Two to Trust: AutoML for Safe Modelling and Interpretable Deep Learning for Robustness

Mohammadreza Amirian[1,2], Lukas Tuggener[1,3], Ricardo Chavarriaga[1,4], Yvan Putra Satyawan[1], Frank-Peter Schilling[1], Friedhelm Schwenker[2], and Thilo Stadelmann[1]

[1] ZHAW Zurich University of Applied Sciences, Winterthur, Switzerland
[2] Ulm University, Ulm, Germany
[3] Università della Svizzera italiana, Lugano, Switzerland
[4] Confederation of Laboratories for AI Research in Europe, Zurich, Switzerland
{amir, tugg, chav, saty, scik, stdm}@zhaw.ch
friedhelm.schwenker@uni-ulm.de

Abstract. *With great power comes great responsibility.* The success of machine learning, especially deep learning, in research and practice has attracted a great deal of interest, which in turn necessitates increased trust. Sources of mistrust include matters of model genesis ("Is this really the appropriate model?") and interpretability ("Why did the model come to this conclusion?", "Is the model safe from being easily fooled by adversaries?"). In this paper, two partners for the trustworthiness tango are presented: recent advances and ideas, as well as practical applications in industry in (a) Automated machine learning (AutoML), a powerful tool to optimize deep neural network architectures and fine-tune hyperparameters, which promises to build models in a safer and more comprehensive way; (b) Interpretability of neural network outputs, which addresses the vital question regarding the reasoning behind model predictions and provides insights to improve robustness against adversarial attacks.

Keywords: automated deep learning (AutoDL) · adversarial attacks

1 Introduction

The recent success of machine learning (ML) and deep learning (DL) has triggered enormous interest in practical applications of these algorithms in many organizations [23, 24]. The emergence of automated ML (AutoML), which includes automated DL (AutoDL), further expands the horizons of such machine learning applications for non-experts and broadens the feasibility of exploring larger search spaces during development. Establishing trust in ML and DL models is thereby vital before they can be applied to real-world problems. Accordingly, trustworthiness has been recognized as the core concept for the applicability of ML algorithms during the first TAILOR workshop at the European Conference on Machine Learning (ECML 2019[1]).

[1] https://ecmlpkdd2019.org

One approach in recent years has been focused on understanding the behavior of deep Convolutional Neural Networks (CNNs) by visualizing the network response in order to boost its trustworthiness [1]. However, there has always been a gap between state-of-the-art research in network architecture development and interpretability, with interpretability often lagging behind. Only more recent attempts incorporate interpretability into architecture design [30, 20, 21]. For example, architectures for image classification have been proposed which produce decisions in a more human-interpretable manner and hence shift the paradigm from maximizing performance to learning "the right for the right reasons" [20]. Consequently, incorporating interpretable model design into the growing domain of AutoML is likely to shorten its path towards practical applications.

In this paper, we present preliminary results that contribute to trustworthy neural network development and application in two respects: First, we report recent advances in AutoML and propose a unified architecture for multi-modal input data (audio and video) through an automated and thus repeatable development process, leading to safer architectures. Secondly, we introduce visualization techniques that improve the interpretability of a model's decision and show how they allow detection of adversarial attacks, improving the model's robustness and design. We argue that the feasibility and effectiveness of deploying AutoML methods contributes to improved trustworthiness. However, many challenges are still to be addressed in order to solve the tension between algorithmic automation and trustworthiness, especially in the case of algorithmic ensembles.

2 Automated Machine Learning

In this section, we present recent advances in automating the development and deployment of machine learning models, in particular CNNs, which have improved state-of-the-art performances by a significant margin in a wide range of applications such as audio processing [17], image processing [4] and natural language processing [8]. These successes come with the challenge of exploring a broad search space for hyperparameters and model designs. Therefore, an efficient search is not only a challenge in practical applications for non-experts, but also drives the need for automation in the research community. Under the umbrella term of AutoML, respective methods have already shown to be fruitful in hyperparameter optimization and model selection for traditional machine learning models [11], as well as in optimizing deep neural network architectures for computer vision [9].

Traditional AutoML aims at solving the Combined Algorithm Selection and Hyperparameter (CASH) optimization problem [11] and to build an ensemble of resulting models downstream to achieve the best possible performance with minimum computational and time resources. An intuitive and effective solution is random search. It reaches competitive results compared to more sophisticated algorithms when those are not pretrained [27]. Using the performance of previous runs on a given dataset to guide further model search motivates the idea of evolutionary optimization of the preprocessing and training pipeline [18]. Additionally,

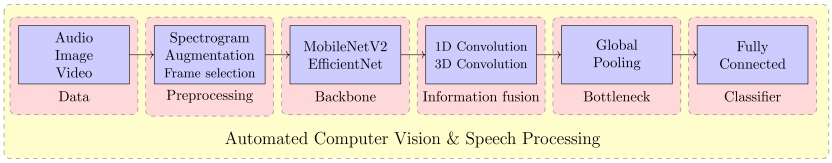

Figure 1. Block diagram of proposed automated audio-visual deep learning approach.

Datasets	Automated Computer Vision 2 Challenge					Final
	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Rank
Winner (kakaobrain)	$\mathbf{0.6277 \pm 0.0628}$	$\mathbf{0.9048 \pm 0.0517}$	0.4076 ± 0.0139	$\mathbf{0.4640 \pm 0.0443}$	0.2091 ± 0.0122	1/20
Runner up (tanglang)	0.6231 ± 0.0449	0.8406 ± 0.0461	$\mathbf{0.4527 \pm 0.0270}$	0.3688 ± 0.0260	$\mathbf{0.2363 \pm 0.0130}$	2/20
Proposed (team_zhaw)	0.5418 ± 0.0340	0.8355 ± 0.0915	0.4110 ± 0.0072	0.3970 ± 0.0298	0.1677 ± 0.0052	8/20

Table 1. Results for the automated computer vision challenge, comparing the proposed idea to other approaches based on the area under the learning curve metric.

model selection and hyperparameter search can also benefit from the information of previous experiments on similar datasets through meta-learning [6].

With respect to automating deep learning, using a unified architecture for the automated design of CNNs in the context of computer vision for image and video data is proposed in [2] as an attempt to overcome the wasteful practice in ML to develop models independently for every new problem and data modality. However, meta-learning [28] and multi-task learning [5] demonstrate that the optimization process can profit even more from using many different tasks in similar modalities. We thus propose here an extension of the multi-modal architecture that encompasses audio data as spectrograms [17]. The resulting generic audio-visual architecture (Figure 1) is appealing due to the following aspects: 1) it extends the state-of-the-art computer vision architectures to different modalities of data besides images; 2) the core information processing block (backbone) can profit from audio-visual information via multi-modal learning when the tasks are related; 3) the information fusion block can learn to combine multi-modal information using attention mechanisms.

The proposed approach aims at finding efficient models for a wide range of tasks on diverse datasets as fast as possible. Therefore, the generic architecture is accompanied with task-specific pre- and post-processing per modality to reduce architecture design burdens in practical applications. An earlier version of the approach competed promisingly in parts of the recent AutoDL 2019[2] challenge. This approach demonstrated a competitive performance compared to state-of-the-art in terms of training speed and generalizing to unseen data. Table 1 presents the performance of the proposed method compared to the winning approaches on unseen datasets for automated computer vision where we achieved the 8^{th} position out of final 20 entries. Similarly, the proposed architecture in Figure 1 achieved the 4^{th} rank amongst 9 entries in the AutoSpeech challenge for general automated audio processing. The proposed method is described in more detail and well investigate with extensive experimental results in [26].

[2] https://autodl.chalearn.org/

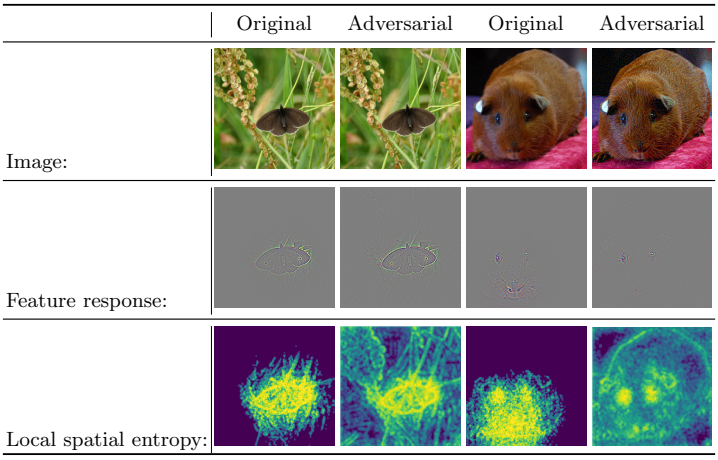

Figure 2. Original and adversarially perturbed images from the ImageNet dataset [7], with attacks being clearly visible in entropy space but not yet in image space.

Method	Dataset	Network	Attack	Performance		
				Recall	Precision	AUC
Uncertainty density estimation [10]	SVHN [14]	LeNet [15]	FGSM	-	-	0.890
Adaptive noise reduction [16]	ImageNet (4 classes)	CaffeNet	DeepFool	0.956	0.911	-
Feature squeezing [29]	ImageNet-1000	VGG19	Several attacks	0.859	0.917	0.942
Statistical analysis [12]	MNIST	Self-designed	FGSM ($\epsilon = 0.3$)	0.999	0.940	-
Feature response (our approach)	ImageNet validation	VGG19	Several attacks	0.979	0.920	0.990

Table 2. Performance of similar adversarial attack detection methods. The Area Under Curve (AUC) is the average value of all attacks in the third and last row (this table is adopted from our previous research presented in [3]).

3 Interpretable and Robust Deep Learning

In this section, we present recent advances towards more interpretable deep neural networks, leading to increased robustness. One key in building trust in ML algorithms is to develop methods that explain the inner workings in a human-interpretable manner. Understanding the reasoning behind decisions of a trained model invariably improve the trust of domain experts. To achieve this for CNNs, several methods have been proposed [1] which can be used as guidelines to modify CNN architectures in order to obtain human interpretable decisions [30]. One of the best understood methods is the analysis of feature response maps computed using e.g. guided backpropagation [22]. Figure 2 illustrates how feature responses of CNNs, computed by guided backpropagation, can be used to visualize the regions where the network focuses at to take a decision.

Such methods to visualize the behavior of CNNs are mostly used to evaluate and compare the decision-making process of networks. Additionally, they also provide key insights to improve model design and robustness. For example, previous research demonstrate [3] that feature response maps are quite informative to detect adversarial attacks [25], which is an essential threat to the robustness

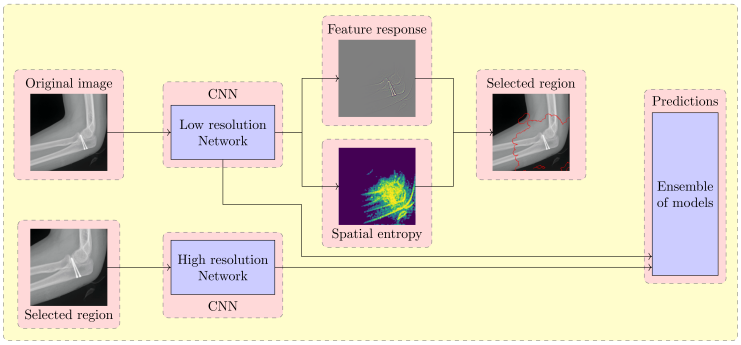

Figure 3. Block diagram for a feature response-based adaptive zooming-in classifier.

and security of deep learning (Figure 2). The feature maps depict the regions in the original image that contribute to the final decision of CNN, while the local spatial entropy images visualize the entropy of feature map activations in every 3×3 image patch. Simple thresholding of the latter yields a veritable detector for otherwise invisibly perturbed adversarial examples. The average local spatial entropy depicted in Figure 2 provides a measure to detect adversarial attacks with competitive results compared to the state-of-the-art (compare Table 2).

Interpreting the decisions of CNNs can be applied beyond ensuring robustness and enabling trust to facilitate novel classifier architecture designs. As a demonstrator, we propose here the following multi-resolution classifier based on two CNN models (Figure 3): Both models have the same input size but operate on the same image in different resolutions. The low-resolution model is trained first for an anomaly detection task based on the original full images. The high-resolution model then learns the finer details of detected anomalies based on a high-resolution crop of the region of interest around the first model's center of feature response. An ensemble of the two models can achieve promising performance in anomaly detection on the MURA dataset [19] of medical images. Guan et al. present an alternative implementation of this idea, thereby improving the performance of thorax disease classification accuracy by using multi-scale information fusion [13].

4 Conclusion

When considering the future impact of deep learning in practical scenarios like medical image processing for diagnosis support, the issue of trust is paramount: Trust of the user/expert in the generated decision, and trust of the developer in the engineering process that currently is often unsystematic and difficult to repeat due to manual "grad student descent". We propose combining these two emerging ideas to address this trustworthiness tango, and present the following corresponding ideas: (a) AutoML to automate the model building process, thus making the vast design space searchable in a systematic manner. Our ad-

dition suggests a unified multi-modal architecture could be trusted for any audio/video/image classification task. (b) Visualizing feature responses of neural networks to give insight into the reasons behind classification results, thus helping concerned parties with the interpretation of a result in addition to providing robustness against adversarial attacks. Our addition shows that such interpretability measures can furthermore be beneficial to build novel classifier architectures by adaptively focusing attention on relevant portions of the input in a user-interpretable manner.

We see potential in further exploring this idea of combining the benefits of interpretability and automation in deep learning. Instead of building manually tweaked model architectures and attempting to interpret them afterward, let an AutoDL system optimize the hyperparameters of a more general architecture with built-in interpretability. This interpretability may also result from designing model-based methods that learn explainable representations [30, 20, 21].

Acknowledgements We are grateful for support by Innosuisse grants 25948.1 PFES-ES "Ada" and 26025.1 PFES-ES "QualitAI".

References

1. Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K.R., Dähne, S., Kindermans, P.J.: iNNvestigate neural networks. JMLR **20**(93), 1–8 (2019)
2. Amirian, M., Rombach, K., Tuggener, L., Schilling, F.P., Stadelmann, T.: Efficient deep CNNs for cross-modal automated computer vision under time and space constraints. In: ECML-PKDD 2019, Würzburg, Germany, pp. 16-19 (2019)
3. Amirian, M., Schwenker, F., Stadelmann, T.: Trace and detect adversarial attacks on cnns using feature response maps. In: IAPR ANNPR. pp. 346–358. Springer (2018)
4. Bianco, S., Cadene, R., Celona, L., Napoletano, P.: Benchmark analysis of representative deep neural network architectures. IEEE Access **6**, 64270–64277 (2018)
5. Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)
6. Chen, Y., Hoffman, M.W., Colmenarejo, S.G., Denil, M., Lillicrap, T.P., Botvinick, M., De Freitas, N.: Learning to learn without gradient descent by gradient descent. In: ICML. pp. 748–756 (2017)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255 (2009)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. pp. 4171–4186 (2019)
9. Elsken, T., Metzen, J.H., Hutter, F.: Neural architecture search: A survey. arXiv preprint arXiv:1808.05377 (2018)
10. Feinman, R., Curtin, R.R., Shintre, S., Gardner, A.B.: Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410 (2017)
11. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and robust automated machine learning. In: NIPS (2015)

12. Grosse, K., Manoharan, P., Papernot, N., Backes, M., McDaniel, P.: On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)
13. Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y.: Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:1801.09927 (2018)
14. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009)
15. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
16. Liang, B., Li, H., Su, M., Li, X., Shi, W., Wang, X.: Detecting adversarial examples in deep networks with adaptive noise reduction. arXiv preprint arXiv:1705.08378 (2017)
17. Lukic, Y., Vogt, C., Dürr, O., Stadelmann, T.: Speaker identification and clustering using convolutional neural networks. In: 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP). pp. 1–6. IEEE (2016)
18. Olson, R.S., Urbanowicz, R.J., Andrews, P.C., Lavender, N.A., Moore, J.H., et al.: Automating biomedical data science through tree-based pipeline optimization. In: European Conference on the Applications of Evolutionary Computation. pp. 123–137. Springer (2016)
19. Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., et al.: MURA: Large dataset for abnormality detection in musculoskeletal radiographs. In: 1st Conference on Medical Imaging with Deep Learning (2018)
20. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: IJCAI. pp. 2662–2670 (2017)
21. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: NIPS. pp. 3856–3866 (2017)
22. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR (workshop track) (2015)
23. Stadelmann, T., Amirian, M., Arabaci, I., Arnold, M., Duivesteijn, G.F., Elezi, I., Geiger, M., Lörwald, S., Meier, B.B., Rombach, K., Tuggener, L.: Deep learning in the wild. In: IAPR ANNPR. pp. 17–38. Springer (2018)
24. Stadelmann, T., Tolkachev, V., Sick, B., Stampfli, J., Dürr, O.: Beyond ImageNet: deep learning in industrial practice. In: Applied Data Science, pp. 205–232. Springer (2019)
25. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. ICLR (2014)
26. Tuggener, L., Amirian, M., Benites, F., von Däniken, P., Gupta, P., Schilling, F.P., Stadelmann, T.: Design patterns for resource-constrained automated deep-learning methods. AI **1**(4), 510–538 (2020)
27. Tuggener, L., Amirian, M., Rombach, K., Lörwald, S., Varlet, A., Westermann, C., Stadelmann, T.: Automated machine learning in practice: state of the art and recent results. In: 6th Swiss Conference on Data Science. pp. 31–36. IEEE (2019)
28. Vanschoren, J.: Meta-Learning, pp. 35–61. Springer (2019)
29. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks (2018)
30. Zhang, Q., Nian Wu, Y., Zhu, S.C.: Interpretable convolutional neural networks. In: CVPR. pp. 8827–8836 (2018)