Naval Research Laboratory
Washington, DC 20375-5320

Final Report: Explainable Deep Networks

WALLACE "ED" LAWSON
Email: ed.lawson@nrl.navy.mil

Navy Center for Applied Research in AI (NCARAI)
Information Technology Division

Keywords: Deep Learning, Explainability, Interpretability, Surveillance, Pedestrian Attributes, Interactive Learning, Few-Shot Learning, Meta-Learning, Anomaly Detection, Sense and Sense Making, Understanding

August 30, 2021

This **NCARAI Technical Note** is the final project report for the NRL Base Program project reducing the burden of massive training data for deep learning.

Introduction

As recently as December of 2016, terrorist groups such as ISIS have called for an increase in attacks against US Naval Bases. To prevent such incidents, video surveillance cameras can be used to monitor critical areas for potential threats. However, monitoring video surveillance cameras can be a challenging task. A typical setup employs many cameras which all produce a significant amount of information. The manpower required to monitor these cameras is tremendous, which can then lead to a reliance on automated video analytics. At issue is the effectiveness of these analytics in communicating important information to the warfighter. State-of-the-art approaches can locate potential threats, but are unable to explain why something might be a threat. The warfighter needs actionable intelligence: explanations are important as they can provide this important information.

In this project, we explored several ways of providing such explanations. The first form of explanation comes from descriptions built using a domain expert-centric knowledge. In our approach, we organize this information using an ontology. We train a deep network to produce these descriptions on a person. This then provides the ability to filter information to look for something of interest or to look for things that are out of the ordinary.

We have also explored methods to search for an object of interest. In our proposed approach, when an object is detected it provides both the classification as well as the keypoints that were used in making this classification. Few-shot learning can learn to recognize both object and provide these keypoints using only a few labeled training instances. We have found this approach to be useful in cases when descriptions are not appropriate.

Technical Objective

The objective of this project is to create explainable deep networks within the context of automated surveillance and to demonstrate how this approach can be used to locate and explain predictions.

Technical Approach

We focus on two aspects of automated surveillance: detecting objects of interest and finding things that are out of the ordinary. In both tasks, we provide a user understandable explanation behind the classification suggested by the deep network.

Finding things that are out of the ordinary relies on verbal desciptions, or explanatory factors (XF). An XF is a list of attributes that explain how observations (e.g., people and objects) differ from others. They are conceptually similar to latent semantic codes (LSCs) [4], but learned in a different way.

An issue with building XF from surveillance data is the vast quantity of information, much of which is redundant. Labeling all of this information is not only tedious, it's also unnecessary. There are a great number of labeled datasets each of which cover a different aspect of this problem. One dataset will explore identity, another labels actions, clothing labels may be spread across 3 or 4 datasets that label things at slightly different levels of granularity.

We propose a novel way to leverage all of this labeled data in order to build a deep convolutional neural network with a wide range of explanatory factors related to people. This information is organized into a

hierarchical ontology, broadly grouped into the affected regions of the person (e.g., whole body vs. upper body). We call this ontology the Hierarchical Ontology for Personal Attributes Recognition (HOPAR).

During training, the attributes from each dataset are mapped from their original labels onto the HOPAR framework. When learning, we analyze each data source using only those labeled attributes; attributes not included in each dataset are marked with special values to indicate that their truth is unknown. Using this approach, we have combined several well known person attributes recognition datasets (PETA, PA-100K, and RAP) to learn over 148 attributes describing a person and their appearance.

In some cases, we wish to look for a specific object of interest. XF are not as useful, and it can be much easier to provide an image of the object instead. We provide interpretability here in the form of keypoints, which indicates regions of the object that should be considered to be disctinctive. This idea represents a signficiant departure from common approaches that learn using the whole image without providing any ability to control which aspect of the object is important. We then recognize the object using a few-shot learning approach known as meta-learning which provides a technique to learn from only a few labeled images. When keypoints are not provided, our framework relies on a separate classifier [6] that labels the visual saliency of each pixel in the original input image. We cluster the saliency map to locate keypoints in the original image. We then crop a small region around each of these salient keypoints, which is then resized to a canonical resolution (84 × 84) and given to the meta-learner. Multiple keypoints are combined using a long-short-term memory (LSTM). Although this concept may be applied to a number of different meta-learning approaches, in this paper we apply this to prototypical learning [5] .

Technical Progress

Explainable Factors

In much of computer vision, recognizing visual and semantic information has been generally considered to be a label-learning problem. Labels represent concepts in which do not exist in isolation. There are inherent relationships between concepts. Current approaches generally do not focus on learning these relationships. To explore this, we proposed a hierarchical ontology-based approach to learning attributes about people and objects. Our ontology was hand-coded using structural guidelines to capture hierarhical relationships such as parts-to-whole and category-to-instance. For people, inherent attributes (e.g., gender,age) are closer to the root, whereas extrinsic attributes (e.g., clothing) are deeper in the ontology. Leaf nodes are grouped to capture mutual-exclusion vs. independence, and other relationships between features.

We demonstrated this idea using an ontology of attributes to describe person. We called our person attributes ontology the hierarchical ontology of personal attributes recognition (HOPAR). HOPAR learns to recognize 148 personal attributes using labels from several computer vision datasets (PETA[1], PA100K[2] and RAP[3]). As is the case with many militarily relevant datasets, labeling of attributes is a challenge. Each dataset will only label a percentage of these attributes. For those that are unlabeled with modify our learning rule to only consider those labels that are used, ignoring all others. Our work on pedestrian attributes recognition [7] demonstrated that deeper convolutional neural networks recognize these attributes more reliably and that they tend to use related attributes to boost accuracy.

One of the important conclusions from this project is how explainable factors can be used, and when they may be a distraction. We demonstrated how the attributes can be used to search using descriptive terms

or how attributes can be used to explain why a classification was made. If given a significant amount of surveillance data, for example, search terms can be used to sort through this data to find potential matches.

In some cases when it may be difficult to exactly determine the object type, human operators found explainable factors to be a distraction in their own decision making. Thus, we believe that XF can be limited to explaining things that are unusual and to filter through data.

Interpretable Object Detection

Next we discuss how we can search for an object of interest. In many applications, an extensive amount of training data is either unavailable or difficult to collect. We propose a novel approach where we enable an annotator to control this object detection process by selecting important aspects of the image. As images often contain irrelevant details, background clutter, and small or subtle differences this can be a great benefit. In our recent work [10], we propose a method where we use few-shot object recognition in combination with these keypoints. The key idea behind our proposal was rather than to process the image as a whole, instead we process small regions around each of the keypoints. Keypoints can be selected either manually by an annotator or automatically using an estimation of human saliency.

Our technique can be applied to a wide variety of problems, one such case study was performed on teaching a robot to recognize tools for shipboard maintenance. A separate work [8] handled the problem of how to direct the attention of the robot to the object of interest. This case study addressed the problem of teaching an autonomous robot to learn new things. In our scenario a human instructor presents an image of an object with a good viewpoint, pose, etc. The robot is, however, expected to perform equally well when recognizing the object with different poses, backgrounds, lighting, etc. Rather than using human selected keypoints, instead we relied on saliency to manually select keypoints.

The robot is also expected to both learn and recognize objects in a reasonable amount of time. Both learning and recognition must be performed fast enough to be tolerable to the human instructor. This should be accomplished within a time period that we refer to as interaction time[8], which can be a short time window, but must be sufficiently fast enough to permit interaction.

We consider a scenario where an instructor teaches a robot to recognize handheld tools: screwdriver, hammer, pliers, and cutters. In our case study, there are 2 variants of each tool. Each of the tools are functionally the same and have the same general shape, but may vary in the color of the handle, the exact size, etc. Additionally, each is presented in front of the human instructor, without any contextual cues that may be used to help recognize the object. The human instructor intentionally varies the pose of the tool so that we can evaluate performance of different poses.

We evaluate performance both with and without keypoints. We collected images over three training sessions where the human instructor shows the tools to the robot, while intentionally varying the orientation of the tool. The three training sessions were conducted over different days. Using keypoints, we are able to recognize objects 69.7% of the time correctly.

References and Publications

References

1. Y. Deng, P. Luo, C. C. Loy, X. Tang, “Pedestrian Attribute Recognition At Far Distance”, International Conference on Multimedia, 2014.
2. X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, X. Wang, “HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis”, 2017.
3. D. Li, Z. Zhang, X. Chen, K. Huang, “A Richly Annotated Pedestrian Dataset for Person Retrieval in Real Surveillance Scenarios”, IEEE Transactions on Image Process, 2019.
4. X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets”, Proceedings of International Conference on Neural Information Processing Systems, 2016.
5. J. Snell, K. Swersky, R. Zemel, “Prototypical Networks for Few-Shot Learning”, Advances in Neural Information Processing Systems, 2017.
6. R. Droste, J. Jiao, J. A. Noble, “Unified Image and Video Saliency Modeling”, European Conference on Computer Vision, 2020.

Publications

7. E. Bekele, W. Lawson, “The Deeper, the Better: Analysis of Person Attributes Recognition”, IEEE International Conference on Face and Gesture Recognition, 2019.
8. W. Lawson, A.M.Harrison, LT E. Vorm, J.G.Trafton, “Joint Attention Estimator”, ACM Conference on Human-Robot Interaction (Late-Breaking report), 2020.
9. W. Lawson, E. Bekele, “Hierarhical Ontology for Attributes Recognition”, In Prep.
10. W. Lawson, M.L Chang, A.M. Harrison, W. Adams, J.G. Trafton, “Salient Meta Learning”, Under review.
11. P. Sandoval-Segura, “AutoProtoNet: Interpretability for Prototypical Networks”, Under review.

Naval/Marine Corps Needs

We have demonstrated the usefulness of our approach in several important application domains.

- In video surveillance, explainable factors can be used to greatly enhance the warfighter’s ability to search through relevant data and to identify things that are unusual. Using keypoint detection, we further provide the ability to provide an interpretable method to look for an object of interest in surveillance data.
- Shipboard maintenance tasks might be performed autonomously by a robot, in our case it is possible to train this robot in the target environment. This provides tremendously flexibility for the robot to learn new things and to correct mistakes that are made.

Summary of Findings

We summarize the findings of this project below

- Explainable factors (XF) can be useful as a to search for objects of interest. A small number of factors can be trained and evaluated quite easily, but as the number increases the training process becomes a bit more challenging. Knowledge engineering is necessary to properly organize data, but when done this can result in a great number of terms that can be used for searching.
- It’s possible to expand XF over time. By annotating only a small amount of the information on each data source, and then using our approach to identify when some factors are unknown we can easily deal with partially labeled information

- Keypoints provide a way to provide a visually interpretable reason for object recognition. These keypoints can be provided either by a human annotator or by another approach such as visual saliency.
- In our case study we showed that keypoints not only provide visually interpretable information, they can also increase accuracy.

Conclusions

Over the course of this project we have explored explainable factors and visually interpretable object recognition. Our work on explainable factors has resulted in an approach that can recognize a wide number of pedestrian attributes with a high level of accuracy.

Interpretability through keypoint detection can be a useful way to direct the attention of an object recognition to critical details. It can also provide some insight into the reasons behind why an object was classified. Keypoint selection can be done either manually or automatically through a saliency estimation approach.

Finally, some of the planned data collection was adversely affected by the ongoing COVID pandemic. Several planned studies were not performed, but could be resumed at some later date. We had ongoing plans to collaborate with visual surveillance companies to explore the potential of keypoints and explainable factors in real-world situations. A planned data collection was paused due to the pandemic but could be resumed when appropriate. Such a data collect could provide additional information on how easy it is to define new explainable factors on the fly and how well these factors generalize to new environments. A second data collect using eye-gaze as a method of keypoint selection was also delayed due the pandemic. In our work we use visual saliency, which is generally what attracts the attention of the human eye. A domain relevant evaluation can provide additional cues on whether task-driven saliency can further improve our results.