Exploring Explainability in Video Action Recognition

Avinab Saha*, Shashank Gupta*, Sravan Kumar Ankireddy*, Karl Chahine, Joydeep Ghosh
The University of Texas at Austin

Abstract

Image Classification and Video Action Recognition are perhaps the two most foundational tasks in computer vision. Consequently, explaining the inner workings of trained deep neural networks is of prime importance. While numerous efforts focus on explaining the decisions of trained deep neural networks in image classification, exploration in the domain of its temporal version, video action recognition, has been scant. In this work, we take a deeper look at this problem. We begin by revisiting Grad-CAM, one of the popular feature attribution methods for Image Classification, and its extension to Video Action Recognition tasks and examine the method's limitations. To address these, we introduce Video-TCAV, by building on TCAV for Image Classification tasks, which aims to quantify the importance of specific concepts in the decision-making process of Video Action Recognition models. As the scalable generation of concepts is still an open problem, we propose a machine-assisted approach to generate spatial and spatiotemporal concepts relevant to Video Action Recognition for testing Video-TCAV. We then establish the importance of temporally-varying concepts by demonstrating the superiority of dynamic spatiotemporal concepts over trivial spatial concepts. In conclusion, we introduce a framework for investigating hypotheses in action recognition and quantitatively testing them, thus advancing research in the explainability of deep neural networks used in video action recognition.

1. Introduction

Understanding human actions in videos is crucial for various applications like behavior analysis, video retrieval, and human-robot interaction. Human action understanding involves recognizing, localizing, and predicting human behaviors. The task to recognize human actions in a video is termed as *Video Action Recognition*. In recent years, significant research efforts have focused on developing effective models for this task, including I3D [6], SlowFast [8], and Video Swin Transformer [13]. Consequently, an intriguing avenue for further exploration lies in understanding the decision-making processes of these networks.

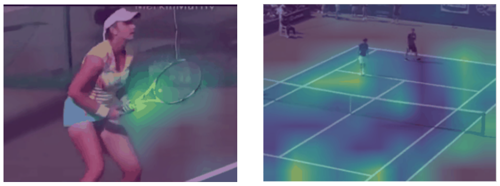
Recently, multiple works have focused on analyzing decisions made by neural networks in the context of image classification tasks by developing feature attribution methods: Integrated gradients [15], Class Activation Mapping (CAM) [18], and Grad-CAM [14]. A widely embraced alternative for studying the explainability of deep neural networks in image classification tasks over traditional feature attribution methods is TCAV [11]. TCAV is a global explanation method focusing on more abstract details instead of granular, pixel-level changes typically associated with feature attribution methods. Although there has been significant progress in post-training explainability methods for deep learning-based image classification, very little research has been done on the applicability of these methods in the context of video action recognition. In this work, we aim to investigate this direction by exploring a feature attribution method for video action recognition, discussing its limitations, and further introducing a video counterpart to TCAV, which we refer to as Video-TCAV. Specifically, we opt for *playing tennis* class from the Kinetics-400 dataset [10] to visualize the outcomes of the Grad-CAM as well as to formulate the proposed Video-TCAV framework.

The rest of the paper is organized as follows. Section 2 discusses the performance of Grad-CAM when employed in the context of video action recognition to set up a baseline representing popular feature attribution methods. In Section 3, we introduce Video-TCAV, which includes an automated pipeline for generating high-level concepts and are evaluated in the proposed Video-TCAV framework. Section 4 concludes the paper by summarizing the ideas explored in this paper and discussing future research directions.

2. Grad-CAM revisited

We first revisit Gradient-weighted Class Activation Mapping (Grad-CAM)[14], a popular feature attribution method renowned for its effectiveness in understanding image recognition models. Class Activation Maps (CAMs) visually represent the focus areas for an image recognition

[*] Equal Technical Contribution. Correspondence to Avinab Saha, Shashank Gupta and Sravan Kumar Ankireddy. Email: {avinab.saha,shashank.gupta,sravan.ankireddy}@utexas.edu



(a) Positive Illustration (b) Limitation

Figure 1. Grad-CAM outputs with respect to the class *playing tennis* for Video Swin Transformer model. (a): Grad-CAM correctly highlights the regions of movement for tennis rackets. (b): Grad-CAM focuses on the tennis court in the background and ignores the players in the frame.

model during prediction by utilizing the weights of the model's final convolutional layer. Building on this, Grad-CAM has emerged as a popular technique for generating visual explanations for various image recognition models by leveraging the gradient of the predicted class concerning the activations of the final neural network layer to produce the heatmap, thereby offering class-specific visualizations.

2.1. Extending Grad-CAM to Videos

A straightforward approach to adapting feature attribution methods from images to videos is to consider each frame as an individual image. However, this bears a significant limitation: it neglects the temporal interplay among frames, a crucial aspect for understanding the actions in videos.

This limitation can be overcome by utilizing a collection of frames as input, thus preserving the temporal connections. These frames are input to the Video Swin Transformer model [13], which was pre-trained on the Kinetics-400 dataset. Subsequently, gradients are computed with respect to the corresponding activation maps. Illustrative instances displaying Grad-CAM outputs for inputs related to the class *playing tennis* for the Swin Transformer model are presented in Figure 1a. The overlaid heatmap demonstrates that the outputs of the Grad-CAM are concentrated in the moving regions associated with the tennis racket.

2.2. Limitations in Feature Attribution Methods

Extending feature attribution methods designed for image recognition models to work well with videos is a non-trivial task and raises concerns regarding robustness. A significant bottleneck lies in processing temporal data in the correct semantic order. For example, the same set of frames played in reverse order should mean the output should be reversed for video action recognition *i.e.,* a person *picking up* an object and *placing down* an object would exhibit identical spatial data, but opposite temporal directionality. While coarse details suffice for overall action labeling in videos, they may not meet the requirements of explaining complex models such as the Video Swin Transformer. To ensure reliable and smooth label explanations across frames, a finer time domain analysis becomes necessary. Background elements further complicate Grad-CAM decisions, as seen in Figure 1b, where the focus shifts to the tennis court rather than the players. Further, Grad-CAM remains a local explanation method, implying that the analysis needs to be done individually for each attribution to draw any class-level conclusions. While this process is feasible for images, which can be viewed in batches, it becomes impractical for videos.

Beyond these domain-specific issues, attribution methods exhibit several other weaknesses. In [1], the authors demonstrated that networks with random weights generate similar attribution maps as trained networks. Additionally, attribution methods are vulnerable to adversarial attacks [9], changes in data preprocessing [12], and even the introduction of random noise [2]. Their widespread use poses particular concerns in high-risk applications such as medical imaging [3]. Due to these limitations in feature attribution methods, interest has grown in more meaningful methods tailored for human understanding, such as concept attribution, introduced in Testing with Concept Activation Vectors (TCAV) [11], that we extend to videos in Section 3.

3. Video TCAV

3.1. TCAV Basics

TCAV offers a popular alternative to feature attribution methods like Grad-CAM. Unlike granular, pixel-level changes, TCAV focuses on abstract details. It is widely accepted that the human brain converts raw pixel values into high-level concepts like texture, specific objects, and their interactions [7]. Similarly, neural networks encode analogous high-level concepts in the output of embedding layers. TCAV attempts to derive explanations by analyzing properties within these embedding spaces.

In TCAV, a concept is delineated by a collection of inputs that characterize it. For instance in image classification, the concept of *stripedness* can be delineated by a set of images portraying striped objects (ⓐ in Figure 2). When mapped to the corresponding feature space, the representations of this collection would ideally be aligned with a specific *stripedness* axis (ⓒ and ⓓ in Figure 2). On the other hand, a collection of random images would not display any such alignment. Thus, if we were to find a hyperplane that separates the embeddings of a set of concept-specific images and a set of random (or control) images, the normal to that hyperplane should give us a representation of the concept. This normal may also be considered a basis vector of the feature space associated with that concept (e.g., a degree of *stripedness*), and is called the Concept Activation Vector (CAV). Now, given any input for our specific task, say a picture of a zebra (ⓑ in Figure 2), the directional derivative of its embedding along the CAV would measure the sensitivity of the embedding to the concept. Figure 2 taken from [11] shows a graphical representation of the entire process discussed above. This scalar quantity may be aggregated differently for multiple inputs to get a more robust value. Moreover, since any learned hyperplane will return a CAV, even if it does not separate the data well, this process is usually repeated for multiple control sets to ensure the sensitivity is statistically significant. Thus, TCAV allows us to quantify the importance of any given concept to a specific step in the neural networks. It is especially useful for comparing concepts, which is hard with attribution maps.

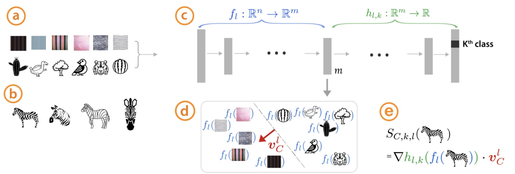

a
b
c
$f_l : \mathbb{R}^n \rightarrow \mathbb{R}^m$
$h_{l,k} : \mathbb{R}^m \rightarrow \mathbb{R}$
Kth class
m
d
$f_l($ $)$
v_C^l
e
$S_{C,k,l}($ $)$
$= \nabla h_{l,k}(f_l( )) \cdot v_C^l$

Figure 2. TCAV process in Image Classification. Image taken from [11]. Best viewed zoomed.

3.2. Video-TCAV Components

This section discusses the various components of the proposed Video-TCAV framework.

Video Action Recognition Model: Similar to TCAV, Video-TCAV is a post-training explanation method, requiring a pre-trained Video Action Recognition model. We opt for the state-of-the-art Video Swin Transformer model trained on the Kinetics-400 dataset for quantitative evaluation of post-training explanations. Figure 3 illustrates the schematic diagram of the Video Swin Transformer, including the three layers whose activations we utilize in testing CAVs in Video-TCAV.

Concepts: The most important component of Video-TCAV, similar to TCAV, is the concepts we want to test. In contrast to image classification, where selecting concepts is relatively simple, generating concepts for video classification poses a more intricate challenge. This complexity arises because choosing clips to represent concepts for Video-TCAV, similar to image crops in the case of TCAV, is not trivial. We generate two categories of concepts:

- *Spatial Concepts*: These are simply images of objects repeated temporally to form videos. This is done as the Video Swin Transformer can only take video input.
- *Spatiotemporal Concepts*: These are objects or humans in motion. These are obtained using YOLO-v7 [16] object detector frame-wise to track a concept of our choice.

3.3. Generating Video-TCAV Concepts

We adopt a machine-assisted method to generate spatial and spatiotemporal concepts for Video-TCAV. Utilizing YOLO-v7, we detect objects in videos and produce spatial and spatiotemporal crops for concept testing in Video-TCAV.

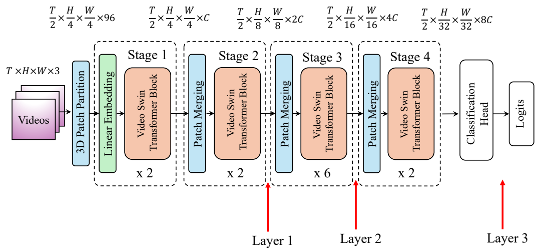

$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times96$
$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{T}{2}\times\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{T}{2}\times\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{T}{2}\times\frac{H}{32}\times\frac{W}{32}\times 8C$
$T\times H\times W\times 3$
Videos
3D Patch Partition
Stage 1
Linear Embedding
Video Swin Transformer Block
x 2
Stage 2
Patch Merging
Video Swin Transformer Block
x 2
Stage 3
Patch Merging
Video Swin Transformer Block
x 6
Stage 4
Patch Merging
Video Swin Transformer Block
x 2
Classification Head
Logits
Layer 1
Layer 2
Layer 3

Figure 3. Video Swin Transformer block diagram. The 3 layers with red arrows whose activations we study while testing CAVs are marked. Best viewed zoomed.

Given that the object detector is not flawless, we manually verify all concept video crops to ensure accuracy in our experiments. Following [11], we present our visualization and results for a single action class *playing tennis* for visualization and results demonstration. Figure 4 illustrates sample detections generated by YOLO-v7 on a video frame from the *playing tennis class*. Notably, YOLO-v7 provides accurate detections of a tennis racket, sports ball, and individuals, which we consider potential concepts aiding in explainability of videos belonging to *playing tennis* class.

Generating Spatial Concepts: We execute the YOLO-v7 object detector on all videos from concept generating split of the test set videos of the Kinetics-400 dataset labeled as *playing tennis'*. Since YOLO-v7 is trained on object detection with 80 different object classes, it leads to multiple object detections in each video frame. For our experiments, we considered all the per-frame detections from the three classes, i.e., person, tennis racket, and sports ball, as concepts that we want to test, and all the other detections are used to generate random concept sets. Figure 5 illustrates our generated spatial concepts. The generated spatial concepts are temporally repeated to generate a static video of 3-minute duration.

Generating Spatiotemporal Concepts: Contrary to spatial concepts, generating spatiotemporal concepts is non-trivial. The basic idea of generating spatiotemporal concepts is tracking a particular instance of an object across frames, like the person playing tennis, a tennis racket, or a sports ball across video frames. This is non-trivial as there might be multiple instances of the same object class in the frame, and YOLO-v7 results in multiple detections and it is necessary to choose the same instance of the object in consecutive frames to prevent the concept video from changing in terms of content. We ensured only one instance of the object was tracked across frames, and we took a spatiotemporal crop based on the detections returned by YOLO-v7. Since the size of the detections of a particular object changes across frames, we generate the concept video with the largest-sized object detected in consecutive frames and pad the other smaller-sized frames, positioning them at the center. Figure 6 illustrates our generated spatiotemporal concepts.

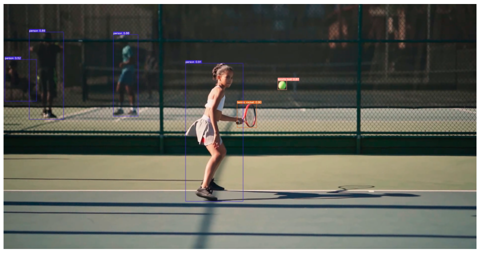


Figure 4. YOLO-v7 detections on a frame of a video from the *playing tennis* class. Best viewed zoomed.

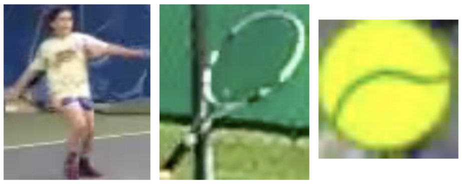


Figure 5. Exemplars of Spatial Concepts: person playing tennis, tennis racket, and sports ball generated from Kinetics-400 dataset.

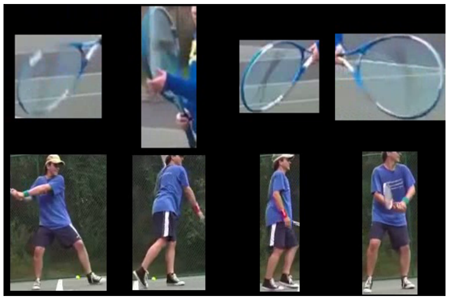


Figure 6. Exemplars of Spatiotemporal Concepts: movement of tennis racket in a match and person playing tennis. We show some sampled frames across each concept video.

3.4. Video-TCAV Experiments

Experimental Setup: We replicate the experiments of [11] to demonstrate the advantages of our pipeline. We collect 30 videos that have the ground truth label as *playing tennis* and generate 25-30 videos per concept. We use these videos and concepts in our Video-TCAV framework. For random concepts, we randomly select 30 videos from our entire corpus of concepts, including those not related to playing tennis e.g. dining and dancing. All experiments use the relative sign-count variant of TCAV. In this, we consider a group of multiple concepts at a time, and learn a one-vs-rest classifier for each concept, which gives us the CAV for that concept. The fraction of data points positively affected by the CAV is the relative TCAV for that concept. Using relative TCAVs helps disentangle the effect of correlation between different concepts. A more detailed discussion may be found in [11].

3.5. Results & Discussion

Spatial Concepts: The generated spatial concepts of static frames with no movement information serve as our baseline. Figure 7 shows the relative TCAVs of the tested static concepts in different layers. While all concepts show some relative importance, it is hard to distinguish their importance from a random concept, especially in the initial layers. This is in line with our expectations, as static concepts should provide some information the context of the video (e.g. the presence of a tennis racket indicates tennis being played) but should not be able to pinpoint the exact action performed. Another aspect to note is that the importance of the random concept persists until the last layer, which validates the static features are not fully discriminative for the action recognition task.

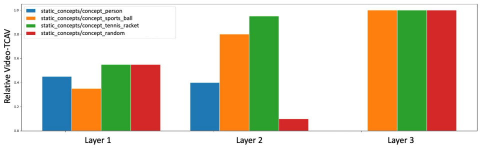


Figure 7. Relative TCAVs for static concepts.

Spatio-Temporal Concepts: The dynamic spatio-temporal concepts, on the other hand, have more predictive power than the static concepts. Figure 8 shows the relative TCAVs of the tested dynamic concepts in different layers. In the initial layers, the importance of the dynamic concepts is more distinguishable from random, but is overall similar to static case, which is in line with the understanding that initial layers extract general-purpose features. However, in later layers, the importance of temporal concepts is dramatically more pronounced. In contrast with the static concepts, the random concept holds negligible importance in the last layer. Interestingly, this increase of importance of temporal information with depth correlates well with how the human brain perceives motion, with visual features first processed by the V1 area, then motion being handled later in the V5/MT area [17].

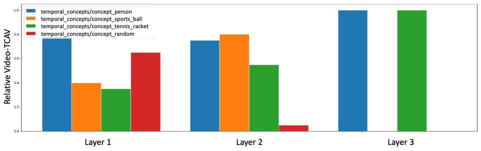


Figure 8. Relative TCAVs for dynamic concepts

Spatial vs Spatio-Temporal Concepts: Next, we directly compare the spatial and spatio-temporal versions of a specific concept. Figure 9 shows their relative TCAVs in different layers. Evidently, when compared side-by-side, the temporal concepts dominate over static concepts in most layers of the network, with this effect becoming more pronounced with depth. It is particularly interesting to note how the final layer exclusively prefers the temporal concept for prediction compared to the static version of the concept.

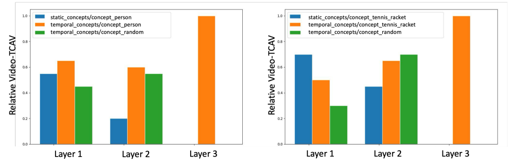


Figure 9. Relative TCAVs for static vs dynamic concepts

Statistical Testing: We tested the validity of the scores with 10 other random sets and applied a two-sided t-test with Bonferroni correction, as in [11]. The results are in the Figure 10. The $p < 0.05$ for our concepts implies that the CAVs we found are robust and statistically significant, where as the CAVs learnt by separating random concepts are not meaningful ($p > 0.05$).

4. Conclusion and Future Work

In this work, we proposed Video-TCAV, a new approach for generating human-friendly concepts and quantitatively measuring their impact on the decision process of deep networks used for action recognition. Further, we provide preliminary results on the Video Swin Transformer to demonstrate the success and pitfalls of our method. However, more experiments using other video action recognition models would strengthen the effectiveness of Video-TCAV. Additionally, Video-TCAVs may also share some of the weaknesses of attribution methods, such as [5], which demonstrates adversarial attacks on TCAV that can artificially raise or lower the importance of a particular concept, leading to bizarre conclusions. Generating concepts for Video-TCAV is challenging. In this work, we proposed a method using YOLO-v7 object detector. An alternative research direction could involve utilizing text-to-video diffusion models, as in [4], to generate concept clips based on human-interpretable text prompts. We hope that this work serves as a starting direction for developing scalable explainable methods for Video Action Recognition.

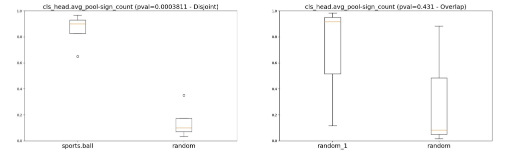


Figure 10. Results of hypothesis testing with spatiotemporal concepts. Best viewed zoomed.

References

[1] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307*, 2018. 2

[2] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018. 2

[3] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6), 2021. 2

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 5

[5] Davis Brown and Henry Kvinge. Brittle interpretations: The vulnerability of TCAV and other concept-based explainability tools to adversarial attack. *CoRR*, abs/2110.07120, 2021. 5

[6] 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 (CVPR)*, 2017. 1

[7] Russell A. Epstein and Chris I. Baker. Scene perception in the human brain. *Annual Review of Vision Science*, 5(1): 373–397, 2019. 2

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[9] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3681–3688, 2019. 2

[10] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 1

[11] 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. 1, 2, 3, 4, 5

[12] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019. 2

[13] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, 2021. 1, 2

[14] 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. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[15] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1

[16] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022. 3

[17] J. D. G. Watson, R. Myers, R. S. J. Frackowiak, J. V. Hajnal, R. P. Woods, J. C. Mazziotta, S. Shipp, and S. Zeki. Area V5 of the Human Brain: Evidence from a Combined Study Using Positron Emission Tomography and Magnetic Resonance Imaging. *Cerebral Cortex*, 3(2):79–94, 1993. 4

[18] 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. 1