Stratégies de Compression et d'Optimisation des Modèles d'Intelligence Artificielle

Moez Krichen

Stratégies de Compression et d'Optimisation des Modèles d'Intelligence Artificielle

Moez Krichen

Laboratoire ReDCAD, Université de Sfax, Tunisie
moez.krichen@redcad.org

Résumé. Dans un contexte où la complexité et la taille des modèles d'intelligence artificielle (IA) augmentent de manière exponentielle, l'optimisation et la compression des modèles émergent comme des nécessités cruciales pour maintenir la viabilité des déploiements d'IA. Cet article explore diverses stratégies visant à réduire les exigences computationnelles et de mémoire des modèles d'IA sans compromettre leur performance. Nous examinons l'élagage des modèles, la quantification, la distillation des connaissances, l'approximation de faible rang, la conception d'architecture, la distillation de modèle et les avancées dans AutoML et la recherche d'architecture neuronale. Chacune de ces techniques offre une voie pour alléger les modèles d'IA, facilitant leur mise en œuvre sur des dispositifs avec des ressources limitées et améliorant leur efficacité et scalabilité. Nous discutons également de l'impact de ces stratégies sur la performance des modèles, soulignant comment elles peuvent être utilisées pour créer des solutions d'IA plus accessibles et plus durables.

1 Introduction

Le développement de l'apprentissage automatique (ML) et de l'intelligence artificielle (AI) a entraîné des changements significatifs dans plusieurs secteurs, menant à l'automatisation des emplois, à l'extraction de perspectives à partir de vastes ensembles de données et à la facilitation de processus de prise de décision avancés. L'utilisation des systèmes d'AI est de plus en plus répandue dans de nombreux domaines, englobant l'identification d'images, le traitement du langage naturel, les voitures autonomes et les recommandations personnalisées (89; 27). Néanmoins, la demande pour améliorer l'efficacité de ces systèmes d'AI a émergé comme un champ d'étude crucial, étant donné leur sophistication croissante et leurs besoins en ressources (57; 70).

À l'ère numérique, l'intelligence artificielle (IA) est devenue une pierre angulaire des avancées technologiques, révolutionnant des secteurs allant de la santé et de l'éducation à l'industrie et au divertissement (66; 42; 43; 38; 36). La quête incessante pour améliorer les performances des modèles d'IA a conduit à des architectures de plus en plus complexes et à des modèles de grande taille, engendrant des progrès significatifs dans la précision et l'efficacité des tâches d'IA (40; 75; 5; 72; 14; 53; 71; 2; 37; 44; 6). Cependant, cette tendance vers des modèles plus grands et plus complexes soulève

d'importants défis, notamment en termes de demandes computationnelles et de besoins en mémoire (64; 39; 54; 55; 34; 90; 7; 3; 13; 60; 22; 8; 50; 61). Ces exigences croissantes limitent la capacité à déployer des technologies d'IA sur des dispositifs aux capacités matérielles restreintes, tels que les smartphones, les drones, et les appareils IoT, restreignant ainsi l'accès à l'IA pour une large gamme d'applications potentielles (69; 41; 56; 68; 4; 17; 16; 9; 30).

Dans ce contexte, la compression et l'optimisation des modèles d'IA émergent comme des domaines de recherche essentiels, visant à surmonter ces obstacles en développant des modèles plus légers sans sacrifier leur performance. Les avantages de tels modèles sont multiples : ils permettent une mise en œuvre plus large de l'IA dans des environnements aux ressources limitées, réduisent les coûts associés à l'infrastructure de calcul et d'énergie, et améliorent la réactivité et la fiabilité des applications d'IA en temps réel.

Le rôle de l'efficacité est d'une importance capitale dans la création et la mise en œuvre des systèmes de ML et d'AI (15). Le concept incorpore divers éléments fondamentaux, tels que l'application prudente des ressources informatiques, la réduction de la complexité temporelle, la consommation efficace de l'énergie et l'amélioration globale des performances du système (20; 1). Améliorer l'efficacité ne résulte pas seulement en des solutions d'AI accélérées et économiquement efficaces, mais joue également un rôle dans la promotion de la durabilité et de la scalabilité. Cela, à son tour, facilite la mise en œuvre de systèmes d'AI sur des appareils à ressources limitées et dans des applications étendues.

Cet article se propose de faire le point sur les stratégies actuelles de compression et d'optimisation des modèles d'IA, en se concentrant sur les méthodes d'élagage, de quantification, de distillation des connaissances, d'approximation de faible rang, de conception d'architecture, de distillation de modèle, ainsi que sur les avancées dans le domaine de l'AutoML et de la recherche d'architecture neuronale. En examinant l'efficacité de ces approches, nous visons à mettre en lumière comment elles peuvent contribuer à la création de modèles d'IA plus accessibles et plus efficaces, capables de fonctionner dans une variété de contextes avec des contraintes matérielles. En abordant ces défis, nous ouvrons la voie à une nouvelle génération de technologies d'IA qui sont non seulement puissantes et précises mais également inclusives, durables et adaptées aux besoins de la société moderne.

Avec la croissance continue de la taille et de la complexité des modèles d'IA, il devient impératif de réduire leurs demandes computationnelles et de mémoire tout en maintenant les niveaux de performance (48). Les approches de compression et d'optimisation des modèles abordent cette difficulté en réduisant la taille, la complexité et les exigences de traitement des modèles d'IA tout en préservant leur précision et efficacité (65; 21). Ces stratégies facilitent la mise en œuvre de modèles d'IA sur des dispositifs aux ressources limitées, résultant en une amélioration de la vitesse d'inférence et une augmentation de la scalabilité (11).

2 Élagage

L'objectif de l'élagage des modèles est de réduire les dimensions des modèles d'IA en supprimant les caractéristiques superflues ou relativement moins significatives (83; 52). Cette méthodologie implique d'identifier et de supprimer les connexions, neurones ou couches entières qui contribuent peu à la performance globale du modèle. L'élagage peut être effectué soit pendant la phase d'entraînement, soit comme une action subséquente à la fin de l'entraînement. Pendant le processus d'entraînement, il est courant de remettre à zéro les petits poids ou d'éliminer les connexions basées sur certains critères, tels que l'élagage basé sur l'importance (47; 28). Les stratégies d'élagage post-entraînement (45; 46), telles que l'élagage des poids basé sur l'importance ou l'élagage itératif, sont employées pour évaluer le modèle appris et éliminer les paramètres redondants. L'élagage a pour double objectif de réduire la taille du modèle et d'améliorer la performance de calcul lors de l'inférence en réduisant le nombre de paramètres à évaluer.

3 Quantification

Les stratégies de quantification visent principalement à réduire le niveau de précision des paramètres du modèle, passant souvent de représentations en virgule flottante à des représentations à point fixe (25). La quantification est cruciale pour réduire l'utilisation de la mémoire et les demandes de traitement des modèles d'IA en codant les valeurs numériques avec moins de bits. Des méthodes telles que la quantification uniforme (88; 10) ou la quantification basée sur le regroupement (85; 86) sont employées pour regrouper des valeurs comparables et les représenter à l'aide d'un livre de codes commun. Cette approche vise à minimiser la perte d'informations. Les techniques de formation consciente de la quantification impliquent l'intégration de contraintes de quantification tout au long de la procédure d'entraînement, garantissant la résilience du modèle à la réduction de précision. Récemment, il y a eu des progrès notables dans le domaine, notamment en quantification uniquement entière et en quantification de précision mixte. Ces développements ont atteint un équilibre harmonieux entre la réduction de la précision et l'amélioration de la performance du modèle. En conséquence, ils ont facilité le déploiement efficace des modèles sur du matériel aux capacités de calcul limitées.

4 Distillation des Connaissances

Le processus de distillation des connaissances implique le transfert de connaissances d'un modèle enseignant, caractérisé par sa structure grande et complexe, à un modèle étudiant, conçu pour être plus petit et plus efficace (51; 26). Le modèle enseignant sert de mécanisme directeur en offrant des cibles douces ou des distributions de probabilité au lieu de cibles rigides pendant le processus d'entraînement (18). Le modèle étudiant s'efforce de répliquer les actions du modèle enseignant en réduisant l'écart entre leurs prévisions respectives (79). Le processus de distillation des connaissances permet au

modèle étudiant d'acquérir les informations et capacités de généralisation du modèle enseignant tout en affichant une forme plus condensée et efficace sur le plan computationnel (63). Cette technique s'avère particulièrement avantageuse dans les scénarios où les modèles d'IA doivent être déployés sur des dispositifs aux capacités de calcul limitées ou lorsqu'il est nécessaire d'atteindre une performance semblable à celle d'un ensemble avec un seul modèle (31).

5 Approximation de Faible Rang

L'objectif des approches d'approximation de faible rang est de réduire la complexité et les demandes de calcul des modèles d'IA en approximant leurs matrices de poids avec des matrices de rang inférieur (62). La méthodologie actuelle tire parti de la reconnaissance que les matrices de poids présentent souvent une redondance et peuvent être efficacement approximées en utilisant un ensemble réduit de vecteurs de base. Des méthodes telles que la décomposition en valeurs singulières (SVD) (19; 24), la décomposition tensorielle (73), et la factorisation matricielle (80; 76) sont employées pour décomposer les matrices de poids en composants de rang inférieur. En utilisant des représentations de rang inférieur pour approximer les matrices de poids originales, l'utilisation de la mémoire et les demandes de calcul du modèle sont considérablement réduites tout en maintenant ses propriétés fondamentales et sa performance.

6 Conception d'Architecture

Les stratégies de conception d'architecture mettent principalement l'accent sur le développement de modèles d'IA à la fois plus efficaces et plus légers, dès les premières étapes du processus de conception. Ces méthodologies englobent la réévaluation du cadre du modèle, la mise en œuvre de modifications architecturales, ou l'adoption de concepts de conception innovants pour améliorer l'efficacité. Diverses techniques, telles que l'élagage du réseau pendant la recherche d'architecture (87), la construction de blocs de réseau compacts, l'utilisation de connexions de saut ou de connexions résiduelles (58; 74; 84), et la mise en œuvre de convolutions séparables en profondeur (29; 78; 35), ont le potentiel de résulter en des modèles plus efficaces en termes de paramètres et de ressources de traitement. Le processus de conception d'architecture prend en compte la performance du modèle ainsi que les limitations imposées par l'environnement de déploiement prévu. Cela permet le développement de modèles personnalisés pour répondre aux exigences d'applications spécifiques ou de plateformes matérielles.

7 Distillation de Modèle

La distillation de modèle fait référence à la formation d'un modèle à plus petite échelle pour répliquer la fonctionnalité démontrée par un modèle plus grand et plus complexe (12; 82). Cette approche se distingue de la distillation des connaissances puisqu'elle met l'accent sur la formation d'un modèle autonome plutôt que sur le transfert

de connaissances à partir d'un modèle existant. Tout au long de la procédure de formation, le modèle plus grand adopte la position d'un enseignant, tandis que le modèle plus petit s'efforce de répliquer ses prédictions ou représentations internes. Le processus de distillation de modèle suit une approche similaire à celle de l'apprentissage supervisé traditionnel mais avec les sorties du modèle enseignant utilisées comme valeurs cibles. Cette méthodologie facilite la création de modèles simplifiés qui démontrent une efficacité computationnelle tout en maintenant une performance comparable à des modèles plus complexes.

8 AutoML et Recherche d'Architecture Neuronale

Les méthodologies AutoML et de recherche d'architecture neuronale (NAS) facilitent l'automatisation des procédures de compression et d'optimisation des modèles à travers l'utilisation d'algorithmes qui recherchent des architectures idéales ou effectuent une construction automatisée de modèles (32; 59; 23; 33). Ces méthodologies visent à explorer une vaste gamme de configurations de modèles potentielles, d'hyperparamètres et d'algorithmes d'optimisation pour découvrir les configurations qui sont à la fois hautement efficaces et performantes (77; 81; 67; 49). AutoML et NAS emploient diverses stratégies telles que l'apprentissage par renforcement, les algorithmes évolutifs et l'optimisation basée sur le gradient pour faciliter le processus de recherche. En utilisant des procédures automatisées pour la construction et l'optimisation des modèles, ces techniques facilitent le développement de modèles personnalisés qui adhèrent aux limitations de performance et de ressources prédéterminées. Ceci résulte en des économies substantielles de temps et de travail pour les praticiens en IA.

9 Conclusion

La compression et l'optimisation des modèles d'IA représentent des domaines de recherche vitaux pour répondre aux défis posés par la croissance exponentielle de la taille et de la complexité des modèles. Les techniques examinées dans cet article, y compris l'élagage, la quantification, la distillation des connaissances, l'approximation de faible rang, la conception d'architecture, la distillation de modèle et les avancées en AutoML et NAS, offrent des voies prometteuses pour alléger les modèles d'IA. Ces stratégies non seulement permettent une mise en œuvre efficace sur des dispositifs à ressources limitées mais aussi ouvrent la porte à une IA plus accessible et durable. En continuant à explorer et à améliorer ces techniques, la communauté de recherche peut s'attendre à une évolution significative dans la manière dont les modèles d'IA sont conçus, optimisés et déployés, rendant la technologie d'IA plus intégrée dans notre vie quotidienne sans les contraintes actuelles de matériel et de coût.

Références

[1] Abdiansah Abdiansah and Retantyo Wardoyo. Time complexity analysis of support vector machines (svm) in libsvm. *Int. J. Comput. Appl*, 128(3) :28–34, 2015.

[2] Q Abu Al-Haija and M Krichen. A lightweight in-vehicle alcohol detection using smart sensing and supervised learning. computers 2022, 11, 121, 2022.

[3] Qasem Abu Al-Haija, Moez Krichen, and Wejdan Abu Elhaija. Machine-learning-based darknet traffic detection system for iot applications. *Electronics*, 11(4) :556, 2022.

[4] Qasem Abu Al-Haija and Moez Krichen. Analyzing malware from api call sequences using support vector machines. In *International Conference on Cybersecurity, Cybercrimes, and Smart Emerging Technologies*, pages 27–39. Springer International Publishing Cham, 2022.

[5] Omar Azib Alkhudaydi, Moez Krichen, and Ans D Alghamdi. A deep learning methodology for predicting cybersecurity attacks on the internet of things. *Information*, 14(10) :550, 2023.

[6] Hamoud Alshammari, Karim Gasmi, Ibtihel Ben Ltaifa, Moez Krichen, Lassaad Ben Ammar, and Mahmood A Mahmood. Olive disease classification based on vision transformer and cnn models. *Computational Intelligence and Neuroscience*, 2022, 2022.

[7] Hamoud Alshammari, Karim Gasmi, Moez Krichen, Lassaad Ben Ammar, Mohamed Osman Abdelhadi, Ammar Boukrara, and Mahmood A Mahmood. Optimal deep learning model for olive disease diagnosis based on an adaptive genetic algorithm. *Wireless Communications and Mobile Computing*, 2022 :1–13, 2022.

[8] Hashem Alyami, Wael Alosaimi, Moez Krichen, and Roobaea Alroobaea. Monitoring social distancing using artificial intelligence for fighting covid-19 virus spread. *International Journal of Open Source Software and Processes (IJOSSP)*, 12(3) :48–63, 2021.

[9] Rubby Aworka, Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Jérémie Thouakesseh Zoueu, Franck Kalala Mutombo, Charles Lebon Mberi Kimpolo, Tarik Nahhal, and Moez Krichen. Agricultural decision system based on advanced machine learning models for yield prediction : Case of east african countries. *Smart Agricultural Technology*, 2 :100048, 2022.

[10] Manijeh Bashar, Kanapathippillai Cumanan, Alister G Burr, Hien Quoc Ngo, Erik G Larsson, and Pei Xiao. Energy efficiency of the cell-free massive mimo uplink with optimal uniform quantization. *IEEE Transactions on Green Communications and Networking*, 3(4) :971–987, 2019.

[11] Anthony Berthelier, Thierry Chateau, Stefan Duffner, Christophe Garcia, and Christophe Blanc. Deep model compression and architecture optimization for embedded systems : A survey. *Journal of Signal Processing Systems*, 93 :863–878, 2021.

[12] Max Biggs, Wei Sun, and Markus Ettl. Model distillation for revenue optimization : Interpretable personalized pricing. In *International Conference on Machine Learning*, pages 946–956. PMLR, 2021.

[13] Wadii Boulila, Maha Driss, Eman Alshanqiti, Mohamed Al-Sarem, Faisal Saeed, and Moez Krichen. Weight initialization techniques for deep learning algorithms in remote sensing : Recent trends and future perspectives. *Advances on Smart*

and Soft Computing : Proceedings of ICACIn 2021, pages 477–484, 2022.

[14] Zakaria Boulouard, Mariyam Ouaissa, Mariya Ouaissa, Farhan Siddiqui, Mutiq Almutiq, and Moez Krichen. An integrated artificial intelligence of things environment for river flood prevention. *Sensors*, 22(23) :9485, 2022.

[15] Erik Brynjolfsson and ANDREW Mcafee. Artificial intelligence, for real. *Harvard business review*, 1 :1–31, 2017.

[16] Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Rubby Aworka, Jérémie Thouakesseh Zoueu, Franck Kalala Mutombo, Moez Krichen, and Charles Lebon Mberi Kimpolo. Crops yield prediction based on machine learning models : Case of west african countries. *Smart Agricultural Technology*, 2 :100049, 2022.

[17] Oumaima Chakir, Abdeslam Rehaimi, Yassine Sadqi, Moez Krichen, Gurjot Singh Gaba, Andrei Gurtov, et al. An empirical assessment of ensemble methods and traditional machine learning techniques for web-based attack detection in industry 5.0. *Journal of King Saud University-Computer and Information Sciences*, 35(3) :103–119, 2023.

[18] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11933–11942, 2022.

[19] Julio Cesar Stacchini de Souza, Tatiana Mariano Lessa Assis, and Bikash Chandra Pal. Data compression in smart distribution systems via singular value decomposition. *IEEE transactions on smart grid*, 8(1) :275–284, 2015.

[20] Payal Dhar. The carbon impact of artificial intelligence. *Nat. Mach. Intell.*, 2(8) :423–425, 2020.

[21] Nahid Eddermoug, Abdeljebar Mansour, Mohamed Sadik, Essaid Sabir, and Mohamed Azmi. klm-ppsa v. 1.1 : machine learning-augmented profiling and preventing security attacks in cloud environments. *Annals of Telecommunications*, pages 1–27, 2023.

[22] Mourad Ellouze, Seifeddine Mechti, Moez Krichen, Vinayakumar Ravi, and Lamia Hadrich Belguith. A deep learning approach for detecting the behaviour of people having personality disorders towards covid-19 from twitter. *International Journal of Computational Science and Engineering*, 25(4) :353–366, 2022.

[23] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search : A survey. *The Journal of Machine Learning Research*, 20(1) :1997–2017, 2019.

[24] George W Furnas, Scott Deerwester, Susan T Durnais, Thomas K Landauer, Richard A Harshman, Lynn A Streeter, and Karen E Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *ACM SIGIR Forum*, volume 51, pages 90–105. ACM New York, NY, USA, 2017.

[25] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC,

2022.

[26] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation : A survey. *International Journal of Computer Vision*, 129 :1789–1819, 2021.

[27] Michael Haenlein and Andreas Kaplan. A brief history of artificial intelligence : On the past, present, and future of artificial intelligence. *California management review*, 61(4) :5–14, 2019.

[28] Wenjing Hong, Peng Yang, Yiwen Wang, and Ke Tang. Multi-objective magnitude-based pruning for latency-aware deep neural network compression. In *International Conference on Parallel Problem Solving from Nature*, pages 470–483. Springer, 2020.

[29] Rashidul Hasan Hridoy, Tarek Habib, Ismail Jabiullah, Riazur Rahman, and Farruk Ahmed. Early recognition of betel leaf disease using deep learning with depthwise separable convolutions. In *2021 IEEE region 10 symposium (TENSYMP)*, pages 1–7. IEEE, 2021.

[30] Olfa Hrizi, Karim Gasmi, Ibtihel Ben Ltaifa, Hamoud Alshammari, Hanen Karamti, Moez Krichen, Lassaad Ben Ammar, and Mahmood A Mahmood. Tuberculosis disease diagnosis based on an optimized machine learning model. *Journal of Healthcare Engineering*, 2022, 2022.

[31] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35 :33716–33727, 2022.

[32] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning : methods, systems, challenges*. Springer Nature, 2019.

[33] Yesmina Jaafra, Jean Luc Laurent, Aline Deruyver, and Mohamed Saber Naceur. Reinforcement learning for neural architecture search : A review. *Image and Vision Computing*, 89 :57–66, 2019.

[34] Hajra Khan, Imran Fareed Nizami, Saeed Mian Qaisar, Asad Waqar, Moez Krichen, and Abdulaziz Turki Almaktoom. Analyzing optimal battery sizing in microgrids based on the feature selection and machine learning approaches. *Energies*, 15(21) :7865, 2022.

[35] Zahid Younas Khan and Zhendong Niu. Cnn with depthwise separable convolutions and combined kernels for rating prediction. *Expert Systems with Applications*, 170 :114528, 2021.

[36] Moez Krichen. How artificial intelligence can revolutionize software testing techniques. In *International Conference on Innovations in Bio-Inspired Computing and Applications*, pages 189–198. Springer Nature Switzerland Cham, 2022.

[37] Moez Krichen. Les méthodes formelles sont-elles applicables à l'apprentissage automatique et à l'intelligence artificielle. 2022.

[38] Moez Krichen. Comment l'intelligence artificielle peut révolutionner les techniques de test de logiciels. 2023.

[39] Moez Krichen. Convolutional neural networks : A survey. *Computers*, 12(8) :151, 2023.

[40] Moez Krichen. Deep reinforcement learning. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2023.

[41] Moez Krichen. Generative adversarial networks. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2023.

[42] Moez Krichen. Renforcer la sécurité des contrats intelligents grâce à la puissance de l'intelligence artificielle. 2023.

[43] Moez Krichen. Strengthening the security of smart contracts through the power of artificial intelligence. *Computers*, 12(5) :107, 2023.

[44] Moez Krichen, Alaeddine Mihoub, Mohammed Y Alzahrani, Wilfried Yves Hamilton Adoni, and Tarik Nahhal. Are formal methods applicable to machine learning and artificial intelligence ? In *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pages 48–53. IEEE, 2022.

[45] Woosuk Kwon, Sehoon Kim, Michael W Mahoney, Joseph Hassoun, Kurt Keutzer, and Amir Gholami. A fast post-training pruning framework for transformers. *Advances in Neural Information Processing Systems*, 35 :24101–24116, 2022.

[46] Ivan Lazarevich, Alexander Kozlov, and Nikita Malinin. Post-training deep neural network pruning via layer-wise calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 798–805, 2021.

[47] Guiying Li, Peng Yang, Chao Qian, Richang Hong, and Ke Tang. Stage-wise magnitude-based pruning for recurrent neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[48] Qi Liu, Tao Liu, Zihao Liu, Yanzhi Wang, Yier Jin, and Wujie Wen. Security analysis and enhancement of model compressed deep learning systems under adversarial attacks. In *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 721–726. IEEE, 2018.

[49] 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*, 2021.

[50] Seifeddine Mechti, Moez Krichen, Dhouha Ben Noureddine, and Lamia H Belguith. A decision system for computational authors profiling : From machine learning to deep learning. *Concurrency and Computation : Practice and Experience*, 34(7) :e5985, 2022.

[51] Hefeng Meng, Zhiqiang Lin, Fan Yang, Yonghui Xu, and Lizhen Cui. Knowledge distillation in medical data mining : a survey. In *5th International Conference on Crowd Science and Engineering*, pages 175–182, 2021.

[52] Gaurav Menghani. Efficient deep learning : A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, 55(12) :1–37, 2023.

[53] Saeed Mian Qaisar, Nehal Alyamani, Asad Waqar, and Moez Krichen. Machine

learning with adaptive rate processing for power quality disturbances identification. *SN Computer Science*, 3 :1–6, 2022.

[54] Saeed Mian Qaisar, Dalila Say, Salah Zidi, and Krichen Moez. Automated categorization of multiclass welding defects using the x-ray image augmentation and convolutional neural network. 2023.

[55] Alaeddine Mihoub, Moez Krichen, Mohannad Alswailim, Sami Mahfoudhi, and Riadh Bel Hadj Salah. Road scanner : A road state scanning approach based on machine learning techniques. *Applied Sciences*, 13(2) :683, 2023.

[56] Alaeddine Mihoub, Hosni Snoun, Moez Krichen, Riadh Bel Hadj Salah, and Montassar Kahia. Predicting covid-19 spread level using socio-economic indicators and machine learning techniques. In *2020 first international conference of smart systems and emerging technologies (SMARTTECH)*, pages 128–133. IEEE, 2020.

[57] Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, and Edward Meinert. The effectiveness of artificial intelligence conversational agents in health care : systematic review. *Journal of medical Internet research*, 22(10) :e20346, 2020.

[58] Ricardo Pio Monti, Sina Tootoonian, and Robin Cao. Avoiding degradation in deep feed-forward networks by phasing out skip-connections. In *Artificial Neural Networks and Machine Learning–ICANN 2018 : 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 447–456. Springer, 2018.

[59] Akram Mustafa and Mostafa Rahimi Azghadi. Automated machine learning for healthcare and clinical notes analysis. *Computers*, 10(2) :24, 2021.

[60] Pierre Stanislas Birame Ndong, Wilfried Yves Hamilton Adoni, Tarik Nahhal, Charles Kimpolo, Moez Krichen, Abdeltif EL Byed, Ismail Assayad, and Franck Kalala Mutombo. A face-mask detection system based on deep learning convolutional neural networks. In *Advances on Smart and Soft Computing : Proceedings of ICACIn 2021*, pages 273–283. Springer Singapore Singapore, 2021.

[61] Dhouha Ben Noureddine, Moez Krichen, Seifeddine Mechti, Tarik Nahhal, and Wilfried Yves Hamilton Adoni. An agent-based architecture using deep reinforcement learning for the intelligent internet of things applications. In *Advances on Smart and Soft Computing : Proceedings of ICACIn 2020*, pages 273–283. Springer Singapore, 2021.

[62] Kazuki Osawa, Akira Sekiya, Hiroki Naganuma, and Rio Yokota. Accelerating matrix multiplication in deep learning by using low-rank approximation. In *2017 International Conference on High Performance Computing & Simulation (HPCS)*, pages 186–192. IEEE, 2017.

[63] Dae Young Park, Moon-Hyun Cha, Daesin Kim, Bohyung Han, et al. Learning student-friendly teacher networks for knowledge distillation. *Advances in neural information processing systems*, 34 :13292–13303, 2021.

[64] Saeed Mian Qaisar, Alaeddine Mihoub, Moez Krichen, and Humaira Nisar. Multirate processing with selective subbands and machine learning for efficient arrhythmia classification. *Sensors*, 21(4) :1511, 2021.

[65] Qing Qin, Jie Ren, Jialong Yu, Hai Wang, Ling Gao, Jie Zheng, Yansong Feng, Jianbin Fang, and Zheng Wang. To compress, or not to compress : Characterizing deep learning model compression for embedded inference. In *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, pages 729–736. IEEE, 2018.

[66] Shalli Rani, Ali Kashif Bashir, Moez Krichen, Abdulaziz Alshammari, et al. A low-rank learning based multi-label security solution for industry 5.0 consumers using machine learning classifiers. *IEEE Transactions on Consumer Electronics*, 2023.

[67] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A comprehensive survey of neural architecture search : Challenges and solutions. *ACM Computing Surveys (CSUR)*, 54(4) :1–34, 2021.

[68] Shashidhar Rudregowda, Sudarshan Patil Kulkarni, Gururaj HL, Vinayakumar Ravi, and Moez Krichen. Visual speech recognition for kannada language using vgg16 convolutional neural network. In *Acoustics*, volume 5, pages 343–353. MDPI, 2023.

[69] Dalila Say, Salah Zidi, Saeed Mian Qaisar, and Moez Krichen. Automated categorization of multiclass welding defects using the x-ray image augmentation and convolutional neural network. *Sensors*, 23(14) :6422, 2023.

[70] Terrence J Sejnowski. The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(48) :30033–30038, 2020.

[71] Souhir Sghaier, Moez Krichen, Abir Othman Elfaki, and Qasem Abu Al-Haija. Efficient machine-learning based 3d face identification system under large pose variation. In *International Conference on Computational Collective Intelligence*, pages 273–285. Springer International Publishing Cham, 2022.

[72] R Shashidhar, S Patilkulkarni, Vinayakumar Ravi, HL Gururaj, and Moez Krichen. Audiovisual speech recognition based on a deep convolutional neural network. *Data Science and Management*, 2023.

[73] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on signal processing*, 65(13) :3551–3582, 2017.

[74] Hyeongseok Son and Seungyong Lee. Fast non-blind deconvolution via regularized residual networks with long/short skip-connections. In *2017 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2017.

[75] Wiem Souai, Alaeddine Mihoub, Mounira Tarhouni, Salah Zidi, Moez Krichen, and Sami Mahfoudhi. Predicting at-risk students using the deep learning blstm approach. In *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pages 32–37. IEEE, 2022.

[76] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Björn W

Schuller. A deep matrix factorization method for learning attribute representations. *IEEE transactions on pattern analysis and machine intelligence*, 39(3) :417–429, 2016.

[77] Lorenzo Vaccaro, Giuseppe Sansonetti, and Alessandro Micarelli. An empirical review of automated machine learning. *Computers*, 10(1) :11, 2021.

[78] Chandra Sekhar Vorugunti, Viswanath Pulabaigari, Rama Krishna Sai Subrahmanyam Gorthi, and Prerana Mukherjee. Osvfusenet : online signature verification by feature fusion and depth-wise separable convolution based deep learning. *Neurocomputing*, 409 :157–172, 2020.

[79] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence : A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6) :3048–3068, 2021.

[80] Shiping Wang, Witold Pedrycz, Qingxin Zhu, and William Zhu. Subspace learning for unsupervised feature selection via matrix factorization. *Pattern Recognition*, 48(1) :10–19, 2015.

[81] Jonathan Waring, Charlotta Lindvall, and Renato Umeton. Automated machine learning : Review of the state-of-the-art and opportunities for healthcare. *Artificial intelligence in medicine*, 104 :101822, 2020.

[82] Zach Wood-Doughty, Isabel Cachola, and Mark Dredze. Model distillation for faithful explanations of medical code predictions. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 412–425, 2022.

[83] Sheng Xu, Anran Huang, Lei Chen, and Baochang Zhang. Convolutional neural network pruning : A survey. In *2020 39th Chinese Control Conference (CCC)*, pages 7458–7463. IEEE, 2020.

[84] Jin Yamanaka, Shigesumi Kuwashima, and Takio Kurita. Fast and accurate image super resolution by deep cnn with skip connection and network in network. In *Neural Information Processing : 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24*, pages 217–225. Springer, 2017.

[85] Jianquan Yang, Yulan Zhang, Guopu Zhu, and Sam Kwong. A clustering-based framework for improving the performance of jpeg quantization step estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4) :1661–1672, 2020.

[86] Shuyuan Yang, RuiXia Wu, Min Wang, and Licheng Jiao. Evolutionary clustering based vector quantization and spiht coding for image compression. *Pattern Recognition Letters*, 31(13) :1773–1780, 2010.

[87] Seul-Ki Yeom, Philipp Seegerer, Sebastian Lapuschkin, Alexander Binder, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Pruning by explaining : A novel criterion for deep neural network pruning. *Pattern Recognition*, 115 :107899, 2021.

[88] Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit : Post-training quantization for vision transformers with twin uniform quantization.

In *European Conference on Computer Vision*, pages 191–207. Springer, 2022.

[89] Caiming Zhang and Yang Lu. Study on artificial intelligence : The state of the art and future prospects. *Journal of Industrial Information Integration*, 23 :100224, 2021.

[90] Salah Zidi, Alaeddine Mihoub, Saeed Mian Qaisar, Moez Krichen, and Qasem Abu Al-Haija. Theft detection dataset for benchmarking and machine learning based classification in a smart grid environment. *Journal of King Saud University-Computer and Information Sciences*, 35(1) :13–25, 2023.