WISDOM ACADEMIC

Transactions on Computing Science

ISSN: 3105-0646 | E-ISSN: 3105-0654 *Volume 2 , Issue 1*

Article

Image recognition and classification deep learning models: A journey, problems, and the ways ahead

Yang Liu[1,*]

[1]Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

[*]**Corresponding author:** Yang Liu, pf332371392@163.com.

CITATION

Liu Y. Image recognition and classification deep learning models: A journey, problems, and the ways ahead. Transactions on Computing Science. 2026; Vol 2 (No. 1): 291.

https://doi.org/10.63808/tcs.v2i1.291

ARTICLE INFO

Received: 18 December 2025
Accepted: 24 December 2025
Available online: 21 January 2026

Abstract: Digital imaging tech paired with so much visual material must mean that the field researching image recognition and classification was far ahead of all other artificial intelligence research, development and application fields. Deep learning, Convolutional neural networks shook it up pretty good, reached human levels and above. Recognize an object, identify a face, get another viewpoint on medical imagery, assist autos in driving themselves, all of these. In this paper I have given a comprehensive analysis from multiple aspects on deep learning-based vision understanding model. It takes us from the most basic type of network right up to today's complex CNNs and even up to the revolutionary vision transformers. The study carefully examines some important technical regulations and improved methods of training and the strictness of the requirements for performances. And also, there's digging into deep, long-term issues like endless demand for labeled info, seriously big computing and environment cost from training, difficult to understand "black box" situation with the model, concerns about bad people trying to trick the model on purpose, and heavy ethical problems with being unfair, not treating everybody equally, and watching over who you share your information with. and I see so many case studies of diagnostic imaging in healthcare and e-commerce on visual search from our research today we've seen so many models just being thrown to do incredibly transformative

things but incredibly constrained as well. Finally, the paper concludes with a roadmap of research suggestions for what to tackle next, some exciting new frontiers in data-efficient learning (FE and self-supervised), automated NAS, explainable AI (XAI), building more energy efficient 'green AI' models that are sustainable for long-term use, sound ethical governance. Discussion expands to include the emerging area of multimodal foundation models that combine vision and language and other sensory modes. This area both creates new possibilities and new problems. Putting together all of these results suggests that we need continuous, full-spectrum innovation in all of these areas: novel algorithms, large and sustainable infrastructure, and principled ethics if we really want to break through the problems we currently face and use deep learning to transform every part of society.

Keywords: deep learning; CNNs; image recognition; image classification; computer vision; AI; model architecture; adversarial robustness; ethical AI; vision transformers; self - supervised learning; model efficiency

1. Introduction

Now we are living in a world with no questions where there is lots of data and the picture is taken up a lot of space on the internet. satellite photographs, social media streams, medical scans, and all varieties of industry sensor feeds have seen their visible parsing go from a research curiosity to a critical technical need. Image recognition and image classification these most basic of tasks which identify objects, scenes and patterns within a series of pixels are the foundation of an immensely large network of applications. For example, biometric security systems, autonomous vehicle perception, precision agriculture, automated industrial quality inspection, augmented reality, etc. and visual assistance for the blind.

For decades, traditional computer vision had always followed a pipeline of a handcrafted feature extractors (SIFT, HOG), and classic machine learning classifiers (SVMs). It worked for such constrained environment but failed in real life with huge variation from viewpoint, lighting, occlusions and deformations, which needs vast domain knowledge to design feature. But really the big transformational breakthrough is in 2012 when deep learning comes back and pretty much it's just like it blew everyone else out of the water on ILSVRC 2012 with ImageNet using a model called

Alexnet. The event proved empirically that deep, hierarchically structured CNNs can learn powerful, hierarchical features directly from raw pixels all on their own, and beat all previous ones by a lot.

For decades, traditional computer vision had always followed a pipeline of a handcrafted feature extractors (SIFT, HOG), and classic machine learning classifiers (SVMs). It worked for such constrained environment but failed in real life with huge variation from viewpoint, lighting, occlusions and deformations, which needs vast domain knowledge to design feature. But really the big transformational breakthrough is in 2012 when deep learning comes back and pretty much it's just like it blew everyone else out of the water on ILSVRC 2012 with ImageNet using a model called Alexnet (Krizhevsky et al., 2012). The event proved empirically that deep, hierarchically structured CNNs can learn powerful, hierarchical features directly from raw pixels all on their own, and beat all previous ones by a lot.

In this next decade there was another very exciting era of architectural advancements where we kept on making the architecture result even more accurate, fast and able. In next, we have landmarks: 1.VGG (Simonyan and Zisserman, 2014), 2. Inception (GoogLeNet) (Szegedy et al., 2015), 3. ResNet (He et al., 2016), which added depth, width, multi-scale aspects. The thing that really started working was when someone realized the need for residual connections - this fixed vanishing gradients, and suddenly it was possible to train networks with hundreds, if not thousands, of layers. At that point, with the requirements of deploying with limited resources came architectures like MobileNet (Sandler et al., 2018) and EFFICIENTNet (Tan and Le, 2019). Most recently, it's been revitalized by the success of the Transformer architecture out of NLP, being applied to vision. Vision Transformers (ViTs) see images as a sequence of patches, they use self-attention to capture the dependencies over the entire space, challenge the decades-old inductive biases for Convolution, and achieve state-of-the art on most benchmarks (Dosovitskiy et al., 2021).

Yet still, there has been an astounding amount of progress, and it's all about to come apart, interrelated problems. Deep model's performance still depends on a huge amount of precisely labeled data, and getting this kind of data is costly, and it takes lots of time, and sometimes it's just not possible. State - of - the - art model training requires lots of computation which costs too much money and causes huge environmental damage with too much energy consumption (Schwartz et al., 2020).

Deep neural networks inherent lack of transparency, or to more casually put it, a “black box problem”, causes a weaking of faith and stops larger use with important groups like healthcare and finance, these types need to know how and why the decisions in court were made, this is as important as the decision itself. and they’re also very easily fooled by adversarial attacks (Goodfellow et al., 2014) - tiny changes to the input that will completely change what the model thinks it is, even though we made almost no change! This is really scary for safety. And then society adopts them, so that’s like the ethical question of what biases in the data training get even worse (Buolamwini and Gebru, 2018); the potential for mass surveillance and loss of privacy, and accountability and fairness and then as well when you think of putting vision models on top of large language models, putting them into large multilingual/multimodal models (Bommasani et al., 2021; Liang et al., 2023), do these models align with each other? Are there ways to control them? How do errors propagate differently depending on what input is given?

This paper is aiming for a comprehensive, thorough, and very critical analysis of those deep learnings specially crafted for image recognition and classification. The first move towards this is a deep dive on how the area has historically developed - from very simple notions about neurons, through to the more complex structures we see now that have replaced them, e.g. from conv nets (classical) to vision transformers to hybrids, and also involving different training techniques. So that includes looking at different sorts of algorithms to choose, like SGD versus AdamW; different ways to regularize things, like Dropout, BatchNorm, Stochastic Depth etc. Losses that are custom-made for seeing tasks would go into this group as well. On top of all that technical basis work, we’re going to dig very deep into every kind of challenge stopping our current progress -- things like how much data is needed, how powerful the computers need to be, whether or not it makes sense for people who might use these tools, dangers if someone intentionally messes with models’ inputs, ethical stuff related to trusting AI’s outputs in day-to-day life, all those sorts of issues will fit under bigger umbrella problems.

Secondly, we will also be showing some detailed cases such as those being used for medical diagnostics (Gulshan et al., 2016) and more commercial visual searches, so that we will finally start to move away from theory, and try to see how well they do in the real world, trying to model and handle these real, complicated and messy situations. {A small third short on autonomous driving perception showing what a

chaotic and dangerous situation can be.} Showing what a gap there is between lab and deployed scenarios. Finally, with all of this historical literature, problems, and case studies together this paper will summarize the findings into a roadmap that looks forward into what the next steps of future R&D is. It would emphasize the urgency of the equilibrated advancement and full-fledged progress, needing a balance between the improvements of the main-algorithmic aspect and the computational efficiency, model transparency and understandability, security defense from danger, and morality ahead of time. and the end goal in a sense is pushing the industry to build systems that can be much more capable, better, more trustworthy and more socially responsible systems.

2. The Evolution and Core Architectures of Deep Model

Deep learning for vision had always been a saga of continuous building of architecture, each time the leap goes over the constrains of previous versions and extends the boundary of learnable visual information computationally.

2.1. Pioneering Era: LeNet and CNNs Dawn

The beginning of CNNs today, may be traced all the way back to the LeNet-5 which developed by the Yann LeCun and his co-workers late 1990s for practical recognition of Handwritten Numbers (LeCun et al., 1998). LeNet-5 has set up an important architectural outline that can still be seen today. The convolutional layer is used for spatial local feature extraction, and the sub-mapping layer is used for spatial subsampling and translation. Finally, it uses a fully connected layer to classify features. The convolutional layers of LeNet are filter learning, and the parameters in the same spatial position are the same, which is very different from the hand-tuned one. Although constrained by the computational power (CPU based) and size of datasets that it could use, LeNet-5 achieving success on MNIST showed that it was possible to train a multi-layer network using backpropagation on spatial data and set the important basis for everything else to follow. Local connectivity and shared weights, central to the CNN paradigm, are inspired from biology (visual cortex), and very powerful inductive biases for grid-like data such as images.

2.2. The Modern Revolution: AlexNet, VGG & Inception Family

And the field got a real renaissance with AlexNet in 2012 (Krizhevsky et al., 2012). made it possible thanks to certain breakthroughs like how ReLU solved the vanishing gradient and sped up the training procedure, using dropout as an easy yet extremely useful type of regularization that keeps overfitting from happening, lastly, having Graphics Processing Units which we could use for massive parallelism and thus be able to train deep networks in a practical way. AlexNet's bigger, bolder architecture compared to LeNet-5 hinted that scale might lead to better results.

The idea is and as a case study the VGG network makes this idea resound (Simonyan and Zisserman, 2014), is that you make a network really, really deep using super skinny (3x3) convolutions to do it, and what you end up getting is a fantastic and amazing network to start with. Its uniform modularity is simple and deep, but vgg's cost grows quadratically with depth due to massive full connected layers, which highlights an early accuracy - efficiency trade-off. Meanwhile, google net (inception V1) (Szegedy et al., 2015). Instead of simply adding more layers, it created a "network in a network", that did convolutions in different sizes (1x1, 3x3, 5x5) as separate layer modules. And crucially, it used 1x1 convolutions for dimensional reduction before more costly operations to make the network both wider and computationally cheaper. later versions (inception v2/v3, inception-resent) used the batch normalization as well as the residual connection, which improved the architecture. Batch normalization stabilized the distribution of inputs going into each layer which allowed for much larger learning rate, and is now in nearly all modern architecture.

2.3. Depth Breakthrough: ResNet, DenseNet, and More

Using VGG-style stacking to push networks deeper created problems; as we made the network deeper, it eventually saturated and then began falling off sharply. The problem isn't just overfitting, or vanishing gradients, it's the result of making plain networks very deep which is an optimization question. The mapping the layers have to learn also becomes more complicated making it even harder for the network to be able to approximate identity mappings when they are encountered if we keep applying non-linearizes. In 2015 ResNet (Residual Network) gave the first elegant

and influential way to solve it, by doing Residual Learning (He et al., 2016). ResNet used shortcut connections that do identity mapping (which people usually refer to as "skip connections") and restated the learning objective. Instead of learning a stacked set of layers which want to directly learn the desired underlying map, it learns a residual function such that the original mapping is. This simple re-statement has huge implications: it makes it easier to learn, because if the identity is optimal, the network can instead force the residual function to 0, rather than fitting an identity function with lots of nonlinearities. It can greatly alleviate the vanishing gradient problem when doing backpropagation. And Gradients can pass through skipping connections, which is why ResNets of 100s or 1000s of layers like ResNet-152 can be successfully trained. Residual Learning became a cute idea that rapidly turned into something essential, the default building blocks of just about any modern deep network. {Variations like ResNeXt started sprinkling even more flavor in with a twist on yet another architectural dimension: the number of parallel transformations or 'cardinality' sitting alongside the other choices of depth and width. ResNeXt achieved improved model capacity and accuracy using grouped convolutions within the residual blocks, without the increased cost in the same way. This showed us that multi-branch transformations were able to capture lots of useful features successfully.

Taking the help of gradients flow, we will further reuse our ideas and run with them but this time with a stronger connection known as DenseNet (Huang et al., 2017). A DenseBlock has its layers take in all the feature maps from all previous layers and give all its feature maps to all future layers. This is dense connectivity making short paths from earlier to later layers, lots of feature re-use in the network. so, things get a lot easier in terms of gradient flowing and so vanishing gradient problems don't matter as much anymore, and you usually tend to get models that are a lot more parameter efficient. And so each layer can actually be very skinny because at each level, we can access something which is already been figured out. Can also reach good performance with fewer parameters as much as with Resnet-style model, but it usually needs to be the same accuracy as well as less computing cost, calculated based on FLOPs. But the memory bandwidth required to concatenate all the previous feature maps is extremely challenging to implement, especially with high-resolution input or limited memory environment. This was the impetus for investigating memory efficient approaches like channel-wise concept or conditionals or partial dense. To try to keep many benefits and benefits of dense connection but save on memory.} The

impact of DenseNet shows the wider trend towards networks which provide better gradient flow and parameter efficiency, a trend which persists as current architecture search and model scaling.

2.4. Efficiency Drive: Mobile-friendly architecting and neural expanding

Deep learning application is slowly transferring from cloud server to the edge device - phones, drones, IoT sensors, so there comes to be more demands for lightweight, low-latency models. The main obstacle here is keeping recognition right when using little computer energy, memory, and power. In this context, MobileNet Series is an epoch-making attempt for systematically introducing depth-wise separable convolution as a main component (Sandler et al., 2018). This op is a standard conv op but split into two ops: first do a depth wise conv which does a depth wise conv of each channel of the input with a filter for spatial convolutions. It does a pointwise conv which is just another set of 1x1 convolutions to mix the channels up again. And this factorization also greatly reduced computational cost and the number of parameters. For example, 3x3 standard convolution is decreased by about an order of magnitude. This is at an ok balance between fast and accurate. MobileNetV2 made further innovation based on this and put forward the inverted residual blocks and the linear bottlenecks. First, increase the dimensionality in a low-dimensional space via a point-wise convolution operation, followed by a depth wise convolution operation in a higher-dimensional space for rich information extraction, and then conduct a reduction operation in a linear manner. The structure guarantees to preserve the important feature after go through a low dimension bottle nick, which give the model much more express power but efficient, also one of the baseline networks for the mobile vision task.

Apart from the approach through convolutional factorization, there exist some other kinds of efficient architectural designs. ShuffleNet uses the channel shuffling operation to shuffle the feature information between different channels after the group convolution operation. Such a method can also eliminate the information flow obstruction brought by the group convolution, which results in good fusion of global features while saving the computational cost. More and more applications have complex and different requirements. It is impossible for people to design the structure

of a network well, so NAS was born. NAS makes architecture designing a hunt for the ideal network structure by using reinforcement learning, evolutionary functions, or gradient techniques to find automatically within a large number of architecture areas the best network structure within some limitations. It is the start of a new age of automated model development.

And this trend of automating designs and EfficientNeT had their big break! (Tan and Le, 2019). Its most significant contribution is a systematic process to scale up the neural network. To my knowledge, for the first time, we have shown that to increase the network, its width (channel), depth (layer), and size of input image all need to be scaled by the same number instead of changing one factor after the other as was done prior. This kind of compound scaling strategy has greatly improved the model's position on the accuracy --efficiency tradeoff line. This gives a scientific basis for designing models of different scales with better efficiency. The next thing is a dynamic change and easy adaptation. For example, once for all networks and super net training attempt to separate architecture search and training; Train a huge super-net which consists of a lot of sub-networks, extract different sub-networks for different hardware platforms (mobile phone or server) / resources without rebinding. Also, can use to deploy tailor-made efficient model to different edge device, and I believe it will be a good direction of future efficient deep learning.

2.5. Paradigm Shift: Vision Transformers & beyond

The biggest recent shift in the architectural space for computers in vision was Vision transformers (ViTs) (Dosovitskiy et al., 2021): ViTs are diverging completely from the convolutional inductive bias that had long defined the field --- translation equivariance, local receptive fields, etc. --- and are completely reframing how we think about images. How do people look at images. It is as series of patches of fixed size, non-overlapping, linearly embedded into high dimensional vector. In order to keep the important spatial information which is lost during such serialization, we add learnable/fixed positional embedding to the patch embedding. And all of these sequences of the vectors get sent to the normal Transformer encoder, which relies totally on attention. Self-attention makes it so that each patch can pay attention to all patches in image. So it can instantly pick up the long range dependence and the full context right from the first layer. , which is very different from a CNN gradually building its larger receptive field out of stacked small ones. And this paradigm shift

gave ViTs an extremely powerful, a very comprehensive understanding of how images are composed, but it came at a cost---lack of innate spatial priors, so it was something like having to learn every single pixel and visual structure from scratch, so these models require a massive amount of training data.

Early ViTs looked promising but showed a big need for very big sets to train, things like the JFT-300M with piles of hundreds of millions of pictures, to do better than good CNNs on ImageNet and other tests. and these were all not easy to find and to use: After that, the following studies tackled the issue swiftly; models like DeiT that show that we can actually do really well and compete with the extra data DeiT does this by using a fancy kind of knowledge sharing called “knowledge distillation”. A very powerful type of neural network, CNN, acts like a teacher, and it guides the student, which is another type of neural network called ViT to learn, passing on some of its own knowledge. Parallel architectural improvements intended to benefit the effectiveness of the computation and make the ViTs to suit the sort of dense predictive task object detection and segmentation.

Swin was a breakthrough variant that returned a hierarchical feature pyramid like CNNs by merging more and more patches at deeper layers (Liu et al., 2021). Critically, it swapped the global self-attention for shifted window based self-attention, where attention is computed within small windows, and cross window connections are achieved through window-shifting in alternating layers. It would be like you have this like this like something very big but then all of a sudden it got way much easier and way much cheaper (less computation) for you to have bigger images than before and in fact this restores the wonderful property known as translation equivariance, which is this great property that makes a lot of these different kinds of viTs much better and much more scalable able to deal with a whole bunch of different kinds of different vision problems out there.

the fast-growing of ViTs makes it a frontier which combines the strength and benefits of these two paradigms, and makes further development and research on even more efficient and data-friendly training method. One big direction is creating hybrid models that cleverly put convolutional layers with Transformer blocks together. ConvNeXt, CoAtNet and other such. And they might toss in some depth wise convs, maybe stage things out, and add in other kinds of improved CNN parts along with the self-attention, in an attempt to blend in CNN’s strong local feats pulling and special prior work with the big context reach of Transformers. Many of these hybrids are

Pareto efficient w.r.t. accuracy, robustness and cost. On the other side, self-supervised pre-training also makes a breakthrough at this time, reducing the dependence of ViTs on labeled data. Masked Autoencoder (MAE) framework is quite strong (He et al., 2022). Following the idea of masking language modeling from the NLP domain, MAE will mask 75% of patches randomly and expect the model to reconstruct the masked-out pixels based upon the seen pixels. This pretraining objective allows the model to learn a good all-around semantic understanding of how objects are structured and what is in an image, so that it can do very well on other tasks using images without having to retrain it again. put together, these advances with hybrid architecture design and self-supervised learning are starting to make transformers more than simply an alternative and competitor to CNNs, they have become a powerful complement and sometimes superior approach evolving alongside CNNs to define the next generation of visual intelligence systems.

3. Critical Challenges in Current Models

The big success of deep learning comes with big problems that makes deep learning less reliable for everyone.

3.1. Data Dependency and Annotation Bottleneck

Deep learning models can be real data hogs. To even get kind of state-of-the-art performance, it could take a whole bunch of pictures that a person has looked at and tagged to do something like a medical imaging type of thing or finding flaws in something like making a machine. Sometimes, we have to look at photos that have been labelled more than a hundred thousand photos! annotation process is too expensive and take time and interannotator is also too variable. This makes it difficult for new apps to get in on situations where not much of what people have shared is known or where folks don’t know too much about the stuff (rare sicknesses, unique ways of making things) But doing data augmentations (crop, rotate, color jitter) or semi supervised learning or just using synthetic images from gan/Diffusion model (Ramesh et al., 2022), it just gives us some temporary respite but does not solve the problem of distribution shift and the reality gap. Depend on huge labeled datasets in the field will lead to the “rich get richer” situation. Only the highly fundable

companies with access to large amounts of proprietary datasets can afford to build better competition. This is even worse in tasks that have to do with understanding something over time (like analyzing what's happening in video), or understanding something in 3D space (like what's happening around a robot in its environment). The amount of work it will take to label something and the size of the data set required both get much worse. More so, the need for more datasets can result in the scrapping of web pages without any concern of legality or ethics and not regarding copyright or any consen.t

3.2. Calculating cost, environment, and unequal access

Train modern large models like Large ViTs & Multimodal model. For some training, the energy used would be enough for many cars to last their whole lives. That's very troublesome when trying to think about how green it is for AI research (Schwartz et al., 2020). And for using cloud computing or having your own special-purpose hardware cluster, i.e., NVIDIA DGX and Google TPUs, is very expensive. So big an economic wall makes it so a few big tech companies and just some of the most fancy and well-moneyed schools can take all the lead on cool AI. This would sideline smaller firms, research labs in less-prosperous parts, and individual researchers, and stifle diversity of thoughts and fairness of access to the rewards of AI innovation. To keep making bigger models no matter if it's efficient or not -- sometimes also known as "red AI" -- is becoming less appealing, and a subtler but key one is that the cost of training vs inference. Very compact model can be cheap to run but can be very expensive to create and train using NAS. Lifecycles (both) is needed for environmental impact to be complete. And also, the location of the data centers physically and whether its power grid is green will also have an impact on the true environmental impact.

3.3. Black box problem and interpretability crisis

Our Deep Neural Networks have inner works we can't see as they're black box It's because its high dimension, nonlinear, and have many parameters (in the orders of millions): If we have something like this, and it comes up with the wrong answer, let's say, marking a cancerous tumor as harmless, missing someone walking by in a car's self-driving system, wrongly saying no to someone borrowing money - it becomes

really hard to find where the logic went wrong, to see what tiny parts made the mistake happen. The fact that we cannot ask why a model decided something is an enormous harm to the people, the clinicians and clinicians, the regulators, and people who are in the public at large. Now it's difficult for us to be responsible. And so it seems that we find ourselves in an ironic spot - where some of the strongest AIs are just being openly used, adopted and utilized by the public and the industry in non-regulated areas such as content moderation, and where similar technology may be entirely avoided and outright banned from other areas that are much more important, like healthcare and finance, and criminal justice, where the "why" is often just as important, if not more so, than what was decided to do.

In response, the field of Explainable AI(XAI) has created post-hoc explanations to make models understandable like gradient-weighted class activation mapping (Grad - CAM), which produce a visual heatmap indicating the parts of an image which were most important to the model's output classification. Local Interpretable Model-agnostic Explanations (LIME) approximates the complex model in the area around a specific data point with a different simple, interpretable model, such as a linear model, to explain each prediction. SHapley Additive exPlanations (SHAP) is based on game theory and is based on assigning an importance value for each feature relative to a specific prediction. These techniques help but only give an approximation which is only indirect reasonings of the model. Their explanations can be unstable (slightly different for similar inputs), incomplete (miss important reasoning chains), and even wrong (e.g., post-hoc on the model instead of explaining the rationale the model used to arrive at its decision). Like perhaps it would show a salient map highlighting watermarks or background texture related to a class in training data instead of what is really important. Gives an illusion of understanding. How "faithful" we can be to these explanations is an open question.

Therefore, while these post-hoc tools may provide useful intermediary answers in practice, the more formidable challenge that XAI aims to take on over the long term, is that we can be given interpretable models, or interpretable architectures from the ground up: reasoning is both transparent, intelligible, and included as a part of the model. Which means going from explaining black boxes to explaining glass boxes, where the decision structure is constructed using meaningful semantically concepts, logical rules, or causes {such as Concept Bottleneck Models (CBMs) that force the model to predict a set of human understandable concepts such as "has stripes, " "is

orange" before making a final prediction, allowing interventions at the concept level.} Prototype based networks learn parts of the training images which can be compared with new inputs to make a decision. Achieving such a shift is still a big challenge and we'll have to make fundamental change in our model architecture and how we train them and evaluate them so they will be able to be as accurate but at some level computationally efficient and can be explained in an interpretable way by humans.

3.4. Prone to Adversarial Attacks and Security Risks

Deep learning models are surprisingly frail: it's easy to fool them with weird examples --- small, tricky changes to an image that no one would notice, but that cause a deep net to predict something totally wrong, with huge confidence (Goodfellow et al., 2014). These kinds of attack can hurt us to live. Autonomous Vehicles: Stop sign if there was a small change could be mistaken for speed limit(sign) In the domain of facial recognition system, adversaries with adversarial makeups or glasses could be granted access. medical imagery, malignant noise is hiding a tumor. Defense such as, Adversarial Training (train on adversarial example) and Input Preprocessing are computationally expensive and can be circumvented by adaptive adversaries. Robust against a range of attacks is an arms race in AI security: {beyond digital attacks, physical-world adversarial examples like stickers on road signs to trick the system or clothing patterns that do the same are proof that digital isn't the only realm of concern. Backdoor attacks are also a danger for pre-trained models, in which a model is given a backdoor during its training so that it will normally function except for one particular pattern. The field of adversarial robustness seems to be more linked with a broader notion of out-of-distribution generalizability and uncertainty estimation.}

3.5. Ethics--Bias Algorithm Issues, Fairness Issues, Damage to Privacy Rights

Machine learning models learn from data, and that data has biases in it from history and society existing in that data. Many audits have found commercial face analysis and recognition systems to be orders of magnitude more problematic for women, specifically for women with darker skin, and also for various other

marginalized groups (Buolamwini and Gebru, 2018). This kind of bias actually makes things worse if they're going to be using it to do something as straightforward as hiring someone, or deciding whether to give someone a loan, and if they are using this biased system. Secondly, the widespread use of Image Recognition by both public and private surveillance, i.e., 'smart cities', and employers can lead to the fact that we become used to living under surveillance, thereby infringing upon the civil rights and the loss of privacy for the general public, which cannot be undone Another big issue which has been mentioned is 'Representational harms', some group is likely just to be wrongfully represented or completely missing from a gen model, or train set. Text-to Image models like would default to pictures of CEOs being men or nurses being women (Ramesh et al., 2022). These issues go beyond technical fixes, they need interdisciplinary work, a complete algorithm check, transparency in results, and there should be laws and rules controlling how we can use AI ethically. like federated learning, differential privacy etc., they need to be trained with decentralized data which leads to lower performance and more complicated system design.

4. Performance evaluation & case studies

4.1. Benchmark Performance and Metrics

Model performance is largely about using the big, standard testbed datasets, which are public. ImageNet (1.2M images,1000 classes), best general image recognition now. Object detection & segmentation usage: MS COCO, PASCAL VOC In autonomous driving research the datasets that are used are Cityscapes for semantic segmentation and BDD100K.

Key evaluation metrics include:

Classification: Top-1 Accuracy, Top-5 Accuracy, Precision, Recall, F1 Score

Object Detection mAP: Mean average precision over IoU and class.

Segmentation: Mean IoU, Pixel Accuracy.

Efficiency: Inference Latency(ms/image)， Throughput(images/s)， Params, FLOPs

While a leaderboard performance on these benchmarks looks good for research progress, it doesn't always get us to the real world with messy lighting conditions, occlusions, new object types, and domain shift (training during the day, running at

night). Emphasize Proper Robustness Testing Under Distributional Shift and Real-World Stress {Some newer benchmark like image net - c (corrupted image), image net - r (different rendition), object net (diff viewpoint/bg) has been built to test the robustness of the model for some common corruptions, as well as change in style or background}

4.2. Case Study 1: Medical image analysis for DR detection

Clinical Challenge: Diabetic retinopathy, one of the biggies for having no more eyesight. Routine screening of retinal fundus photograph for early detection needs very well-trained ophthalmologist, therefore resource is poor setting bottle neck.

Deep Learning approach: Train a custom CNN model like ResNet or DenseNet on a large dataset of fundus images that are graded according to their DR severity (no DR, mild, moderate, severe, or proliferative) by experts. The transfer learning from ImageNet is common. The model does a classification task, giving out a severity grade or a referral recommendation. more sophisticated ones use U-Net-like segmentation models first to find out the lesions (micro aneurysms, hemorrhages) and only then to perform classification, which makes it more understandable: Ensemble with lots of models and test-time-augmentation are two simple ways to make the final result better.

Outcome and Impact: Work done by Gulshan et al. (2016) and others have demonstrated that deep learning models are able to perform as well as, if not better than, humans in sensitivity and specificity for retrospectively studied cases. Commercial and academic interest has swelled for AI-assisted diagnosis to augment the clinician's eye and multiply the reach and uniformity of screening: Regulatory nods for AI-based DR-screening devices like the EU and USA have granted is an essential step towards clinical implementation.

Limitations and Hurdles: 1. Data scarcity & quality: Requiring huge amounts of diverse, well-labeled datasets, hard to achieve. {Labels are difficult because graders disagree on the messiness of labels caused by ambiguity in some medical images} 2. Black Box Liability: Can't explain to a medicolegal why a model flagged an image. 3 Domain shift: It fails on images taken on different cameras, different people, different clinical protocols, which requires costly recalibration. 4 Clinical integrations: must effortlessly play with PACS system, engineers' and regulators' nightmare. {Who's the

AI? Triage，second reader，primary screener? To have it correctly defined, a specific pathway is required for one to use the clinical pathway safely and successfully.

4.3. Case Study 2 – Visual Search and Recommendations: E-commerce

Business Challenge: Help someone looking for an online product that has no text describing it, just a picture (search by picture) and help people find products they will like and engage with and buy them.

Deep Learning Approach: Core technology is metric learning or representation learning. We train a deep CNN like ResNet - 50, or EfficientNet as a feature extractor, which maps both the query image (user upload) as well as all the catalog product images into the same high - dimensional embedding space. Train by triplet loss or contrast loss (Chen et al., 2020) so that visually similar products are close in embedding (positives), dissimilar products are far apart (negatives). At Inference time we find our nearest neighbor or neighbors to the query embedding within this space. {Recent system has been more multimodal, fusing image embeddings with text embeddings from product titles and descriptions using a model such as CLIP (Ramesh et al., 2022) to give better results, on more ambiguous queries.}

Outcomes & Impact: Pinterest (Lens), Amazon (StyleSnap), Alibaba, Google etc. has implemented very effective visual search engines. It significantly enriches users' experience, it helps users find products, it boosts sales conversion rate and it helps find fakes. and find images and copyrights (find fakes)

Intraclass variance, interclass similarity The same shirt can be many different colors/patterns (high intra-class variance) and many different black handbags looks very similar (high inter-class similarity).2. Query and catalog mismatch: The user's uploaded photos can be messy and low-qualified, containing objects from weird angles only showing parts of the object. Catalog photos are clean and well-lit. The model needs to be invariant to the nasty background and viewing angle, still. 3) Scale and Freshness: The catalog images could be in the billions and constantly changing, so you'd need super scalable, efficient nearest neighbor search infrastructure like FAISS/HNSW {And keep the embeding index in near real time as the catalog changes - that's one huge engineering problem too}

4.4. Brief Case Study 3: Perception for Autonomous Driving

Operation: An AV must “see” and make sense of an ever-changing world, complex and dynamic and safety-critical, in 3D with cameras, lidar and radar close to real-time. task is car/pedestrian/cyclist detection, road/sidewalk semantic segmentation and lane/depth estimation

Deep Learning, Modern AV stacks use deep, multiple tasks / multiple sensors neural networks Architectures such as BEV Transformers that use information from lots of cameras to make one big 3D view. The tempoal info from video are usually done with 3d convs, or recur layers the system needs to give output with very low latency and great dependability.

outcomes and limitations: {Deep learning has come an extremely long way from only being able to do closed tests and only being able to do some public road trip. however, this example highlights many extremes of the key problems, the “long-tail” of low probability but high consequence scenarios – a pedestrian carrying something unusually large or unusual, the worst adversarial effects – glare, heavy rain, deep snow, the need for interpretability and verifiability for robustness, and the massive ethical/ liability consequences if it fails. It says that even getting image recognition right isn’t an AI problem, as a standalone system, it has to be integrated with some other system which is stable, secure, and accountable – that’s where reality starts.

5. Strategic suggestions for future development

To sidestep every single pointed, critical issue and move onto something bolder and just, one would have to go a long and roundabout way.

5.1. Being the first ones to use data efficiently and without supervision

Be free from large-scale labeled datasets is very important.

Advance Few-shot and meta-learning: Develop models which can promptly adjust to fresh visual notions or jobs through only a few instances by making use of previous knowledge. Research should instead focus on improving the generalization

and robustness of these methods beyond narrow benchmarks, {and studying better ways to initialize them, more expressive ways of adapting it, and ways to learn from instructions or natural language descriptions in a zero-shot way.}

Scaling self-supervised learning: SSL methods like SimCLR/MoCo, MAE/BeiT show that it is viable for models to learn strong general-purpose visual representations using just unlabeled data. Further work will be done on speeding up SSL pretraining and transferring representations to down-streams with minimal fine-tuning. One direction will be to unify different SSL pretext tasks and figure out what makes a good pretraining task from a theoretical standpoint. SSL with small amounts of labeled data in a semi-supervised setting holds much promise

Improve Synthetic Data Realism and Utility: Better GANs & Diffusions & Neural Radiance Fields that make photorealistic, diverse and plausibly-physical synthetic data automatically. Research needs to focus on adapting technique to close the 'sim-to-real' gap, how can we train the model in synthesize world but when it uses in real world it works better? That's mean building better metric to measure the quality and diversity of synthesize data and better way to synthesize the data that focus on hard and rare case.

5.2. Promoting efficiency of the model and the "Green AI" movement

We need to change our target from just being accurate to being as accurate as possible for a given amount of computational energy or effort we want to use.

Democratize neural architecture search (NAS): Creating more effective, accessible and ecologically-sustainable NAS algorithms that can automatically find the most suited architecture for a specific set of hardware limitations and task requirements without the huge computing resource requirement NAS methods of the previous generation like weight sharing NAS, predictor NAS, and taking advantage of low fidelity performance predictors. Make NAS tools open source for non-experts.

Mainstreaming Advanced Model Compression: Structured and unstructured pruning, quantization (ultra-low bit like INT4), and knowledge distillation – they should all be standard in production, especially the edge. Research should also be performed on hardware aware compression techniques that can be applied during

training (quantization aware training) to reduce accuracy degradation Automated compression pipelines are also a big thing.}

Promote Hardwae-Algorithm Co-design: Let AI researchers interact with hardware engineer more so they'll make new AI accelerator (ASS IC/FPGA) for sparse, irregular computation like recent models and efficient compressing techniques {such as sparse matrix multiply, low-precise arithmetic, attention} Software frameworks have to adapt in order to depict hardware capabilities to software modeling developers

5.3. Opening the Black Box: Explainable to Interpretable AI

To truly create trustworthy AI, we have to stop giving post facto explanations and being inherently transparent.

Research Inherently Interpretable Architectures: There has to be a shift away from explaining black box models after the fact and towards designing interpretable models from the start. We require huge investments in new archiitures which interpretability is built within the forward pass. Prototype-based networks: decisions are derived by comparing input features against patterns that the model has learned to represent (concept bottleneck models); these make the model output a set of human-understandable concepts (has stripes, is metallic) before arriving at a final decision, allowing humans to potentially observe and intervene at the concept level; disentangled representations: this is when different latent dimensions correspond to different and independent factors of variation. Crucially, developing these models should be done using a multi-objective evaluation that tries to find some trade–off between how accurate the models are and how human–interpretable these models are—the model that is so interpretable that it has to make suboptimal predictions just to explain its decision to people will not be acceptable, and the model that no one can understand will not have any trust at all.

To achieve this will take research problems. A big problem right now is relying on very large amounts of expensive, person labeled concept tags to build concept-based model. Thus, we need to develop novel training paradigms capable of learning an aligned set of semantically meaningful concept representations with weaker supervision, or without it. Additionally, it would have to be more than just finding what features were correlated (ie: add causal reasoning frameworks into the model) to really understand. And building models that can work out what

interventions and counterfactuals would be ("what if we took whiskers out of this picture? would still classify it as a cat?") So, if you could have these kinds of cause structures built into models, we'd have models that are much, much less likely to fall for this kind of spuriousness, but we can have models where our model would give us a reason that actually makes sense for how something caused it instead of just giving us some association that happened to be associated well together, which was really important in making very, very big decisions.

Create Standardized Auditing Frameworks: If we want our interpretable fair robust models to be anything more than just buzzwords, industry and academia must develop and follow standards for auditing. Make it so that universally accepted protocols, benchmarks, and quantities for these kinds of attributes exist, making them just as rigorous and significant as traditional accuracy benchmarks. In particular, it needs explanation benchmarks with ground-truth rationales (e.g., regions in images a human would attend to during classification, such that one could assess whether an explanation method had correctly identified the reasoning used to make a prediction) It's true, it's not good enough to just do auditing when you're like oh, did you check this static dataset well? Like you have to stress out about data shift, so that when you use the model, you're not being unfair to an underrepresented group or to a new domain where your model has shifted away from training.

And it's not some sort of fun for academics to do, and it actually has a real-world effect on regulation and actually having to do the job. With the addition of more and more AI for finance & health, criminal justice, etc. It's possible some form of independent audit against some standard safety and fairness criteria before passing, it would set a standard, making it simpler to do this kind of work and get certified. A goal to strive towards. In the end, standardizing auditing turns those subjective opinions about model trustworthiness into fact-checkable, comparable things, which creates an accountable way for people, regulators, and creators to look at the whole model-report-card, not just how well it does its job.

5.4. Solidifying the strength and security of the model as a subject

Robustness has to be a first-class design goal, not something you tack on at the end after all the really enjoyable parts are done.

For us to create secure models, we need to plant adversarial resilience into the training pipeline. The most empirically strong baseline for making models better at

avoiding such attacks is to do adversarial training, which means including more examples that were created by the network as it was trained: but the main problem is the high price of computing, because during training, it must repeatedly solve a small subproblem to generate attacking samples. Therefore, there is an effort to get to more robust and widely robust paradigms in the future work. Explore certified defenses which give mathematical guarantees about the model's stability under any input modifications inside a specific norm bound, using techniques such as randomized smoothing and interval bound propagation. Furthermore, robustness is not to be defined w.r.t. synthetic, worst-case perturbations only. Another parallel-and-equal goal is making models more robust, with natural corruptions & perturbations that are common in the real world like motion blur, sensor noise, and change in lighting/weather. Real world is very unpredictable; the real task is far from adversarial robustness, which requires closing the gaps between adversarial attack and natural distributions as a step towards developing robust models for this unpredictable real world.

Many places, where the application of this software is dangerous, when it breaks down, such as self-driving cars, diagnosis of medicine, industrial control in this place, a regular test can't prove its good. Formal Verification here's important + hard. This is like making sure that a brain made of wires (neural network) would do something we want for all the numbers a watch (sensor) can show when its working. take the network and its characteristics and convert them into constraints that solvers can use to prove that an autonomous car's perception system would never misclassify a stop sign under certain lighting and obstruction conditions. While most other methods are limited by the combinatorial explosion to small networks or certain properties, we need to invest here. Scalable verification algorithms, abstraction methods, and hardware-aware verification tools should be developed so that such guarantees can be made about the highly complex models that real-world safety-critical systems would use. credible ai systems must have metacognitive abilities so as to see themselves at limits. This would need models to be able to do both robust OOD detection, as well as to have reliable uncertainty estimates. OOD detection allows for recognizing inputs which are semantically different from the data it was trained on (like a new object for a classifier). Uncertainty estimation estimates how confident is the model in its prediction. In tandem, they make systems able to trip the right safety protocols — defer to a human operator, enter safe mode, etc. The following technical instruction

show some results, first is Bayesian Deep Learning: it models the weights and prediction uncertainty using methods like Variational Inference, Monte Carlo Dropout. then second are Ensemble methods: uncertainty comes from the fact that different models disagree. But as far as I'm concerned, the great barrier to them being used more in situations where time really matters is that they're slow on computers. So, then moving forward, work will be needed to make lightweight and scalable approximations to these frameworks – making single model proxies that can act as a diverse approximation of an ensemble or making fast inference for Bayesian NN's – without losing the quality of our uncertainty signal.

5.5. proactive building on an ethically framed and interdisciplinary governance

Technological progress is to be responsible as a steward.

Do end -to - end bias mitigation: Toss in good bias detection and easing tactics everywhere along the way in ML pipeline, from making and markup data for the models to training them, checking them, and using them, by means like fairness restrictions and adversarial easing {Diverse groups building it and fair fairness talk fitting the situation}

Accelerate Privacy-Preserving Tech: Make it so people using federated learning (work with different, not centered up data sets at once) and differential privacy (add some extra math noise to keep each point private) can learn better without all their information in one spot; how well and quickly those ways talk to each other and what they make is still in the works. Homomorphic encryption for secure inference does have possible solutions to look into.

Promote collaboration between different fields and develop policies: Develop lasting means of cooperation amongst AI technologists, social sciences, law, ethics, as well as subject matter specialists. Support sensible, flexible, rules and standards that foster innovation yet also guard interest, privacy, and civil rights, algorithmic impact assessments, public ledgers of high-risk AI systems, and sandboxed regulatory environments to test all have versions being explored across the planet, need help to grow.

6. Conclusion

Deep learning has made deep changes to the area of photos. This started a revolution in how computers could see, and this is now a big part of what we have today. From a long time ago lenet to today's hybrid CNNs+Vits and more architacturally the journey is ever more more powerful, more capable and more efficient models. And here is the new cool stuff that happened and now all the ideas in the old sci fi stuff that we saw actually came true so we got some cool new stuff now for our health and jobs and walking around and everything.

But clearly from this piece, we get that there'll be one really powerful, really hard set of barriers, and the trajectory was set by its unquenchable thirst for data, massive carbon footprint, lack of transparency, easy to manipulate, and the ethical issues it raises are quite troubling. Medical diagnosing, e-commerce, autonomous driving -- all the case studies point to the same sad truth, closing the gap from what we can do today to being able to reliably, safely, equitably get to this potential is an ugly mix of tech and non-tech.

And so, a constantly responsible advance has to entail an intentional switch in strategy. We need to push back against the current singular push towards getting small improvements in accuracy on static benchmarks with an all-hands-on-deck push to build the next generation of data efficient, energy efficient, interpretable, robust, and equitable AI. All of this in a single package: starting from a firm grasp of the foundation in algorithms, moving up the stack toward cross-stack system optimizations for sustainability, then through the various evaluation regimes for trustworthiness to guidelines and infrastructure for ethics and governance. {Integrating vision models with multimodal "and" not "but" interactive AIs that can think and talk about what they see next and, on all fronts,}

Moving towards ways of learning with more efficiency by using less data together, working with green AI ideas, telling others what models decide in a way they can understand, making sure that our systems still perform their job well when things go wrong or when someone wants to trick them, and making sure that things are fair right from the beginning of creating these smart picture-reading machines, we can all, researchers, companies, leaders included - push for better ways to read pictures through deep learning. if we want a society's great advantage, we have to be forward - thinking and discover the danger, then reduce. Image recognition will not be defined by models that see better in the future. They will be defined by building systems that understand more, do more, and better serve all people.

Acknowledgment: We'd like to give a thank for all this huge amount of open-sourced researches, those who are creating publicly available datasets, such as ImageNet and COCO, and the people who share their pre-trained models which make it all possible. and we thank all other academics and industries for their continual critiques on the hardships, the ethics, and the direction of artificial intelligence. All these have a great influence on this work.

Conflict of interest: The author declares no conflict of interest.

Funding: This research received no external funding.

References

[1] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). *On the opportunities and risks of foundation models* (arXiv:2108.07258). arXiv. https://arxiv.org/abs/2108.07258

[2] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html

[3] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 1597–1607). PMLR. https://proceedings.mlr.press/v119/chen20j.html

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

[5] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*. https://arxiv.org/abs/1412.6572

[6] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA, 316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

[7] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16000–16009). https://doi.org/10.1109/CVPR52688.2022.01553

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image

recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

[9] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700–4708). https://doi.org/10.1109/CVPR.2017.243

[10] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4015–4026). https://doi.org/10.1109/ICCV51070.2023.00371

[11] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[12] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791

[13] Liang, W., Zhang, Y., Cao, J., Xie, B., Yu, K., & Wang, F.-Y. (2023). *Can large language models understand context?* (arXiv:2302.07180). arXiv. https://arxiv.org/abs/2302.07180

[14] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022). https://doi.org/10.1109/ICCV48922.2021.00986

[15] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with CLIP latents* (arXiv:2204.06125). arXiv. https://arxiv.org/abs/2204.06125

[16] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520). https://doi.org/10.1109/CVPR.2018.00474

[17] Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63. https://doi.org/10.1145/3381831

[18] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*. https://arxiv.org/abs/1409.1556

[19] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594

[20] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 6105–6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html