This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Building Explainable AI Evaluation for Autonomous Perception

Chi Zhang†, Biyao Shang†, Ping Wei†, Li Li‡, Yuehu Liu†, Nanning Zheng†
†Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, 710049, China
‡Department of Automation, BNRist, Tsinghua University, Beijing, 100084, China
{colorzc, shangbiyao}@stu.xjtu.edu.cn, {liuyh, pingwei, nnzheng}@xjtu.edu.cn, li-li@tsinghua.edu.cn

Abstract

The development of the robust visual intelligence is one of the long-term challenging problems. From the perspective of artificial intelligence evaluation, the need to discover and explain the potential shortness of the evaluated intelligent algorithms/systems as well as the need to evaluate the intelligence level of such testees are of equal importance. In this paper, we propose a possible solution to these challenges: Explainable Evaluation for visual intelligence. Compared to the existing work on Explainable AI, we focus on the problem setting where the internal mechanisms of AI algorithms are sophisticated, heterogeneous or unreachable. In this case, the interpretability of test output is formulated as an semantic embedding to the existing correlation between factors of data variances and test outputs. Dictionary learning is introduced to jointly estimate the semantic mapping and the semantic representations for explanation. The optimal solution of proposed method could be reached via an alternating optimization process. The application of the "Explainable AI Evaluation" will strengthen the influence of objective assessment for visual intelligence.

1. Introduction

Recently we have witnessed a series of success that visual perception and understanding in traffic environment has achieved with the help of emerging artificial intelligence techniques, such as vehicle and pedestrian detection, recognition and semantic segmentation, etc. These data-driven visual methods learn discriminative representations with respect to the factors of visual appearance variances and try to handle all of them. However, the test and applications of such methods reveal that the reliability could not be guaranteed in many real world situations like poor illumination, motion blur, spot area, noises caused by extreme weathers and etc. Therefore, developing autonomous visual perception methods is still one of the most challenging problems.

In the long-run research and development process, it is crucial to explore the blind spots and failure modes of certain visual perception methods. Despite the performance degradation which visual approaches could not handle intrinsically, such disadvantages may be caused by the bias of training set[8] as well as the intrinsic capability of designed architecture of certain approaches. Most recently, several techniques has been developed to address discovering potential training set bias [13] and interpreting the learned visual representations of deep networks w.r.t. the network performance[1]. Based on the hypothesis that interpretability is an property of the learned representations[1], these methods tend to explain the effect of components like convolutional layers or batch normalization of the given network architecture. Here we define this kind of property as **intra-interpretability** since it comes from the its own architecture design.

In this paper we focus on the interpretability comes from the exploration on external test data, i.e. **inter-interpretability**, when the structure and implementation details of certain method are invisible. Such "blackbox" problem setting is common in the benchmarks for visual algorithms [3, 2] or artificial intelligent tests and evaluations [6, 11, 10]. The backbone assumption is that for the same intelligent approach in a certain evaluation process, the data variations are the major influence to the variations in output performance. In this case, if we can find a semantic embedding of human domain knowledge to the factors of data variance, we might be able to establish the descriptive relationship between semantic concepts and output performance.

To this end, we propose "explainable AI evaluation" for mining the inter-interpretability for the performance of autonomous perception intelligence, as shown in Fig. 1. We address first to infer the deterministic relationship by Ridge Regression between factors of data variances (denoted as Disentangled Variables) and test outputs of certain AI method according to its performance variances under variances of test data. Further, the semantic mapping from human domain knowledge to Disentangled Variables, along with the semantic representations (denoted as Explainable Representations) simultaneously, is obtained by subspace

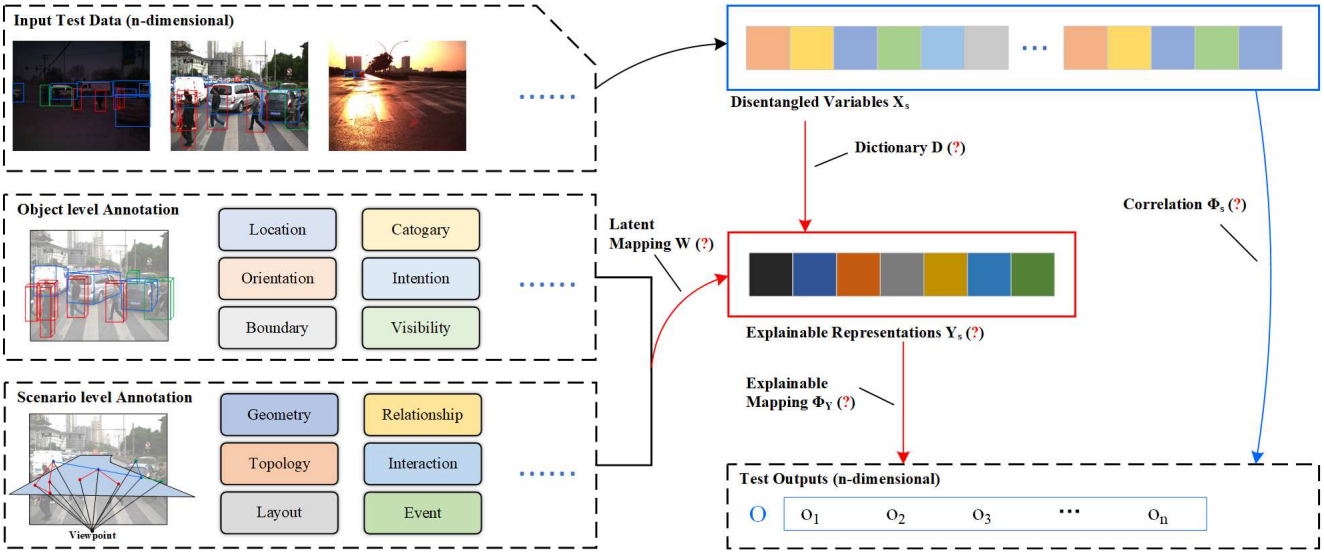

Figure 1. **The flowchart of proposed explainable AI evaluation framework.** The Correlation Φ_s between Disentangled Variables X_s is obtained firstly through Ridge Regression. The rest of problem solving could be viewed as a dictionary learning process where Dictionary D and Explainable Representations Y_S are estimated jointly, under the constraints that D, Φ_Y, Φ_s should be consistent. In this way, we can finally obtain an Explainable Mapping Φ_Y to interpret test outputs by estimated Explainable Representations Y_S.

embedding of semantic concepts in human domain knowledge via dictionary learning. As a consequence, the relationship between Disentangled Variables and test outputs could be interpreted partly by knowledges and concepts in the way that humans would understand.

The contributions of this paper is two-fold:

1. The concept of interpretability is introduced into visual intelligence evaluation, resulting in presentation of "explainable AI evaluation".

2. The problem is formulated as an semantic embedding to the existing correlation between Disentangled Variables and test outputs, the optimal solution of which is obtained via alternating optimization process, as shown in Section. 2.

We believe that "intra-interpretability" and "inter-interpretability" are equally important for visual AI methods. The application of the latter one in intelligence test and evaluation will strengthen the influence of objective assessment in the long-term research and development of visual intelligence and hence contributes to the further improvement of visual AI.

2. Proposed Approach

We propose a framework for explaining the test outputs O (L-dimensional) by Explainable Mapping Φ_Y, as shown in Fig. 1. Intrinsically, it implies to predict how and to what extent does the variations associated with human domain-knowledge influence the unknown testee, resulting in the variations in test outputs.

To this end, we need to firstly extract the factors of data variations from the input test data. These factors, namely Disentangled Variables, have the definite correlation with the test outputs, which is the backbone of machine learning theory.

Disentangled Variables, which encodes the complex variations of the input data, are of high dimensions. Such representations may be beyond human understanding. Hence, we assume without the loss of generality that the variations which could be described by human domain knowledge, i.e. Explainable Representations, are sparse subsets of Disentangled Variables.

We further propose that such sparse mapping need to be consistent with the Explainable Mapping and the correlation mapping between Disentangled Variables and test outpus. In this way, these mappings could be predicted jointly utilizing dictionary learning.

2.1. Notations

Suppose there are n labeled samples $Q_s = \{I_s, A_s, O_s\}$ where I_s denotes the input test images, $\mathbf{A}_s = \{A_S, A_O\}$ is the domain knowledge annotation in both object level (Location, Orientation, and etc.) and scenario level (Geometry, Topology and etc.). O_s is the test output of certain intelligent method under given visual task. For example, O_s could be a set of per-image mean accuracy or F-score in Detection or per-image mean Intersection-over-Union(IoU) in Segmentation, corresponding to the input I_s. Due to the paragraph limitation, here we assume that the d-dimensional Disentangled Variables X_s w.r.t. each I_s has been explored by certain disentangled learning approach

like [7].(Although it is still an open question.) Given Q_s and X_s, the goal of our method is to find an optimal Explainable Mapping Φ_Y from unknown Explainable Representations Y_s to the test outputs O_s.

2.2. Correlation Analysis

We start our approach by performing correlation analysis between the factors of data variations and variable test outputs, which is a multi-variate regression problem inherently. In order to prevent singular matrix caused by multicollinearity, we utilize Ridge Regression to explore this correlation Φ_s by minimizing the following loss, similar as [12]:

$$\arg\min_{\Phi_s} \|\mathbf{O}_s - \Phi_s \mathbf{X}_s\|_F^2 + \lambda \|\Phi_s\|_F^2 \tag{1}$$

The analytic solution could be obtained as follows,

$$\Phi_s = \mathbf{O}_s \mathbf{X}_s^T (\mathbf{X}_s \mathbf{X}_s^T + \lambda \mathbf{I})^{-1} \tag{2}$$

Here we suppose that the pre-computed orthogonal Disentangled Variables X_s could effectively represent factors of data variations. In this case, low rank constraints could be used to select variables which tend to have an obvious influence by forcing the coefficients of variables which perform no effect to be zeros.

2.3. Explainable Mapping

Given Disentangled Variables X_s and corresponding annotation $\mathbf{A}_s$, the key to enable interpretability for variable-output correlation Φ_s lies in the embedding of annotation $\mathbf{A}_s$ in the space of Disentangled Variables X_s. As stated above, X_s is of high dimensions where only a small part of variables could be explained by human domain knowledge.

Considered that the multicollinearity of $\mathbf{A}_s$ may compromise the linear mapping between X_s and $\mathbf{A}_s$, we propose to use a linear transformation of $\mathbf{A}_s$, i.e. Explainable Representations $\mathbf{Y}_s$, to explain the outputs $\mathbf{O}_s$ via the Explainable Mapping $\mathbf{\Phi}_Y$, constrained by the mapping consistency $\|\Phi_Y \mathbf{Y}_s - \Phi_S \mathbf{D} \mathbf{Y}_s\|_F^2$. The loss function of our approach is given as follows,

$$\begin{aligned} \arg \min_{\mathbf{D},\mathbf{Y}_s,\mathbf{W},\Phi_Y} & \|\mathbf{X}_s - \mathbf{D}\mathbf{Y}_s\|_F^2 + \lambda_1 \|\mathbf{Y}_s - \mathbf{W}\mathbf{A}_s\|_F^2 \\ & + \lambda_2 \|\mathbf{O}_s - \Phi_Y \mathbf{Y}_s\|_F^2 + \lambda_3 \|\Phi_Y - \Phi_S \mathbf{D}\|_F^2, \\ & s.t. \|d_i\|_2^2 \leqslant 1, \|w_i\|_2^2 \leqslant 1, \|\phi_i\|_2^2 \leqslant 1, \forall i \end{aligned} \tag{3}$$

where $\mathbf{D}$ is the learned dictionary and $\mathbf{Y}_s$ is sparse representations of $\mathbf{X}_s$ w.r.t. $\mathbf{D}$. $\mathbf{W}$ is the linear transformation mapping between $\mathbf{Y}_s$ and $\mathbf{A}_s$.

2.4. Optimization

From Eq. 3 we are able to determine the non-convex characteristics of the proposed framework. However, alternating optimization could be applied since each term of the loss function is convex by itself, as in [5]. Comparably, our optimization process, as shown in Alg. 1, is a little bit more complex due to the augmentation of an extra term of unknown variables, where

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_s \\ \lambda_1 \mathbf{W}\mathbf{A}_s \\ \lambda_2 \mathbf{O}_s \end{bmatrix}, \tilde{\mathbf{D}} = \begin{bmatrix} \mathbf{D} \\ \lambda_1 \mathbf{I} \\ \lambda_2 \mathbf{\Phi}_Y \end{bmatrix}, \tag{4}$$

and Λ_1, Λ_2 are constructed diagonal matrices.

Algorithm 1 Alternating optimization method for solving explainable mapping

Ensure: Disentangled Variables $\mathbf{X}_s$, Test Outputs $\mathbf{O}_s$, human knowledge annotations $\mathbf{A}_s$, Correlation $\mathbf{\Phi_s}$, Lagrangian multipliers $\lambda_1, \lambda_2, \lambda_3$

Require: Dictionary $\mathbf{D}$, Explainable Representations $\mathbf{Y}_s$, Explainable Mapping $\mathbf{\Phi}_Y$, Latent Mapping $\mathbf{W}$

1: Initialize $\mathbf{D}$,$\mathbf{Y}_s$,$\mathbf{\Phi}_Y$, $\mathbf{W}$.
2: **while** not converge **do**
3: Compute $\mathbf{Y}_s = (\tilde{\mathbf{D}}^T \tilde{\mathbf{D}})^{-1} \tilde{\mathbf{D}}^T \tilde{\mathbf{X}}$.
4: Solve $(\mathbf{X}_s \mathbf{Y}_s^T + \lambda_3 \mathbf{\Phi}_s^T \mathbf{\Phi}_Y) = \mathbf{D}\mathbf{Y}_s \mathbf{Y}_s^T + \lambda_3 \mathbf{\Phi}_s^T \mathbf{\Phi}_s \mathbf{D}$ by [4].
5: Compute $\mathbf{W} = (\mathbf{Y}_s \mathbf{A}_s^T)(\mathbf{A}_s \mathbf{A}_s^T + \Lambda_1)^{-1}$.
6: Compute $\mathbf{\Phi}_Y = (\lambda_3 \Phi_s \mathbf{D} + \lambda_2 \mathbf{O}_s \mathbf{Y}_s^T)(\lambda_2 \mathbf{Y}_s \mathbf{Y}_s^T + \lambda_3 \mathbf{I} + \Lambda_2)^{-1}$.
7: **end while**

3. Experiments

3.1. Details

Dataset details. To sufficiently and profoundly evaluate visual intelligence for recognition and understanding in road traffic environment, we manually selected 1400 inconsecutive images of 14307 annotated vehicle instances from TSD-Max dataset[1] to build a diverse and difficult visual benchmark set, namely Explainable Visual Benchmark (EVB) dataset.

Parameter details.The 4096-dimensional visual feature of each image is extracted by the ImageNet pre-trained VGGNet as the Disentangled Variables $\mathbf{X}_s$. Besides, the per-image attributes $\mathbf{A}_s$ discussed in [11] and [9] are annotated as a 21-dimensional vector, including Road Type, Scenario Type, Acquisition Time, Weather Conditions and Complex Illuminations. Moreover, three different vehicle detection networks, i.e. Mask-RCNN, SSD and YOLO (all pre-trained on MS-COCO dataset), are evaluated on the proposed EVB dataset. For each algorithm, we obtain the per-image performance, including Precision, Recall and F-score as $\mathbf{O}_s$.

Table 1. Top-5 attributes which contributes to the F-scores of Mask-RCNN Detection Performance estimated by Ridge Regression, LASSO and Our Method. (+ means positive influence while - means negative influence)

Methods	1st	2nd	3rd	4th	5th
Ridge Regression	Intersection(+1.00)	Overcast(+0.81)	Sunny(-0.71)	City(-0.54)	Tunnel(-0.51)
LASSO	Tunnel(-1.00)	Residential(+0.57)	Night(+0.36)	Toll(+0.30)	Intersection(+0.27)
Ours	**Tunnel(-1.00)**	**Residential(+0.91)**	**Toll(+0.51)**	**Night(-0.38)**	**Intersection(+0.37)**

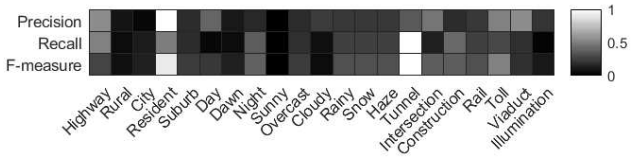

Figure 2. **The estimated interpretable mapping between semantic concepts and output vehicle detection performance of pre-trained Mask-RCNN.**

3.2. Qualitative Results

The interpretability of the output performance $\mathbf{O}_s$ lies in the correlation mapping $\mathbf{\Phi}_Y\mathbf{W}$ with semantic annotations $\mathbf{A}_s$. Fig. 2 illustrates the absolute influence of the semantic concepts (21 dimensions) to the output performance (3 dimensions) of pre-trained Mask-RCNN. For example, the 9th column, "Sunny Days" , has little impact on the performance because the majority of the selected images (944/1400) are acquired under this weather condition. It could be also observed that the 15th column, "Tunnel", however influences the detection performance drastically, which matches the human intuition.

We further compare the Top-5 attributes predicted by Ridge Regression, LASSO and the proposed method respectively, as described in Table. 1. Our method is consistent with the predicted results by LASSO regression in general, but the prediction for the contribution of "Night" attribute to the overall F-score differs from each other. Considering that the average F-score Mask-RCNN achieved on images annotated with"Night" is lower than the mean average F-score, our method characterizes the relationship between the "Night" attribute and the detection performance more reasonably.

4. Conclusion

In this paper, we introduce the concept of "Explainable AI Evaluation" based on the interpretation of test outputs of certain AI method using human domain knowledge. Such problem is formulated as a dictionary learning process, where Explainable Representations and semantic embedding are jointly obtained with the constraints on mapping consistency. Optimal solution is achieved via alternating optimization. The proposed framework could be beneficial to the intelligence test and evaluation for visual AI.

[1] `http://trafficdata.xjtu.edu.cn/index.do`

References

[1] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3319–3327, July 2017.

[2] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan 2015.

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, June 2012.

[4] A. Jameson. Solution of the equation ax + xb = c by inversion of an m × m or n × n matrix. *SIAM Journal on Applied Mathematics*, 16(5):1020–1023, 1968.

[5] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen. Learning discriminative latent attributes for zero-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4223–4232, 2017.

[6] L. Li, Y.-L. Lin, N.-N. Zheng, F.-Y. Wang, Y. Liu, D. Cao, K. Wang, and W.-L. Huang. Artificial intelligence test: a case study of intelligent vehicles. *Artificial Intelligence Review*, 50(3):441–465, Oct 2018.

[7] M. F. Mathieu, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pages 5040–5048, 2016.

[8] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, June 2011.

[9] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *CoRR*, abs/1805.04687, 2018.

[10] C. Zhang, Y. Liu, L. Li, N. Zheng, and F. Wang. Joint task difficulties estimation and testees ranking for intelligence evaluation. *IEEE Transactions on Computational Social Systems*, pages 1–6, 2019.

[11] C. Zhang, Y. Liu, Q. Zhang, and L. Wang. A graded offline evaluation framework for intelligent vehicles cognitive ability. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 320–325, June 2018.

[12] L. Zhang, S. K. Shah, and I. A. Kakadiaris. Hierarchical multi-label classification using fully associative ensemble learning. *Pattern Recognition*, 70:89–103, 2017.

[13] Q. Zhang, W. Wang, and S. Zhu. Examining CNN representations with respect to dataset bias. *CoRR*, abs/1710.10577, 2017.