TITLE

Towards Persuasive Deep Learning for Video Analysis with Limited Supervision

Zhihui Li (z5187111)

Supervisors: Dr. Lina Yao and Prof. Salil Kanhere

Table of Contents

Introduction

Although dramatic success in deep learning (DL) [1] has led to an explosion of new AI capabilities in many real-world problems, including image classification [2], video analysis [3,4], natural language processing (NLP) [5,6], and more, adapting DL methods to perform learning applications with limited examples, such as video monitoring/surveillance systems, is still very challenging. For instance, medical devices are now frequently used in patients' homes, and it has been reported that many medical device-related incidents are caused by user error. Preventing such home medical device errors through video observation – for example, being able to recognise an unusual mistake in the operation of an infusion pump – could save a patient's life. In applications where security is the concern, recognizing/tracking a suspicious stranger in a surveillance system is critical to detecting unauthorised activity and could have a dramatic impact on situational outcome.

Deep learning models consist of hundreds of millions of parameters. Their superior performance results from being trained with millions of examples. The collection of such a large number of training examples for many concepts is very labour-intensive. It is unlikely that a database 10 times larger than the ImageNet[2] database or covering far more general concepts will ever be created. Due to the high cost of labelling and application requirements, however, many novel concepts appearing in daily life applications are represented in only a handful of examples[7], which are insufficient for learning a robust deep learning model. Moreover, the state-of-the-art deep network structure is very complicated and cannot be interpreted [8,9], thus it is non-trivial to design or adapt network structure for novel or unseen concepts. On the other hand, application experts usually have domain knowledge of feature engineering to design sophisticated features for specific concepts [10-15]. For example, region-specific hand-crafted features [16] can be designed to capture the semantics of complex scenes in images and videos. Such features, dubbed as privileged features, provide insight and privileged information about visual concepts. The **aim** of this proposal is to develop theoretical foundations and an interpretable deep learning paradigm that can 1) provide explanations of learning results; 2) facilitate learning with limited examples; and 3) communicate with users/experts using these privileged features. We focus on three limitations which have been largely unaddressed by existing studies:

Limitation (a) Lack of interpretability in deep network structures

Limitation (b) Poor capability to adapt to new concepts that have only a handful of examples

Limitation (c) Lack of communication mechanisms between DL models and experts/users.

To address these limitations, this proposal plans to develop a persuasive deep learning with privileged features model, which will create a suite of new or modified machine learning techniques to produce interpretable models which, when combined with effective explanation techniques. The proposed technologies will enable end users to understand and effectively adapt the learning model to few-shot observations. This research will benefit home patients by allowing them to safely use medical devices, and will enhance cybersecurity by detecting unauthorised access.

Literature Review

We review the progress in image/video analysis and relevant fields as follows.

Prior to 2012, hand-crafted features, such as SIFT [10], were used for image analysis [21]. Major advances have recently been made in deep learning which allow computational models to learn representations of data with multiple levels of abstraction to solve problems that have hitherto resisted the best attempts of the computer vision community [1]. The database of more than 15 million manually labelled digital images built by the Stanford AI lab enables researchers to investigate the effects of deep learning in a large-scale image recognition setting. Stanford researchers proposed a method of generating image descriptions [22] which requires a deeper understanding of images beyond object recognition. With the aim of building a large visual knowledge base with minimum human labelling effort, the Never Ending Image Learning project at CMU [12] automatically extracts visual knowledge from Internet data in a similar fashion to semi-supervised learning [20]. More recent works of interpretable machine learning can be found in [41].

Video analysis is more difficult than image analysis because of the lack of large video datasets for feature learning [14] [7]. Earlier studies usually focused on constrained videos produced by professionals, e.g. news

videos, films or lab- recorded videos [23]. The research focus has gradually shifted to unconstrained videos, which can be produced by anyone [14]. A joint work by CMU, IBM and Columbia University has constructed a video concept ontology [7] for semantic feature learning. Following a pilot evaluation, the U.S. government commenced an extraordinarily large five- year project, the ALADIN project, which seeks to combine researchers from industry and academia to develop a revolutionary way of creating a fast, accurate, robust, and extensible technology that supports video analysis needs. Driven by the increasing needs of government and consumers, academics at CMU, Stanford, MIT, UC Berkeley, the Australian Research Council Centre of Excellence for Robotic Vision and others, as well as researchers from industry such as Google, IBM and Microsoft, have poured resources into research on understanding unconstrained videos.

Advances in DL produced the aforementioned systems, which offer tremendous benefits, but their effectiveness is still limited by the machine's inability to explain its decisions and actions to human users. Interpretable learning methodology will be essential if users are to understand and trust the received information, especially when users transfer the knowledge to new applications (e.g. unusual mistakes in operating a home medical device, a suspicious stranger, abnormal activities, etc.) with only a handful of examples. In contrast, hand-crafted features provide intuitive domain knowledge from experts. Inspired by learning with privileged information [24,25], we will introduce an interpretable deep learning paradigm in this project which incorporates these privileged features to effectively communicate between DL models and experts, and facilitate efficient adaptation of DL models for few-shot learning.

Methodology

Our research will investigate a novel deep learning framework that provides an interpretable interface represented by privileged features. Since such features are hand-crafted by experts, other experts can easily understand their semantics. These privileged features will therefore become a common language that can facilitate communication between the DL model and users. We first present a novel interpretable deep learning architecture using privileged hand-crafted features. We will investigate how to extract and select activated deep network structures, based on these features, which can be used to understand the DL model. We will then study how to reuse the extracted deep network structures to perform rapid adaptation to new concepts with only a handful of examples, and will explore how to facilitate communication between the DL model and experts through the privileged features. Lastly, we will use image classification and video analysis, such as medical device usage detection, as prototype systems to evaluate the proposed framework.

Task T1: Interpretable Structure Extraction

In this task, we plan to study interpretable deep learning models. To facilitate the understanding of deep learning models, we intend to extract deep network structures by developing a deep learning model that combines input features (e.g. RGB pixel values) and privileged features (e.g. salience map, SIFT, segmentation) hand-crafted by experts.

The proposed framework is not-restricted to the visual domain and can be applied in various application domains; for example, we can extract semantic texture features/grammatical features in NLP applications. We can also extract subgraph features in social network analysis and use them as privileged features in applications to learn an interpretable DL model. We will use image classification and video analysis as our prototype system for evaluation, and our hand-crafted features will be designed mainly for image and video domains. The proposed task has two main steps: extraction of explainable hand-crafted image/video representation, and deep learning with combined input and privileged features (referred to below as a joint deep learning model).

Task 1.1 Extraction of explainable hand-crafted image/video representation. For some object concepts, as shown in Figure 2(a), salience provides the hint to identify the location of a kuola, thus a salience map can be used as the privileged information for designing features for object recognition. For some scenes, the SIFT features can easily provide the outline of some structures, therefore experts can design or combine different hand-crafted features for specific image classification tasks. Since video events have both spatial and temporal information, the improved Dense Trajectory (iDT) feature has dominated many video analysis tasks, such as event detection and action recognition, due to its superior performance over other features [11]. However, the iDT feature does not consider events in different orientations. This orientation information can be incorporated into iDT, and this subtask will explore various image/video feature extraction techniques to effectively uncover

the semantic structure of images and videos. We will use visual concepts from such applications as multimedia event detection, action recognition and surveillance event detection to evaluate these privileged features.

Task 1.2 Joint deep learning model with input and privilege features. Many researchers have recently switched to using ConvNet or Recurrent Neural Network (RNN) [32] to learn image/video features instead of using hand-crafted features. Once trained, it is only necessary to perform a forward pass through the pre-trained network to generate image/video representations. Simonyan and Zisserman [8] show that an improvement in prior-art configurations can be achieved by increasing the depth of networks, which leads to a very complicated network structure. Moreover, the performance of DL models critically depends on tuning the sophisticated network structure for a particular task, which is not trivial even for experienced practitioners. To this end, we propose to use hand-crafted features as the privileged features and design an interpretable deep learning model as shown in Figure 2(a), which jointly learns well-defined concepts using both input features and privileged features. To incorporate the domain knowledge of the privileged features, our design uses the privileged features as another output of the original ConvNet model; the privileged feature responses can thus be fed back to the original ConvNet model. This design has NOT been discussed in any existing work. Since the final DL model can capture the information from both types of features for training, the learned model will be more reliable and interpretable. We can use the both features to explain the learning results (e.g. which region representing the object/action/event). We can also use the extracted CNN features for prediction, which will significantly reduce the expensive computational time for extracting hand-crafted features.

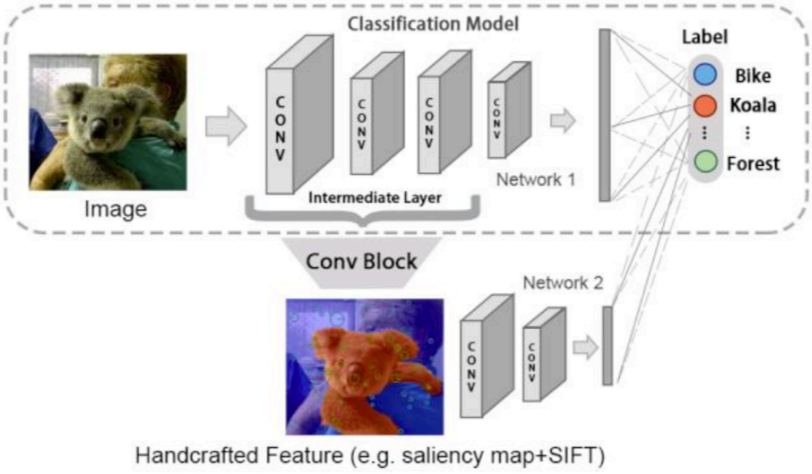

Figure 1 Joint Deep Learning Model.

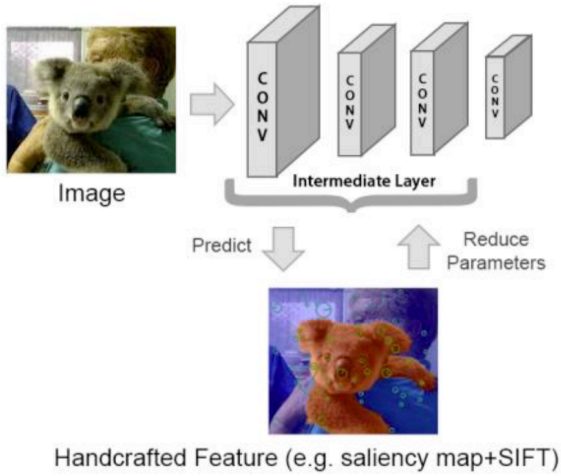

Figure 2 Prediction and Extraction of deep network structure.

Task T2: Efficient Adaption to Unseen Concepts

It is necessary when the learning task changes to redesign the structure and learn the DL model from scratch. Unlike image analysis, a particular problem here is the lack of sufficient labelled videos for learning an accurate model. In this task, we will investigate transfer learning methods to adapt novel concepts with a handful of examples. Existing DL models have hundreds of millions of parameters in the full network structure, and a fine-tuning method [31] is used to adapt the model learned by ImageNet [2] to another applications. However, the update procedure involving all the parameters requires thousands of examples for robust adaptation, which is not feasible for new or unseen concepts. To this end, we will investigate to reuse the interpretable DL model in Task T1, localise informative deep network structures, and adapt the parameters from the extracted deep network structures to novel or unseen concepts.

Task T2.1: Localisation of deep network structures. After the proposed joint deep learning model has been learned, we will investigate the use of gradient localisation techniques [33] or region proposal [38] to extract the activation structure of the network and identify the activation regions. As shown in Figure 2(b), our model has additional privileged features, which provides more domain knowledge than the existing approach [33] for identifying regions with better semantic meaning. We will explore a knockdown strategy for the hand-crafted features; for example, we can use the salience map to remove the background for some well-defined concepts and use it as the input feature. We can then easily localise the activated parts of the network, giving domain experts a better understanding of the network.

Task T2.2: Adaptation based on extracted structures. Instead of updating all the parameters, we propose to reduce the size of the parameter space in the deep network structure. As shown in Figure 2(b), we can identify which parts of the network are important. In this research, we will explore the reduction of the parameter space based on the extracted structured in Task T1 to adapt the new concepts (e.g. unusual mistakes, suspicious strangers, abnormal activities, etc.). Since the parameter space is greatly reduced, the number of required examples will be much fewer. Therefore, adaptation to novel concepts with limited examples becomes feasible.

Since fitting the DL model with limited examples may not be stable and may suffer from overfitting, in this project, we will also study how to improve the stability of adapting the DL model with limited examples. One possible method is to explore the ensemble learning strategy, which repeatedly removes some of network structures and aggregates the DL model. Therefore, a robust DL model could be adapted.

Task T3: Communication Mechanism between DL model and Experts

One major challenge in deep learning is how to design the network structure for good generalisation. The usual resort is a trial-and-error strategy to tune the network structure with distributed servers in a brute force manner, principally due to the lack of understanding of the behaviours and outcomes of the learned deep learning model.

In this research, we will explore how to exploit our interpretable DL model to provide explanations of learning results, which can be used to convey semantic feedback to experts. Our primary idea is to design a bidirectional communication mechanism through a common language that can be interpretable by both the DL model and the experts, which allows the experts to interpret some explainable outcomes of the learned model and perform further analysis, as well as, provides additional information, such as DL readable instructions, to update the DL model.

Since privileged features are designed by the experts based on their domain knowledge, they can understand the semantics of privileged features, which can serve as the common language in this task. As shown in Figure 1, the privileged features are connected to both the output of ConvNet and the output from the labels. Although the privileged features capture human knowledge on the learning task, the initial privileged features may not perfectly fit the training examples. In this task, we will explore how to distil human knowledge, provide feedback to the experts, and involve them into the loop to improve their privileged features and the DL model. As shown in Figure 3, we propose to leverage the backward response from the labels and the forward pass from the CNN model to localise important hand- crafted features. Such localisation of hand-crafted features provides more semantic insight into the outcome of the DL model. For example, as shown in Figure 4, we can use the responses to localise which region of SIFT features representing the kuola. Therefore, the experts can make use of this information to verify their hand-crafted features, analyse the DL model and tune its structure more effectively. Similarly, the experts can update their hand-crafted features and feed additional information as the instruction to tell the DL model how to improve its performance.

Figure 3 Communication using response from DL model

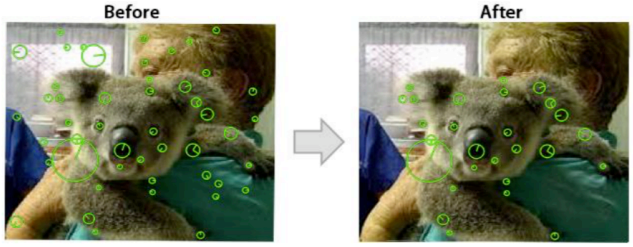

Figure 4 Localisation of hand-crafted features.

Besides this, we will also explore to use textual description (e.g. image caption) to extract NLP feature as the privileged feature, which naturally provides a human language to facilitate communication between the DL model and the users/experts. Thus, we can design an advanced communication mechanism for some complicated events in videos.

Evaluation and prototyping. A prototype system will be developed to showcase the usability of the output. The algorithms and tools developed will be analysed and validated in the following three application scenarios.

- *Event detection*. We will use the TB level TRECVID MED dataset collected by the U.S. National Institute of Standards Technology (NIST). The videos can be downloaded from the Internet, e.g., YouTube. We will also use the YouTube-8M Large-Scale Video Understanding task to test the performance of the output.
- *Activity detection in surveillance videos*. We will evaluate the proposed method using the TRECVID Surveillance Event Detection dataset collected by the NIST. This dataset was recorded at London Airport.
- *Detection of home medical device usage error*. We collected the dataset ourselves [39] and test the performance of the proposed method.

Research Plan

This project will have highly productive outcomes that will enrich the knowledge base in intelligent video processing and machine learning for social and economic benefit. The communication and dissemination of results will be achieved by the following means:

- We will publish papers in relevant discipline-specific journals and conferences such as IEEE TPAMI, IEEE TNNLS, IEEE TIP, ACM Multimedia, CVPR, and ICML. We plan to publish 4 ERA A* papers plus 5 ERA A papers during the life of the project.
- A prototype system will be developed.
- Table 1 presents the timeline for managing this project, providing a detailed schedule, corresponding expected outcomes and communication of results.

Research tasks, key indication factors, and outcomes	Year 1				Year 2				Year 3			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Interpretable Structure Extraction	3 A*A journal conference papers											
Efficient Adaption to Unseen Concepts			3 A*A journal conference papers									
Communication Mechanism between DL model and Experts					3 A*A journal conference papers							

Table 1 Timeline of this project.

Conclusions

In this proposal, we aim to develop theoretical foundations and a persuasive deep learning paradigm that can provide explanations of learning results, facilitate learning with limited examples, and communicate with users/experts using these privileged features. This project will have highly productive outcomes that will enrich the knowledge base in intelligent video processing and machine learning for social and economic benefit. We will publish papers in relevant discipline-specific journals and conferences

REFERENCES

[1] Y. LeCun, Y. Bengio and G. Hinton. Deep learning. Nature, 521: 436-444, 2015.
[2] A. Krizhevsky and I. Sutskever and G. Hinton. ImageNet classification with deep convolutional neural networks. NIPS, 2012.
[3] J. Ng et al. Beyond short snippets: Deep networks for video classification. CVPR, 2015.
[4] Z. Xu, Y. Yang and A. Hauptmann. A discriminative CNN video representation for event detection. CVPR, 2015.
[5] J. Hirschberg and C Manning. Advances in natural language processing. Science, 2015, 349(6245): 261-266.
[6] K. Cho et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. EMNLP, 2015
[7] M. Naphade, J. Smith, J. Tesic, S. Chang, W. Hsu, A. Hauptmann. Large-scale concept ontology for multimedia. IEEE MultiMedia, 13 (3), 86-91, 2006.
[8] K. Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. ICLR, 2015.
[9] X. Zhang, J. Zou, K. He and Jian Sun. Accelerating very deep convolutional networks for classification and detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/TPAMI.2015.2502579, 2016.
[10] D. Lowe. Object recognition from local scale-invariant features. ICCV, 1999.

[11] H. Wang and C. Schmid. Action recognition with improved trajectories. ICCV, 2013.
[12] X. Chen, A. Shrivastava and A. Gupta. Enriching visual knowledge bases via object discovery and segmentation. CVPR, 2014.
[13] S. Gao, L. Duan, I. Tsang. DEFEATnet - A Deep Conventional Image Representation for Image Classification. IEEE Trans. Circuits Syst. Video Techn. 26(3): 494-505 (2016)
[14] Z. Ma, Y. Yang, F. Nie, N. Sebe, S. Yan and A. Hauptmann. Harnessing lab knowledge for real-world action recognition. International Journal of Computer Vision, 109(1-2): 60-73, 2014.
[15] Z. Ma, Y. Yang, N. Sebe, K. Zheng and A. Hauptmann. Multimedia event detection using a classifier-specific intermediate representation. IEEE Transactions on Multimedia, 15(7):1628-1637, 2013.
[16] Z. Ren, S. Gao, L. Chia, I. Tsang. Region-Based Saliency Detection and Its Application in Object Recognition. IEEE Trans. Circuits Syst. Video Techn. 24(5): 769-779 (2014).
[17] L. Fei-Fei, R. Fergus, P. Perona. One-shot learning of object categories. TPAMI 28(4):594-611, 2006
[18] M. Palatucci, D. Pomerleau, G. Hinton, T. Mitchell. Zero-Shot Learning with Semantic Output Codes. NIPS 2009. pp1401-1418.
[19] Z. Ma, Y. Yang, N. Sebe and A. Hauptmann. Knowledge adaptation with partially shared features for event detection using few exemplars. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(9), 1789–1802, 2014.
[20] Y Yang, Z Ma, Z Xu, S Yan and A Hauptmann. How related exemplars help complex event detection in web videos? ICCV 2013.
[21] Y Yang, F Wu, F Nie, HT Shen, Y Zhuang and A Hauptmann. Web and personal image annotation by mining label correlation with relaxed visual graph embedding. IEEE Transactions on Image Processing, 21 (3), 1339-1351, 2012.
[22] A. Karpathy, A. Joulin and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. CVPR, 2015.
[23] I. Laptev, B. Marszałek, C. Schmid and B. Rozenfeld. Learning realistic human actions from movies. CVPR, 2008.
[24] V. Vapnik, A. Vashist. A new learning paradigm: Learning using privileged information. Neural Network 22(5):544-557, 2009.
[25] X. Xu, J. Zhou, I. Tsang, Z. Qin, R. Goh, Y. Liu. Simple and Efficient Learning using Privileged Information. IJCAI 2016 Workshop on BeyondLabeler - Human is More Than a Labeler.
[26] S. Pan, Q. Yang. A Survey on Transfer Learning. TKDE 2010 22(10):1345-1359
[27] S. Pan, I. Tsang, J. Kwok, Q.Yang. Domain Adaptation via Transfer Component Analysis. IEEE TNN 22(2): 199-210, 2011.
[28] J. Zhou, S. Pan, I. Tsang, Y. Yan. Hybrid Heterogeneous Transfer Learning through Deep Learning. AAAI 2014: 2213-2220.
[29] P. Liu, J. Zhou, I. Tsang, et. al. Feature Disentangling Machine - A Novel Approach of Feature Selection and Disentangling in Facial Expression Analysis. ECCV (4) 2014: 151-166
[30] J. Zhou, X. Xu, S. Pan, I. Tsang, Z. Qin, R. Goh. Transfer Hashing with Privileged Information. IJCAI 2016: 2414-2420.
[31] J. Yosinski et al. How transferable are features in deep neural networks? NIPS 2014, pp. 3320-3328.
[32] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997.
[33] B. Zhou et. al. Learning Deep Features for Discriminative Localization, CVPR 2016.
[34] Z Xu, Y Yang, I Tsang, N Sebe, A Hauptmann. Feature weighting via optimal thresholding for video analysis. ICCV‟13, 3440-3447
[35] Z Xu, I Tsang, Y Yang, Z Ma, A Hauptmann. Event detection using multi-level relevance labels and multiple features . CVPR‟14, 97-104 [36]Y Yan, Z Xu, I Tsang, G Long, Y Yang. Robust Semi-Supervised Learning through Label Aggregation. AAAI, 2244-2250
[37] Z. Xu, L. Zhu, Y. Yang and A. Hauptmann. UTS-CMU at THUMOS 2015. CVPR, 2015.
[38] S. Ren, K. He and R. Girshick. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS, 2015.
[39] Y Yang, A Hauptmann, et al. Learning to predict health status of geriatric patients from observational data. IEEE CIBCB 2012.
[40] Y. David, W. Hyman, V. D. Woodruff, and M Howell. Overcoming barriers to success: Collecting medical device incident data. Biomedical Instrumentation and Technology. 41(6):471, 473-5, 2007.
[41] NIPS 2016 Workshop on Interpretable ML for Complex Systems. https://sites.google.com/site/nips2016interpretml/