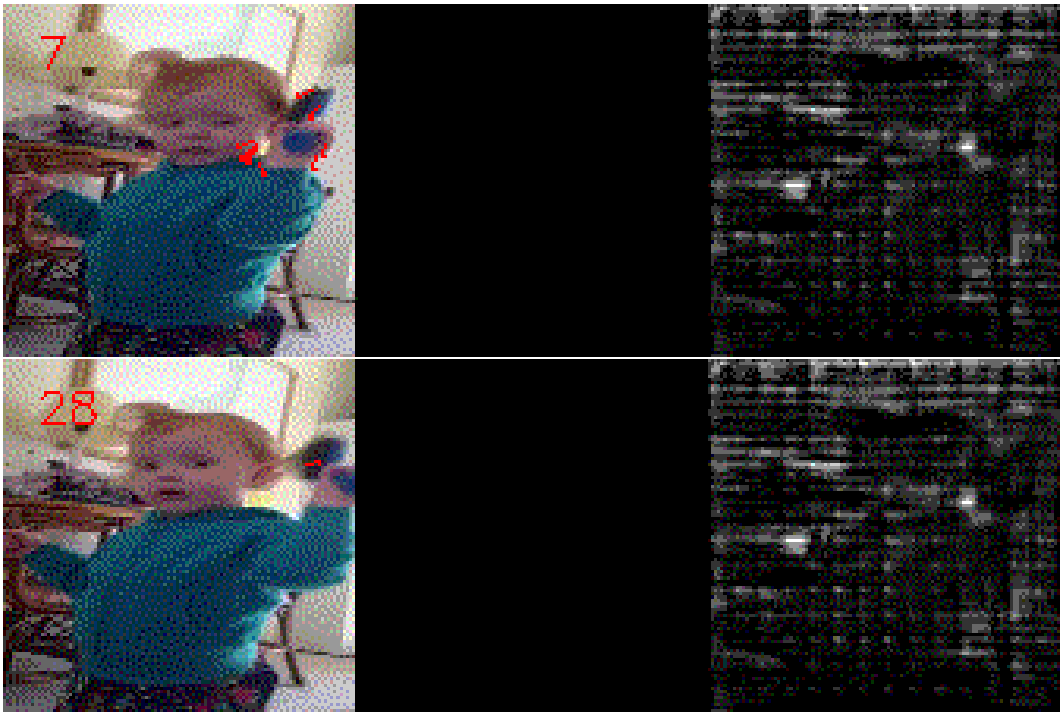

Figure 3: Two frames from a Kinetics explanation for “brushing hair". From left to right, the original video with optical flow superimposed, the importance of the first frame, and the importance of the error correction. For now, the left part looks good, while the other two can still use some work.

References

[1] Julius Adebayo et al. “Sanity checks for saliency maps”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 9505–9515.

[2] Gunnar Farnebäck. “Two-frame motion estimation based on polynomial expansion”. In: *Scandinavian conference on Image analysis*. Springer. 2003, pp. 363–370.

[3] Liam Hiley, Alun Preece, and Yulia Hicks. “Explainable Deep Learning for Video Recognition Tasks: A Framework & Recommendations”. In: *arXiv preprint arXiv:1909.05667* (2019).

[4] Max Jaderberg et al. *Spatial Transformer Networks*. 2015. arXiv: `1506.02025 [cs.CV]`.

[5] David Ouyang et al. “Video-based AI for beat-to-beat assessment of cardiac function”. In: *Nature* (2020), pp. 1–5.

[6] Rory Sayres et al. “Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy”. In: *Ophthalmology* 126.4 (2019), pp. 552–564.

[7] Daniel Smilkov et al. “Smoothgrad: removing noise by adding noise”. In: *arXiv preprint arXiv:1706.03825* (2017).

[8] Heng Wang et al. “Action recognition by dense trajectories”. In: *CVPR 2011*. IEEE. 2011, pp. 3169–3176.

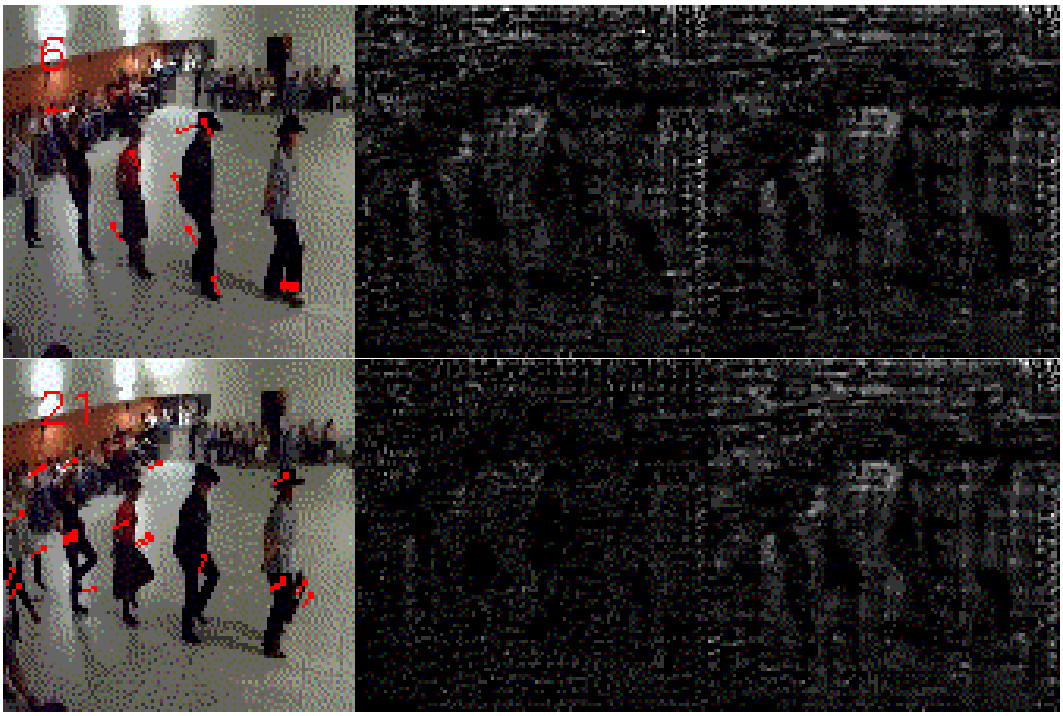

Figure 4: Two frames from a Kinetics explanation for “country line dancing".

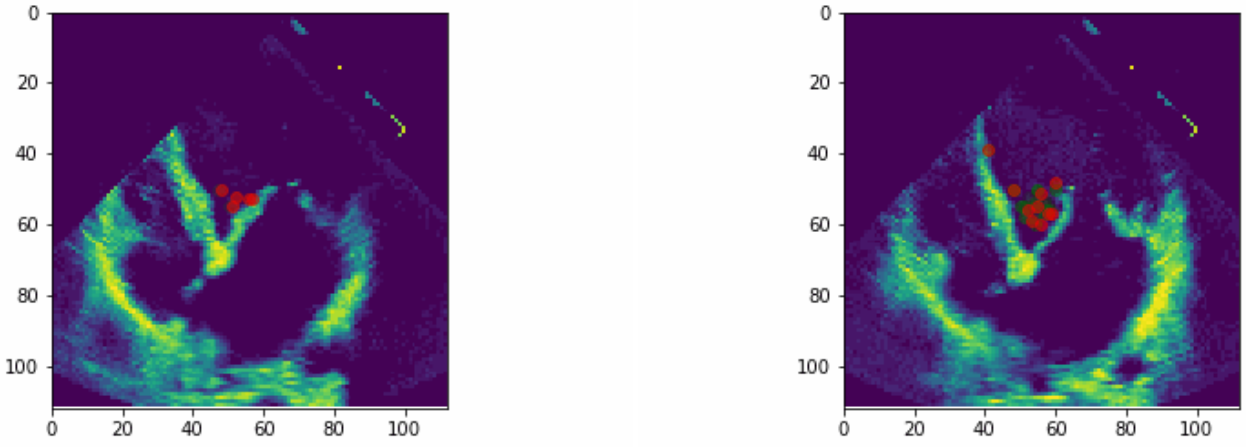

Figure 5: Two frames from a trajectory explanation for ejection fraction. Red points indicate current positions, green points indicate history