Generating Conceptual Explanations for DL based ECG Classification Model

Suparshva Jain, Amit Sangroya, Lovekesh Vig, C. Anantaram, Arijit Ukil, Sundeep Khandelwal
TCS Research, Tata Consultancy Ltd.
{suparshva.jain, amit.sangroya, lovekesh.vig, c.anantaram, arijit.ukil, sundeep.khandelwal}@tcs.com

Abstract
Deep learning techniques are being used for heart rhythm classification from ECG waveforms. Large networks using end-to-end learning such as convolutional neural networks are not easily interpretable by end-users such as doctors. This is because most of the state of art explainability techniques focus on explanations for data-scientists who have technical knowledge. However, end-users of such systems are normally doctors who are not familiar with the technical details of machine learning. Therefore, to address this gap, we propose a framework that provides explanation of ECG classification model in the language of clinicians. We leverage state of art knowledge-based systems to map domain concepts with outcomes of machine learning model. Our results show that domain concepts based explanations are useful for clinicians and can greatly reduce their cognitive load. This shall lead to larger deployment of these models in real world.

Introduction
Cardiovascular diseases are the leading cause of death worldwide and the electrocardiogram (ECG) is a major tool in their diagnoses. Automated computer analysis of standard 12-lead electrocardiograms gained importance in the process of medical diagnosis. Recent work based on DL (deep learning) techniques substantially improve the accuracy and scalability of automated ECG analysis (Rajpurkar et al. 2017). These findings surely demonstrate that an end-to-end deep learning approach can classify a broad range of distinct arrhythmias with high diagnostic performance similar to that of cardiologists. However, explainability of such complex deep networks is a practical limitation in real-world deployment of systems based on these models. Explainability, fairness and accountability are key pillars for the large-scale implementation of AI methods in medical domain.

eXplainable AI (XAI) is a relatively new field that tries to fix the barrier of explainability, an inherent problem of AI techniques brought by sub-symbolism (e.g. ensembles or Deep Neural Networks) (Arrieta et al. 2019). Apart from post-hoc visualization techniques (Wagner et al. 2019) (e.g., feature dependency plots), feature importance techniques based on sensitivity analysis, there have been three main approaches for explainability of AI systems: i) Proxy or Shadow model approaches like LIME, SHAP, ii) Model inspection approaches like Class Activation maps (CAM), Grad-CAM, Smooth-Grad-CAM, etc. and iii) Data based approaches like Decision sets and Formal Concept Analysis (Lakkaraju, Bach, and Leskovec 2016; Sangroya et al. 2019; Petsiuk et al. 2020; Pfau et al. 2019; Iyer et al. 2018). Most of the research work on explainability has followed one of the above approaches (Guidotti et al. 2018). However, each of these approaches have limitations in the way the explanations are generated.

One common issue among all these XAI techniques is that the explanations generated are not very informative for end-users and domain-experts. The explanations generated are more suitable for use by data science team (typically involving data engineers, software developers, machine learning engineers etc.) to improve the accuracy of machine learning model and/or to get insights into the machine learning model. However, the explanation generated do not correlate with domain concepts, and thus, from an end-user perspective, does not assist them in getting a better understanding of machine learning model outcomes. Medical domain is even more critical and challenging, as the cost of a wrong interpretation of model decision can be very huge. Figure 1 shows comparison of example domain concept based explanation with one popular XAI technique known as SHAP. It is clearly visible that a clinician would be more comfortable in choosing explanation based on domain concepts.

Figure 1: Explanation example using a SHAP framework vs. our approach based upon domain concepts
learning, data science, or programming, making some explanation methods which presume users’ prior knowledge in AI (such as gradients, activations, neurons, layers) unviable. 2) Diverse users, tasks, and explainability needs: developing XAI for end-users must adapt to the variability in the end-users’ roles, tasks, and needs for explanation. For example, a doctor may demand distinct explanations from AI when using it as a diagnostic support system, whereas a human resources specialist resorts to explanations to support her hiring decisions (different end-users and tasks).

The main contribution of this paper is an explainability approach that leverage the strengths of a symbolic rule based expert system to explain a complex deep learning model. We demonstrate our approach by providing end-user oriented explanations for a well known deep learning ECG classification model on real world datasets. The scope of this paper is explaining Atrial Fibrillation detection of deep learning model. This can be surely extended to multiple arrhythmia classes. Apart from various quantitative metrics, we get a qualitative evaluation involving a senior practicing cardiologist. Our results show that the explanations are surely end-user oriented and can explain a complex DL model using simple to understand domain concepts.

Background and Preliminaries

In this section, we present some background on deep model for ECG classification and KARDIO expert system (Lavrac et al. 2007) for the electrocardiographic diagnosis of heart arrhythmias.

Deep Model for ECG classification

Many machine learning based techniques have been designed for Arrhythmia classification given ECG data (Ebrahimi et al. 2020). More recent works using deep learning (DL) based techniques greatly improve the classification accuracy. A DL based model from Stanford (Rajpurkar et al. 2017) uses a 34-layer convolutional neural network (CNN) to detect arrhythmias in arbitrary length ECG time-series. Interpreting the outcome of such a complex network is very challenging especially for non-technical users who are not familiar with machine learning. DL model may confuse between various classes e.g. as they report in their paper that common confusions is between Ectopic Atrial Rhythm (EAR) and the sinus rhythm (SINUS). The main distinguishing criteria for this rhythm is an irregular P wave, which is an important and meaningful explanation for a clinician. However, none of state of art XAI techniques can provide such explanations. Conceptual explanations based on domain knowledge is the only possible solution in such scenarios.

KARDIO

Kardio is a medical expert system based upon model of the human heart designed for the diagnosis of cardiac arrhythmias (Bratko, Mozetič, and Lavrač 1990) (Lavrac et al. 2007). Kardio’s performance is estimated by cardiologists to be equivalent to that of a specialist of internal medicine (not a cardiologist) who is highly skilled in the reading of ECG recordings, and it can be used as a diagnostic tool in ECG interpretation. It may also be used for instruction in electrocardiography. The Kardio model was compiled, by means of qualitative simulation and machine learning tools, into various representations that are suited for particular expert tasks. Authors of Kardio investigate a hierarchical organization of a qualitative model and outline an experiment whereby the construction of a deep model is automated by means of machine learning techniques.

The aim of Kardio is to diagnose cardiac arrhythmias from ECG descriptions. To do so, it looks for rules that describe all possible cardiac arrhythmias (single or multiple) corresponding to a given symbolic description of an ECG. Rule learning relies on a qualitative model of the heart that simulates the cardiac electrical activity: over 2,400 heart disorders can be related to over 140,000 ECG descriptions. One interesting aspect about the Kardio expert system is that the explanations of arrhythmias generated by the system are based on domain concept features such as $P$ waves absent, Irregular Rhythm etc. This explanation is highly understandable by clinicians. The decision rules that drive these explanations are not very complex. For example, one of the explanation for Atrial Fibrillation class is $P$ waves absent, Irregular Rhythm, Heart Rate is 120.

Symbolic Description of an ECG

The input to Kardio is comprised of 10 features, which are also characteristic diagnostic features that cardiologists look at in their process of diagnosis. The deep model comprises of 3 impulse generators and 4 conduction pathways. The output comprises of 7 terms: 1) Are the three impulse generators functioning properly. If not, what conditions causes them to fail 2) Similarly are the 4 conduction pathways functioning properly. (see listing 1 and 2) The symbolic description helps to explain the ECG signal using clinical terms and structural causal model of heart.

Listing 2: The 7 output terms in Kardio

domains_arr( arr(SA, Atr, AV, Jun, BB, Reg, Ect) ) :-
domain(SA, [quiet, sa, sad, sb, sr, st] ),
domain(Atr, [aeb, af, afl, at, mat, quiet, wp] ),
domain(AV, [avbl, avb3, lgl, mob2, n, wen, wpw]),
domain(Jun, [jb, jeb, jr, jt, quiet] ),
domain(BB, [llbb, normal, rbbb] ),
domain(Reg, [avr, quiet, vf, vfl, vr, vt] ),
domain(Ect, [quiet, veb] ).

Thus, we have two models at our hand: one is the deep model that is highly accurate but complex and not easily interpretable; and the second model based on Kardio which is easy to understand but not scalable for large deployments. The major challenge is how do we build an explanation pipeline that can take the DL model outcome and provide Kardio level explanations for ECG classifications. In our work, we are building this pipeline that is able to generate conceptual explanations for DL based ECG classifications.
Listing 1: The 10 features that make up the input to Kardio

domains_ecg(
  ecg(Rhythm, P, RateP, P_QRS, PR, QRS, Rate, EctP, EctPR, EctQRS)) :-
  domain( Rhythm, [irregular, regular] ),
  domain( P, [abnormal, absent, changing, normal] ),
  domain( RateP, [between_100_250, between_250_350, between_60_100, over_350, under_60, zero] ),
  domain( P_QRS, [after_P_always_QRS, after_P_some_QRS_miss, independent_P_QRS, meaningless] ),
  domain( PR, [after_QRS_is_P, changing, meaningless, normal, prolonged, shortened] ),
  domain( QRS, [absent, delta_LBBB, delta_RBBB, normal, wide_LBBB, wide_LBBB_RBBB, wide_RBBB] ),
  domain( Rate, [between_100_250, between_250_350, between_60_100, over_350, under_60] ),
  domain( EctP, [abnormal, absent] ),
  domain( EctPR, [after_QRS_is_P, meaningless, normal, prolonged, shortened] ),
  domain( EctQRS, [absent, delta_LBBB, delta_RBBB, normal, wide_LBBB, wide_LBBB_RBBB, wide_RBBB] ).

Figure 2: Example of improving P-wave Feature Extraction using average of ECG segment

ECEC Framework: Explanations in Clinician’s Language for ECG Classification

The Stanford cardiologist-level arrhythmia detector (Rajpurkar et al. 2017), is a deep learning based model that was shown to perform better than an average cardiologist in various ECG classifications such as Atrial Fibrillation (AFib), Sinus Tachycardia, Sinus Bradycardia or Normal etc. Given that the aim of such DL models is to build ECG classification mechanisms such that a general physician is able to use that classification to decide on the next course of treatment for a patient, it is important to generate conceptual level explanations to show how the DL model arrived at a particular classification.

In our work, we focus on this aspect - generating conceptual level explanations for Arrhythmia classifications generated by the Stanford DL model. Our model takes the classification generated by the Stanford DL model for a given input raw ECG signal as the classification to explain. We use the Neurokit-Kardio pipeline as a shadow model for the Stanford DL model to generate such conceptual explanations of the classifications made by the DL model. We describe our steps below.

Domain Features Extraction

The first step is to pass the input ECG signal to the DL model and get its classification. Once this is done, we use the Neurokit library, which is a well known tool for ECG signals (Makowski et al. 2021), for extracting logical features from the input ECG signal. This gives the basic segmentation of the input ECG signal into P-wave, QRS and T-wave segments. However, there are some issues related to noise and changing P-wave morphology. In order to handle this problem, we improve the output of feature extraction using domain knowledge. We extract out the morphology of the signals in terms of P-waves and RR intervals.

We know that in a normal signal the P-wave morphology (i.e. the shape of the P-wave) remains the same throughout the signal, while in an AF signal this may change. Hence, we isolate each heart beat in a signal, synchronise each heart beat around the R-peak and then take an average over each heart beat in a signal (See Figure 2). This averaged out heart beat can be used to detect the absence or presence of P-waves. The heart beat just before the QRS complex is compared against bell curve using DTW (Dynamic Time Warping) and if this distance is relatively large, we can say that the P-wave is absent and if this distance is small, then we can say the P-waves are present.

Other features like heart rate, relationship between P wave and QRS Complex, Rate of P-waves, PR interval can be extracted using simple logic after ECG delineation using Neurokit library.

Feature Softening

The domain features extraction process in the previous step has to contend with the hard thresholds and this leads to issues related to accuracy and makes it more sensitive. To handle this problem, we use a feature softening approach. Rather than having a binary value (yes/no) for the presence of each feature, we get a probabilistic value. Using feature softening for each logical feature, we assign probability val-
ues for each feature which is used to determine if a particular feature is present in the ECG signal or not. The final score in Table 1 is calculated based on multiplying all the probability values of each logical feature in the ECG signal.

Our feature softening procedure is different for each feature. For example, for P-wave feature softening, we use a DTW (Dynamic Time Warping) distance score, this score is compared against a normal distribution to get a probability value. The three possible values for P-wave features are: Absent, Normal, and Abnormal. Through feature softening process, we optimize the selection of one of three possible values for P-wave. Similarly, for Rhythm feature, we use standard deviation in R-R intervals within an ECG signal to see if the rhythm is regular or not.

**Backtracing DL Classification**

The next step we do is to generate a decision table for all possible classifications that can be generated from the Kardio model. Once that is done, for a given DL classification we backtrace that classification from the decision table to find all probable conceptual feature combinations that could lead to such a classification with softening of the conceptual features extracted out of backtracing from Kardio model. An ECG Signal is first passed to the DL Model to get the classification output. Simultaneously we calculate score for each possible entry in the decision table. Then we use the DL output to select the most likely row of this decision table as the explanation.

Kardio Model has 6 categorical features and based on all possible configuration of Kardio Model we created a decision table for selecting the best possible explanation for a given ECG Signal (Lavrac et al. 2007). Since an AFib classification depends on “irregularly irregular RR intervals” and on the “absence of P-waves”, we match the conceptual features with the domain features extracted from the raw ECG signal with a softening of the features. The probability of match is computed for such a classification.

**Generating Explanations using Domain Concepts**

Explanations are generated using the probability of match of the conceptual features with the classification match on the Kardio decision table. The explanation is ranked according to the probability values as the most likely explanations. For example, for the ECG signal belonging to Atrial Fibrillation class, we generate the explanation in terms of domain concepts such as (Rhythm is Irregular, P-wave is Absent, Role of P-wave is Zero, Relation between P and QRS is meaningless, PR interval is meaningless). Such a domain concept based explanation is useful for a clinician and easily understood rather than the traditional explanations of deep models such as gradient of the signal, perturbations etc.

**Experimental Evaluation**

**Experimental Setup**

We perform our experiments on Physionet-2017 dataset. We use Stanford DL model as reference model for getting the ECG signal classification due to the fact that this is model is widely acknowledged as highly accurate model. For training the DL model we used a 80:10:10 split for training, cross validation and testing respectively. In order to evaluate our approach, we use various metrics focusing on: 1) How accurate is the explainability pipeline? 2) Subjective evaluation: percentage of explanations where cardiologist agrees, and 3) Fidelity of explanations: How consistent are the explanations?

**Experimental Results**

As shown in Table 2, we achieve significant accuracy using explainability pipeline for two classes: Atrial Fibrillation and Normal. In order to generate highly accurate explanations, it is important that underlying pipeline is also accurate. It is practically impossible to achieve similar accuracy equivalent to a DL model (See Table 3), but it is very important that underlying pipeline also possess a good amount of accuracy. Precision, recall and F1 scores clearly show that our pipeline is also showing a significant accuracy.

We performed calibration of the DL model and choose only those instances where DL model was highly confident. We employ temperature scaling using netcal open source library (Mozafari et al. 2018; Küppers et al. 2020). The expected calibration error (ECE) before and after calibration was 0.059 and 0.002 respectively. This ensured that only high confidence and accurate samples where taken in our pipeline for explanations. We then did human evaluation of the explanations generated by our system wherein an experienced cardiologist manually validated the explanations generated by our pipeline (See Table 4). We choose 78 samples randomly from our test set and involved the cardiologist to check if he agrees with the explanations or not. Cardiologist marked 91% explanations correct and meaningful. We also got those examples validated where DL model was predicting incorrectly and used the instances where cardiologist has a disagreement to improve our pipeline.

**Comparing the explanations with SHAP output**

SHAP is a popular technique for interpreting the results of a deep neural network (Lundberg and Lee 2017), hence we have here tried to correlate the SHAP values of the DL model with the explanations generated by our pipeline. If we analyse the distribution of the SHAP values in the figure 4, we see that the SHAP values have a very small standard deviation which makes it much harder to interpret, as to which of the regions in an ECG signal appear to be important. Further, SHAP values cannot indicate what exactly is wrong with the signal, making it much harder for the end user to understand the model’s prediction. We have found that cardiologists find it much easier to understand the output from our explainability framework, since our pipeline makes use of domain concepts which can be interpreted easily by a medical expert. Also, our pipeline is able to point out what exactly is abnormal in an ECG signal.

**Related Work**

Various approaches for explainability of black-box models have been proposed (Guidotti et al. 2018). Broadly the
| S.No. | Rhythm  | P wave  | Rate_P | P_QRS   | PR Interval  | Heart Rate | Class | Score |
|-------|---------|---------|--------|---------|-------------|------------|-------|-------|
| 0     | Irregular| Absent  | zero   | Meaningless | Meaningless | between 60 and 100 | AF* | 0.28  |
| 1     | Irregular| Absent  | under 60 | Meaningless | Meaningless | under 60 | AF   | 0.07  |
| 2     | Irregular| Absent  | zero   | Meaningless | Meaningless | 100       | AF*  | 0.001 |
| 3     | Irregular| Normal  | between 60 and 100 | Meaningless | Meaningless | between 60 and 100 | AF* | 0.0006 |

Table 1: Score Calculation for Decision Table

| Arrhythmia class  | Precision | Recall | F1   |
|-------------------|-----------|--------|------|
| Atrial Fibrillation (AF) | 0.73     | 0.72   | 0.72 |
| Normal             | 1.0       | 0.59   | 0.74 |

Table 2: Classification Performance of ECEC Framework

| Arrhythmia class  | AF | Normal |
|-------------------|----|--------|
| Agree with Classification | 30 | 45     |
| Disagree with Classification | 0  | 3      |
| Agree with Explanation  | 30 | 41     |
| Disagree with Explanation | 0  | 7      |

Table 4: Human evaluation results when a senior cardiologist validated the explanations

In contrast to model explanation approaches such as LIME and SHAP (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017), our approach focus on end-users so that the explanations can be generated in their language. To the best of our knowledge, there is no work that directly focus on explainability for end-users more specifically for classification problems in ECG domain. One interesting aspect of our work is integration of a symbolic system with a deep learning model. The pipeline combines the interpretability powers of symbolic system with functional strengths of deep learning model. This enables our architecture to be easily customize to many domains other than ECG classification.

Conclusions and Future Work

State of art XAI frameworks such as LIME, SHAP, Grad-CAM etc. provide explanations for machine learning models, however often these explanations are not easy to comprehend for non-technical end users. This paper tries to address the limitations of XAI techniques for wider acceptability by end-users such as doctors. We focus on a very critical problem of classifying Atrial Fibrillation detection using ECG signal. The proposed explainability approach makes it easier for clinicians to interpret DL model’s outcomes. This leads to faster deployment of such models in real world settings. Our results shows that explanations are consistent and
usable for doctors. In future, we would like to extend this work to other Arrhythmia classes and also richer explanations based on causal and counterfactual techniques. Incorporating clinician’s feedback for improving explainability pipeline is also worth attempting.

References
Anders, R.; Diederich, J.; and Tickle, A. 1995. Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-Based Systems 6:373–389.
Arrieta, A. B.; Díaz-Rodríguez, N.; Ser, J. D.; Bennettot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2019. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai.
Bender, A.; Montavon, G.; Bach, S.; Müller, K.-R.; and Samek, W. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers.
Bratko, I.; Mozetič, I.; and Lavrač, N. 1990. KARDIO: A Study in Deep and Qualitative Knowledge for Expert Systems. Cambridge, MA, USA: MIT Press.
Chen, J.; Song, L.; Wainwright, M. J.; and Jordan, M. I. 2018. Learning to explain: An information-theoretic perspective on model interpretation.
Ebrahimi, Z.; Loni, M.; Daneshtalab, M.; and Gharabaghi, A. 2020. A review on deep learning methods for eeg arrhythmia classification. Expert Systems with Applications: X 7:100033.
Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Gianotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51(5).
Iyer, R.; Li, Y.; Li, H.; Lewis, M.; Sundar, R.; and Sycara, K. 2018. Transparency and explanation in deep reinforcement learning neural networks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES ’18, 144–150. New York, NY, USA: Association for Computing Machinery.
Küppers, F.; Kronenberger, J.; Shantia, A.; and Haselhoff, A. 2020. Multivariate confidence calibration for object detection. In The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.
Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery.
Lavrac, N.; Bratko, I.; Mozetič, I.; Čerček, B.; Grad, A.; and Horvat, M. 2007. Kardio-e-an expert system for electrocardiographic diagnosis of cardiac arrhythmias. Expert Systems 2:46 – 55.
Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., Advances in Neural Information Processing Systems 30. Curran Associates, Inc. 4765–4774.
Makowski, D.; Pham, T.; Lau, Z. J.; Brammer, J. C.; Lespinasse, F.; Pham, H.; Schölzel, C.; and Chen, S. H. A. 2021. NeuroKit2: A python toolbox for neurophysiological signal processing. Behavior Research Methods 53(4):1689–1696.
Mozafari, A. S.; Gomes, H. S.; Janny, S.; and Gagné, C. 2018. A new loss function for temperature scaling to have better calibrated deep networks. CoRR abs/1810.11586.
Omeiza, D.; Speakman, S.; Cintas, C.; and Weldermaram, K. 2019. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models.
Petsiuk, V.; Jain, R.; Manjunatha, V.; Morariu, V. I.; Mehra, A.; Ordonez, V.; and Saenko, K. 2020. Black-box explanation of object detectors via saliency maps.
Pfau, J.; Young, A. T.; Wei, M. L.; and Keiser, M. J. 2019. Global saliency: Aggregating saliency maps to assess dataset artefact bias.
Rajpurkar, P.; Hannun, A. Y.; Haghpanahi, M.; Bourn, C.; and Ng, A. Y. 2017. Cardiologist-level arrhythmia detection with convolutional neural networks.
Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. “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, San Francisco, CA, USA, August 13-17, 2016, 1135–1144.
Sangroya, A.; Anantaram, C.; Rawat, M.; and Rastogi, M. 2019. Using formal concept analysis to explain black box deep learning classification models. In Proceedings of the 7th International Workshop “What can FCA do for Artificial Intelligence”? co-located with IJCAI 2019, Macao, China.
Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In The IEEE International Conference on Computer Vision (ICCV).
Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences.
Wagner, J.; Kohler, J. M.; Gindele, T.; Hetzel, L.; Wiedemer, J. T.; and Behnke, S. 2019. Interpretable and fine-grained visual explanations for convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
Yoon, J.; Jordon, J.; and van der Schaar, M. 2019. INVASE: instance-wise variable selection using neural networks. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.
Zhou, B.; Khosla, A.; A.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. CVPR.