Advances in Artificial Intelligence for Data Visualization: Automate Reading of Diagnostic Plots with Compute Vision Models

Weihao (Patrick) Li
PhD student

Report for
Progress review

5 April 2023

MONASH
University

MONASH
BUSINESS
SCHOOL

**Department of
Econometrics &
Business Statistics**

(04) 0459 1219
weihao.li@monash.com

ABN: 12 377 614 012

keywords: AI, data visualization, residual plot, visual inference, hypothesis testing, computer vision

1 Overview of the thesis

1.1 Background and motivation

Model diagnostics play a critical role in evaluating the accuracy and validity of a statistical model. They enable the assessment of the model's assumptions about the data, identification of outliers or anomalous observations that may have an exaggerated effect on the model, evaluation of how well the model fits the data, and identification of possible approaches to improve the model's performance.

When conducting model diagnostics, despite the availability of numeric summaries with finite or asymptotic properties that have been endorsed, data analysts prefer or require graphical representations of data. The preference for visual diagnostics is attributed to its intuitive nature and the possibility of discovering unexpected abstract and unquantifiable insights. In the context of regression diagnostics, a common practice is to plot residuals against fitted values, which serves as a starting point for evaluating the adequacy of the fit and verifying the underlying assumptions. Other visualization techniques such as histograms, Q-Q plots and box plots can be used to identify potential issues with assumptions made about the data, such as linearity, normality, or homoscedasticity.

In recent times, a novel statistical inferential framework known as visual inference (Buja et al. 2009) has been developed, which relies on the use of graphical representations of data. The visual inference approach capitalizes on the innate capability of the human visual system to rapidly identify patterns and deviations from expected patterns. By exploiting visualization techniques, visual inference provides a more natural and easily comprehensible way of interpreting data and discerning relationships between variables, compared to conventional statistical techniques. This inferential framework can be applied to perform exploratory data analysis, model selection, and hypothesis testing.

In the same paper, Buja et al. (2009) proposes the use of the lineup protocol as a visual hypothesis test, inspired by the police lineup technique employed in eyewitness identification of criminal suspects. The protocol comprises m randomly positioned plots, where on of them represents the data plot, while the remaining $m-1$ plots represent the null plots with the same graphical structure, except that the data in them is consistent with the null hypothesis H_0. Under H_0, the data plot is expected to be indistinguishable from the null plots, and the probability of correctly identifying the data plot by an observer is $1/m$. If the data plot is correctly identified, it provides evidence against H_0.

This method has gained increasing traction in recent years and has already been integrated into data analysis of various topics, such as diagnostics of hierarchical linear models (Loy and Hofmann 2013), geographical research (Widen et al. 2016) and forensic examinations (Krishnan and Hofmann 2021).

With the advent of sophisticated visualization software and tools, visual inference has the potential to revolutionize statistical analysis by providing an innovative alternative to traditional statistical approaches and enabling more lucid and effective communication of research findings.

The incorporation of human judgment is a fundamental aspect of visual testing, but it may restrict its widespread usage. The reliance on human assessment, akin to handicraft production methods employed in pre-industrial societies, renders the lineup protocol unsuitable for large-scale applications, due to its high labor costs and time requirements. Moreover, it presents significant usability issues for individuals with visual impairments, resulting in reduced accessibility.

The advancement of technology is essential to alleviate the workload of individuals by automating repetitive tasks and providing standardized results in a controlled environment. The evaluation of visual tests on a large scale is not feasible without the aid of machines and technology. Modern computer vision models offer a promising solution to this challenge. As a subfield of AI, computer vision with its modern deep learning architectures has successfully resolved numerous critical problems in automation. The development of the convolutional neural network (CNN) by Fukushima and Miyake (1982) was inspired by the vision processing in living organisms. This architecture was first applied to hand-written number recognition trained with back-propagation by LeCun et al. (1989), making it one of the earliest attempts at successfully extracting information from digital images via self-learning algorithms. The modern computer vision model typically utilizes deep neural networks with convolutional layers (Fukushima and Miyake 1982), which leverage the hierarchical pattern in data and provide regularized versions of fully-connected layers. This approach downscales and transforms images by summarizing information in a small space. Numerous studies have shown that it can effectively tackle vision tasks, such as image recognition (Rawat and Wang 2017). The development of graphics processing units and the widespread availability of high-performance personal computers have made computer vision research a new trend in the 21st century. The remarkable achievements, such as computer-aided diagnosis (Lee and Chen 2015), pedestrian detection (Brunetti et al. 2018), and facial recognition (Emami and Suciu 2012), have significantly impacted our daily life.

The utilization of computer vision models for reading data plots is not widely adopted. Nevertheless, certain fields have embraced this notion, as evidenced by the application of computer vision models in reading recurrence plots for time series regression (Ojeda, Solano, and Peramo 2020), time series classification (Chu et al. 2019; Hailesilassie 2019; Hatami, Gavet, and Debayle 2018; Zhang et al. 2020), anomaly detection (Chen, Su, and Yang 2020), and pairwise causality analysis (Singh et al. 2017). However, the assessment of lineups with computer vision models represents a relatively novel area of investigation.

1.2 Research questions

The main objective of this research is to construct an automatic visual inference system to facilitate conducting regression diagnostics visual tests on a large scale. The study will concentrate on three specific projects, namely

1. Exploring the application of visual inference in regression diagnostics and comparing its efficacy with conventional hypothesis tests.
2. Designing an automated visual inference system to assess lineups of residual plots derived from the classical normal linear regression model.
3. Deploying the automatic visual inference system as an online application and and making the pertinent open-source software accessible.

1.3 Current reserach outcomes

3. summarise the work in the first year

- an experiment to compare visual test with conventional test
- a prototype model for lineup evaluation
- points out the limitation of the work

4. summarise the work in the second year

- refine and reconduct the experiment

5. main findings to date

- findings from the first year
- findings from the second year

2 Thesis structure

topic

chapter 1: Introduction

chapter 2 the first paper

chapter 3 build a computer vision system to evaluate residual plot

chapter 4? software publication

chapter 5 discussion and future work

3 Timetable

- April: submit abstract to ASC

- May: submit paper

- June: submit poster/short paper to IEEE vis conf, start working on computer vision model

- July: leave for few weeks

- Aug

- Sep: finalize the computer vision model

- Oct: IEEE vis conf

- Nov: web interface development

- Dec: ASC

- Jan, 2024:

- Feb, 2024:

- Mar, 2024: submit paper

- Aug, 2024: submit thesis

4 Difficulties

things that are actually quite serious

1. identified difficulties
2. suggestions for overcoming these difficulties

References

Brunetti, A, D Buongiorno, GF Trotta, and V Bevilacqua (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing* **300**. Publisher: Elsevier, 17–33.

Buja, A, D Cook, H Hofmann, M Lawrence, EK Lee, DF Swayne, and H Wickham (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**(1906), 4361–4383.

Chen, Y, S Su, and H Yang (2020). Convolutional neural network analysis of recurrence plots for anomaly detection. *International Journal of Bifurcation and Chaos* **30**(01). Publisher: World Scientific, 2050002.

Chu, H, X Liao, P Dong, Z Chen, X Zhao, and J Zou (2019). An automatic classification method of well testing plot based on convolutional neural network (CNN). *Energies* **12**(15). Publisher: Multidisciplinary Digital Publishing Institute, 2846.

Emami, S and VP Suciu (2012). Facial recognition using OpenCV. *Journal of Mobile, Embedded and Distributed Systems* **4**(1), 38–43.

Fukushima, K and S Miyake (1982). "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition". In: *Competition and cooperation in neural nets*. Springer, pp.267–285.

Hailesilassie, T (2019). Financial Market Prediction Using Recurrence Plot and Convolutional Neural Network.

Hatami, N, Y Gavet, and J Debayle (2018). Classification of time-series images using deep convolutional neural networks. In: *Tenth international conference on machine vision (ICMV 2017)*. Vol. 10696. International Society for Optics and Photonics, pp.106960Y.

Krishnan, G and H Hofmann (2021). Hierarchical Decision Ensembles- An inferential framework for uncertain Human-AI collaboration in forensic examinations. *arXiv:2111.01131 [cs, stat]*.

LeCun, Y, B Boser, JS Denker, D Henderson, RE Howard, W Hubbard, and LD Jackel (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation* **1**(4). Publisher: MIT Press, 541–551.

Lee, H and YPP Chen (2015). Image based computer aided diagnosis system for cancer detection. *Expert Systems with Applications* **42**(12), 5356–5365.

Loy, A and H Hofmann (2013). Diagnostic tools for hierarchical linear models. *Wiley Interdisciplinary Reviews: Computational Statistics* **5**(1), 48–61.

Ojeda, SAA, GA Solano, and EC Peramo (2020). Multivariate time series imaging for short-term precipitation forecasting using convolutional neural networks. In: *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, pp.33–38.

Rawat, W and Z Wang (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation* **29**(9), 2352–2449.

Singh, K, G Gupta, L Vig, G Shroff, and P Agarwal (2017). Deep convolutional neural networks for pairwise causality. *arXiv preprint arXiv:1701.00597*.

Widen, HM, JB Elsner, S Pau, and CK Uejio (2016). Graphical Inference in Geographical Research. en. *Geographical Analysis* **48**(2), 115–131.

Zhang, Y, Y Hou, S Zhou, and K Ouyang (2020). Encoding time series as multi-scale signed recurrence plots for classification using fully convolutional networks. *Sensors* **20**(14). Publisher: Multidisciplinary Digital Publishing Institute, 3818.