Neural Architecture Search for deepfake detection

José Nave

Vasco Lopes

João Neves

Universidade da Beira Interior and NOVA LINCS,
Covilhã
Portugal

Abstract

As deepfakes popularity increases, so does their quality. It has become an absolute necessity to be able to distinguish between synthetic and real footage. The research community has dedicated substantial efforts to address this threat, which has led to an explosion in the number of papers related to deepfakes in recent years, and we can see that the most common techniques today involve the use of deep learning and neural networks to detect deepfakes. In spite of the myriad of methods devised to tackle this problem, they fail to generalize to unseen data, raising the need for developing architectures inherently robust to deep fake attacks. For this, Neural Architecture Search (NAS) has started recently to be applied to this problem. This paper reviews the works that have used NAS for deepfake detection. In particular, we explore deepfakes and the processes involved in their creation and detection. It also presents NAS and some fundamental concepts associated with it. In this article, we will explore the research opportunities that arise from the intersection of these two topics.

1 Introduction

Recent advancements in deep learning ignited significant progress over tasks such as image recognition, speech recognition, and machine learning translation. It has been successfully applied to solve various complex problems ranging from big data analytics to computer vision and human-level control.

However, it has also been employed to develop new threats to privacy, democracy, and national security. One of these threats has been materialized in the form of deepfakes. Deepfakes algorithms can be used to create fake images and videos that make the boundary between authentic and synthetic media very thin. Public figures such as politicians and celebrities were the first targets of these attacks since these models required considerable data to train models that could create photo-realistic images and videos. Public figures usually have that amount of data already available. With recent algorithmic and training improvements, less data from the target is required to create synthetic images. Recent methods now allow non-experts to generate synthetic images with just one image of the target. Naturally, these methods have become a considerable concern recently [7, 8, 15, 16].

2 Related Work

2.1 Deepfakes Detection

Deepfakes are digitally altered photos, videos, or audio using deep learning techniques. They are difficult to distinguish from real images [8, 13, 16]. We can divide deepfakes into four categories [2, 15, 16]: Face synthesis consists of generating entirely new human faces that do not exist. Face swapping involves replacing a person's face with another person's. Face attribute consists of modifying some facial attributes, such as adding glasses and altering skin colour. Facial re-enactment, also known as expression swap, consists of transferring facial expression from a source face to a target face and retaining the features and identity of the target face.

Deepfake detection has been customarily assumed to be a binary classification problem. It aims to create a classifier to distinguish between authentic and synthetic images. Deepfakes started to be created manually via traditional visual effects or computer graphics approaches. It is common to use deep learning models to generate synthetic images [14, 15].

Neural networks are effective for deepfake detection, but some problems exist[2, 8, 13, 15, 16]. Most models prioritize accuracy, making them computationally expensive. Another problem is the need for generalization, as they fail in unseen conditions. Deep learning models are prone to adversarial attacks, requiring robust networks [4, 6]. Overfitting can lead to adversarial vulnerability. Innovative techniques are necessary to stay ahead of manipulation methods in this battle between creating and detecting deepfakes.

2.2 Neural Architecture Search

Deep Neural Networks (DNNs) have advanced tasks like image recognition, speech recognition, machine translation, and deepfakes detection [3, 12]. However, designing networks is a trial-and-error process that requires expertise [7, 11]. Neural Architecture Search (NAS) automates network design for a given dataset. It has outpaced manual design on many tasks and is the next step in automating machine learning. NAS allows the discovery of more complex architectures. Designing these networks in a trial-and-error way is tedious and requires architectural engineering skills and domain expertise. Experts use their experience or technical knowledge to create and design a neural network [4, 5, 6, 11, 12, 18].

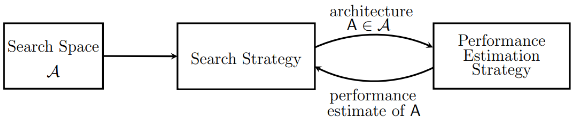

Figure 1: How NAS works [5]

To create a Neural Architecture Search (NAS), three components are needed: a search space, a search strategy, and a performance estimation strategy. The search space consists of a pool of possible operations (layers) and configurations, which outlines the types of architectures that can be designed. The search strategy is responsible for optimizing the NAS algorithm to experiment and find the optimal architecture within the defined search space. The performance estimation strategy is used to estimate the performance of a predicted neural architecture in the search space. In short, the search space defines all the possible candidate neural architectures, the search strategy determines how they are generated, and the performance estimation strategy guides the search process to find optimal models [5, 10, 12, 18]. In Figure 1, we can see how these concepts interact.

There are two categories of search spaces: macro-search and micro-search (also called cell-based search). Micro-search builds cells within a pre-existing architecture, while macro-search develops entire architectures.

Cell-based search spaces are a popular type of search space in neural architecture search (NAS) [12, 18]. In this approach, the network architecture consists of a fixed outer skeleton and searchable cells that make up the microstructure, allowing for faster and more efficient searches [12]. However, human experts must pre-define design choices, limiting the expressiveness of the NAS search space. In cell-based search spaces, the network architecture is divided into two parts: a fixed outer skeleton and a set of searchable cells that comprise the microstructure. Instead of searching for the entire network architecture from scratch, cell-based search spaces propose searching over smaller, modular cells and stacking them in a pre-defined outer skeleton to form the overall architecture [12, 18]. On the other hand, macro search spaces have high representation power but are inefficient to search [12, 18].

Gradient-based methods, such as DARTS [9], are a type of search strategy that relaxes discrete decisions in architecture design to continuous variables, enabling efficient gradient-based optimization. Although differentiable NAS requires little computational resources, it has lower accuracy than other NAS methods. Once a supernet is trained, each architecture can be evaluated by inheriting weights from the corresponding subnet. The scalability and efficiency of supernets are due to a linear increase in computational costs for training [12, 18].

Zero-cost proxies are a popular type of performance estimation strategy that offer an efficient way to estimate the performance of architectures using inexpensive computations or heuristics. These estimators analyze the characteristics or properties of the architectures, such as their design or modelling capabilities, without the need for training until convergence. When combined with other strategies, they can yield excellent results [12, 17, 18].

3 Neural Architecture Search for Deepfake Detection

Exploiting the fact that deepfake detection methods have their performance substantially reduced across different datasets (lack of robustness) and that the architecture design in these methods is usually done manually, which means it is done in a trial-and-error way, Jin et al. [7] decided to use NAS to develop an end-to-end framework capable of designing network architectures without human intervention. They designed a forgery-oriented search space by focusing on Central Difference Convolutions (CDC), which have shown effectiveness for face forgery detection and introduced a novel performance estimation strategy that leads to a selection of more robust architectures achieving competitive results in four benchmark datasets.

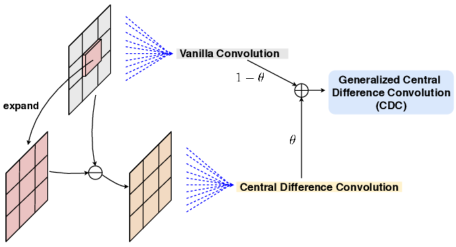
Vanilla Convolution
expand
$1-\theta$
Generalized Central Difference Convolution (CDC)
θ
Central Difference Convolution

Figure 2: Central Difference Convolution [7]

A similar approach [10] also exploits the idea that manually designing neural network architectures is a time-consuming process that involves much prior knowledge that, when incorrect, may deteriorate the model performance. Liu decided to employ NAS to search for an architecture capable of deepfake detection. The authors use NAS to search in a cell-based search space and then employ those cells in a pre-defined architecture. They introduce a strategy of localizing the potentially manipulated region to add robustness to the method. Their method learns two tasks simultaneously, finding the most probable place for the manipulation and differentiating between fake and actual samples.

Another exciting approach [14] employed p-DARTS [1] to search for robust architectures capable of detecting deepfakes. The author concluded that data augmentation techniques can help the model detect deepfake techniques, but a good trade-off must be found to avoid underfitting.

One idea could be to evaluate the impact NAS has on the lack of robustness and efficiency. This could be done by comparing manually designed architectures and those obtained with NAS.

4 Conclusion

As we can see, even though NAS has already been applied to this topic, there is still much to explore. In other fields, NAS already achieves competitive results while discovering lighter networks, spending less time in the process, and requiring much less expertise in the field. Another advantage of NAS is its ability to discover innovative methods, which in the case of deepfake detection is essential. For that reason, we believe NAS can still improve deepfake detection techniques.

5 Acknowledgments

This work is supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT.IP

References

[1] Xin Chen, Lingxi Xie, Jun Wu, and Qi Tian. Progressive DARTS: bridging the optimization gap for NAS in the wild. *CoRR*, abs/1912.10952, 2019. URL `http://arxiv.org/abs/1912.10952`.

[2] Minh Dang and Tan N. Nguyen. Digital face manipulation creation and detection: A systematic review. *Electronics*, 12(16), 2023. ISSN 2079-9292. doi: 10.3390/electronics12163407. URL `https://www.mdpi.com/2079-9292/12/16/3407`.

[3] Shuchao Deng, Zeqiong Lv, Edgar Galvan, and Yanan Sun. Evolutionary neural architecture search for facial expression recognition. *IEEE transactions on emerging topics in computational intelligence*, pages 1–15, 01 2023. doi: 10.1109/tetci.2023.3289974.

[4] Chaitanya Devaguptapu, Devansh Agarwal, Gaurav Mittal, Pulkit Gopalani, and Vineeth N Balasubramanian. On adversarial robustness: A neural architecture search perspective. *Research Archive of Indian Institute of Technology Hyderabad (Indian Institute of Technology Hyderabad)*, 10 2021. doi: 10.1109/iccvw54120.2021.00022.

[5] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21, 2019. URL `http://jmlr.org/papers/v20/18-598.html`.

[6] Minghao Guo, Yuzhe Yang, Rui Xu, Ziwei Liu, and Dahua Lin. When nas meets robustness: In search of robust architectures against adversarial attacks, 06 2020. URL `https://ieeexplore.ieee.org/document/9156305`.

[7] Xiao Jin, Xin-Yue Mu, and Jing Xu. Searching for the fakes: Efficient neural architecture search for general face forgery detection, 2023.

[8] Natalie Krueger, Dr. Mounika Vanamala, and Dr. Rushit Dave. Recent advancements in the field of deepfake detection, 2023.

[9] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. *CoRR*, abs/1806.09055, 2018. URL `http://arxiv.org/abs/1806.09055`.

[10] Ping Liu. Automated deepfake detection. *CoRR*, abs/2106.10705, 2021. URL `https://arxiv.org/abs/2106.10705`.

[11] Yuqiao Liu, Yanan Sun, Bing Xue, Mengjie Zhang, Gary G. Yen, and Kay Chen Tan. A survey on evolutionary neural architecture search. *IEEE Transactions on Neural Networks and Learning Systems*, 34:550–570, 02 2023. doi: 10.1109/TNNLS.2021.3100554. URL `https://ieeexplore.ieee.org/abstract/document/9508774`.

[12] Vasco Lopes. *Improving Neural Architecture Search With Bayesian Optimization and Generalization Mechanisms*. PhD thesis, 06 2023.

[13] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *CoRR*, abs/2004.11138, 2020. URL `https://arxiv.org/abs/2004.11138`.

[14] Jordi Moreno Claver. Neural architecture search for detection of deepfakes. 2020.

[15] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. Deep learning for deepfakes creation and detection. *CoRR*, abs/1909.11573, 2019. URL `http://arxiv.org/abs/1909.11573`.

[16] Rubén Tolosana, Rubén Vera-Rodríguez, Julian Fiérrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *CoRR*, abs/2001.00179, 2020. URL `http://arxiv.org/abs/2001.00179`.

[17] Colin White, Arber Zela, Binxin Ru, Yang Liu, and Frank Hutter. How powerful are performance predictors in neural architecture search. 04 2021.

[18] Colin White, Mahmoud Safari, Rhea Sanjay Sukthanker, Binxin Ru, Thomas Elsken, Arber Zela, Debadeepta Dey, and Frank Hutter. Neural architecture search: Insights from 1000 papers. 01 2023. doi: 10.48550/arxiv.2301.08727.