Chapter 8
The Architectures of Geoffrey Hinton

Ivana Stanko

Abstract Geoffrey Everest Hinton is a pioneer of deep learning, an approach to machine learning which allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction, whose numerous theoretical and empirical contributions have earned him the title the Godfather of deep learning. This chapter offers a brief outline of his education, early influences and prolific scientific career that started in the midst of AI winter when neural networks were regarded with deep suspicion. With a single goal fueling his ambitions—understanding how the mind works by building machine learning models inspired by it—he surrounded himself with like-minded collaborators and worked on inventing and improving many of the deep learning building blocks such as distributed representations, Boltzmann machines, backpropagation, variational learning, contrastive divergence, deep belief networks, dropout, and rectified linear units. The current deep learning renaissance is the result of that. His work is far from finished; a revolutionary at heart, he is still questioning the basics and currently developing a new approach to deep learning in the form of capsule networks.

Keywords Geoffrey Hinton · Restricted Boltzmann machines · Cognitive science · Distributed representations, Backpropagation algorithm, Deep learning

8.1 Context

The field of artificial intelligence has been divided into two opposing camps on the converging paths toward the ultimate quest of "making machines do things that would require intelligence if done by men"[1]—those who want to *program* machines to do things (knowledge-based approach based on manipulating symbols according to the rules of logic) and those who want to *show* machines how to do things (machine

[1] Marvin Minsky.

I. Stanko (✉)
University of Zagreb, Zagreb, Croatia
e-mail: istanko@hrstud.hr

S. Skansi (ed.), *Guide to Deep Learning Basics*,
https://doi.org/10.1007/978-3-030-37591-1_8

learning based on applied statistics and extracting patterns from raw data). After many decades and many a hard-won battle, the winner is clear-cut and accumulated efforts of generation upon generation of great thinkers from many diverse fields crystallized in the approach to machine learning called deep learning which allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction [25]. Credit assignment as the central problem of machine learning aside, it is hard to dispute that the final steps toward the current deep learning renaissance have been spearheaded by the unshakable and deep belief of Geoffrey Hinton, the Godfather of deep learning, and the person behind many of its principal components.

Geoffrey Everest[2] Hinton, an Emeritus Distinguished Professor at the University of Toronto, a Vice President and Engineering Fellow at Google, Chief Scientific Adviser of the Vector Institute, an ACM Turing Award Laureate and one of the Deep Learning Conspiracy[3] trio along with Yoshua Bengio and Yann LeCun, was born on December 6, 1947 in London to a family of distinguished and accomplished individuals. In the context of Hinton's life endeavor, the most prominent node in his family tree represents George Boole whose seminal work *An Investigation of the Laws of Thought* introduced Boolean algebra that has not only been instrumental in the development of digital computing but also offered many less-remembered insights into probability theory, the bedrock of machine learning, that have come to fruition through the research of his great-great-grandson [36]. Hinton's mother was a school teacher who made it clear early that he had choices in life—he could either be an academic (like his father, an entomologist) or a failure[4]—so by the age of 7, he already knew that getting a Ph.D. was nonnegotiable.[5] His father, a Stalinist, sent him to a private Catholic school, an education which in addition to a regular curriculum included mandatory morning prayers, which taught him, among other things, that people's beliefs are often nonsense [31]; a most useful training example that enabled him to generalize well throughout his professional career.

School, in general, was not a particularly happy time for Hinton; he spent the whole 11 years feeling like an outsider which lit up a contrarian spark in him and prepared him for the revolutionary role he would assume in the development of neural networks [31]. It was in high school that his interest took shape during illuminating conversations with his friend Inman Harvey who introduced Hinton to the idea that human memory might work as a hologram, a clear-cut example of distributed representations, where each memory is stored by adjusting connection strengths between the neurons across the entire brain [4]. His first foray into university ended with him dropping out after a month and doing odd jobs here and there while his life goals

[2] Name shared with a relative George Everest, a surveyor and a geographer, after whom Mount Everest was named.

[3] Another popular nickname is Canadian Mafia.

[4] *The Godfather of Deep Learning was Almost a Carpenter*. For full interview see https://www.bloomberg.com/news/videos/2017-12-01/the-godfather-of-ai-was-almost-a-carpenter-video.

[5] *This Canadian Genius Created Modern AI*. For the full interview see https://www.youtube.com/watch?v=l9RWTMNnvi4.

crystallized [31]. The second time around, he started studying Physics and Physiology; Physiology did not even try to explain the how behind the mechanics of the brain function and physics left him drowning in a sea of equations he considered too difficult so he switched to Philosophy for a year hoping to find the answer to the meaning of life—he did not and was left unsatisfied because the framework of the discipline offered no way to verify whether you were right or wrong [31]. the third time was not the charm and his next stop, Psychology, turned out not to be the royal road to understanding the complexities of the mind—the emphasis was on rats in the mazes, models of how the mind worked were hopelessly inadequate and it was only after Hinton spurred other students to protest the content of the course that the department organized a single perfunctory lecture on Freud, Jung, and Adler—his discontent with the entire field continued to grow [31]. Ditching the academia altogether, he tried out the life of a carpenter but having met an expert one Hinton realized his inadequacies [4] and returned to Cambridge where he earned his B. A. Hons in Experimental Psychology in 1970.

Having realized that the only way to truly understand a complex device such as the human brain is to build one,[4] he set his sights on artificial intelligence and went to Edinburgh University where he began his graduate studies under the supervision of Christopher Longuet-Higgins, a renowned chemist who coined the term *cognitive science* and envisioned it as a unification of mathematical, linguistic, psychological, and physiological sciences, with AI as the key ingredient that would elucidate how the human mind worked [26]. It was the shared view on AI's real purpose and Longuet-Higgins' work on an early model of associative memory that made Hinton choose him as his thesis advisor. Unfortunately, just before Hinton arrived Longuet-Higgins' interest in neural networks had started to wane as he found himself impressed with Terry Winograd's dissertation *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language* and its star SHRDLU. Longuet-Higgins assumed Hinton wanted to follow the same path and work on symbolic AI; many fights ensued but convinced that neural networks were the only way to go for what Hinton persisted in doing what he believed in.[6] Despite the disagreements and a lack of belief in Hinton's ideas, Longuet-Higgins still supported his right to pursue them and provided valuable advice about building everything on solid mathematical foundations [31]. Hinton was awarded his Ph.D. in 1978 but the prevailing attitude in Britain was that neural networks were a complete waste of time which prevented Hinton not only from finding a job but also from landing a single job interview; a completely different atmosphere awaited at the University of California where he spent time collaborating with Don Norman and David Rumelhart in a fertile intellectual environment that encouraged the to and fro of artificial neural networks and Psychology.[6]

[6] *Heroes of deep learning: Andrew Ng interviews Geoffrey Hinton.* For full interview see https://www.deeplearning.ai/blog/hodl-geoffrey-hinton/.

8.2 Building Blocks

Psychology had long questioned the localizationist doctrine—one computational element for one entity—and the work of Jackson, Luria, Lashley, Hebb, Rosenblatt, and Selfridge, among many others, pointed toward a different way of implementing propositional knowledge in the processing system like the human brain. It seemed more likely that concepts were implemented by **distributed representations** where each entity is represented by a pattern of activity distributed over many computing elements, and each computing element is involved in representing many different entities [15], which automatically enabled generalization because similar objects are represented by similar patterns and property-inheritance whenever the units that are active in the representation of a type are also a subset of the units active in the representation of an instance of that type [6]. Whatever the brain was doing obviously worked and following Orgel's Second Rule "Evolution is cleverer than you are", Hinton knew that distributed representations provided a promising model of implementing semantic networks in parallel hardware. In order to use them effectively the learning problem of finding a good pattern of activity, i.e., the underlying complex regularities, needed to be solved.

An early deep learning model that attempted to solve that problem appeared in 1983 under an intentionally veiled article title *Optimal perceptual inference*, a necessary decision given that even mentioning neural networks was frowned upon in most academic circles. In it Hinton and Terrence Sejnowski hinted at **Boltzmann machines**—energy-based models, which are essentially stochastic Hopfield networks with symmetrically connected hidden[7] units, they developed during Hinton's time at Carnegie Mellon. The energy minimum of a Boltzmann machine corresponds to a distributed representation and the whole system creates a good collection of them by clamping some of the individual units into certain states that represent a particular input, which allows it to find a good energy landscape that is compatible with that input [1]. This is achieved with an unsupervised learning algorithm that requires only local information and uses hidden units to capture the best regularities in the environment by minimizing the Kullback–Leibler divergence by changing the weights in proportion to the difference between the data-dependent expected value of the product of states at thermal equilibrium and the corresponding data-independent expected value [19]. The problem was that the learning itself was slow, noisy, and impractical; Sejnowski believed it could be vastly improved if there were a way to learn smaller modules independently [12]. He was right, but it would take almost 20 years and additional tinkering to transform Boltzmann machines into an essential component of many deep probabilistic models they are today.

Another seed that would require time to come into full bloom and signal the arrival of the AI spring was planted in 1986 when Rumelhart, Hinton, and Ronald Williams managed to get the *Learning representations by back-propagating errors* paper that has since then become classic published in Nature. That itself required some political work[6]—realizing its importance for the progress in the field of AI, Hinton talked to

[7] The name for those units was actually inspired by Hidden Markov Models [31].

one of the suspected referees Stuart Sutherland, a well-known British psychologist, and explained to him the value of using a relatively simple learning procedure to learn distributed representations which, fortunately, did not fall on deaf ears. **Backpropagation**, or simply backprop, uses the chain rule of calculus to compute the gradient allowing another algorithm, like stochastic gradient descent, to perform the learning [5]. Variations of the algorithm date back to 1960s and the work of Henry J. Kelly, Arthur E. Bryson, and Stuart Dreyfus while in the 1970s, Seppo Linnainmaa and Paul Werbos worked on implementing it. When Rumelhart rediscovered it in 1981 and fine-tuned it with Hinton and Williams, nothing spectacular happened; only after returning to it after his work on Boltzmann machines did Hinton realize its full potential and the trio resumed their work on it, applying backpropagation to word embeddings to learn distributed representations and seeing semantic features emerge from it [4].

The network was given information expressed in sets of triplets *person1-relationship-person2* in a graph structure, a family tree task, and succeeded in converting that information into big feature vectors from which emerged meanings of individual features like the nationalities, generations, and branches of the family tree that represented people, allowing the network to derive new information that was used to predict the third item given the first two [32]. Backpropagation took the network from graph structure (input) to features and their interactions (hidden units) and back again (output) by repeatedly adjusting the connection strengths in the network (weights) in a backward pass, starting from the penultimate layer and making its way down the earlier ones to minimize the total error between the actual and desired output [32]. Equally important was the fact that backpropagation rose up to the challenge presented in Minsky's and Papert's 1969 *Perceptrons* and provided solutions to the problems with Rosenblatt's invention—multilayer neural networks were now able to handle abstract mathematical problems such as the XOR and other parity problems, encoding problems, binary addition, and negation as well as mirror symmetry problems and geometric problems such as discrimination between T and C independent of their translation and rotation [33].

Together with Alexander Waibel, Toshiyuki Hanazawa, Kiyohiro Shikano, and Kevin Lang, Hinton made his first practical application by developing a version of backpropagation called time-delay neural networks, which was used for successful phoneme recognition [38]. A researcher at heart, his interest did not lie in developing applications; they were a means to an end—proving something was useful enough to keep the funding flowing while his sights were set on figuring out how the brain works [31]. Despite it being a major breakthrough that started to turn the tide within the AI community, backpropagation did not work, as it was expected, well at the time. Today, when it is used in probably 90% of commercial and industrial applications of neural networks [28], it is relatively easy to see what held it back in the 80s and 90s. In Hinton's own words, the labeled datasets that were used were thousands of times too small, the computers were millions of times too slow, the weights were initialized in a stupid way, and a wrong type of nonlinearity was used.[8] Time would take care of the first two, while the joint effort of Hinton and his collaborators would be behind rectifying the rest.

8.3 Tinkering

As the temporary hype died out so did the funding. Left with the option of accepting military (Office of Naval Research) money to continue his work and generally dissatisfied with the politics that prevailed in American society, Hinton moved to Canada where he took up a position at the Department of Computer Science at the University of Toronto and was able to continue his research minimally interrupted by teaching obligations—thanks to the funding he received from the Canadian Institute for Advanced Research (CIFAR) [31]. His aspiration was to foster the transfer of ideas between the recent advances in neural networks and methods used in statistics, mainly variational methods that included deterministic approximation procedures that generally provided bounds on probabilities of interest [22].

One of the popular ones was the **expectation maximization** (EM) algorithm—a method for tackling approximate inference problems in undirected graphical models with visible and latent variables where computing the log probability of observed data is too difficult so a lower bound (the negative variational free energy) is computed instead and the entire inference problem boils down to finding an arbitrary probability distribution over latent variables that maximizes the lower bound [5]. The EM algorithm alternates between an expectation (E) step that finds the distribution of the latent variables given the known values of the visible ones and the current estimate of the parameters and a maximization (M) step that adjusts the parameters according to maximum likelihood, assuming that the distribution from the E step is correct [30]. Radford Neal and Hinton improved it by making a generalization of it which can be seen in terms of Kullback–Leibler divergence by showing that there was no need to perform a perfect E step, an approximation of it computed by recalculating the distribution for only one of the latent variables would suffice and would converge faster to a solution in mixture estimation problems [30].

His attention turned to unsupervised learning and **autoencoders**, neural networks trained to first convert an input vector into a code vector using recognition weights and then to convert that code vector into an approximate reconstruction of an input using generative weights [21]. Unsupervised learning is a more plausible model of the way humans learn and can be implemented in machine learning by minimizing the sum of a code cost, the number of bits necessary to describe the activities of the hidden units, a reconstruction cost, the number of bits necessary to describe the difference between the input, and its best approximation reconstructed by the activities

[8] Geoffrey Hinton's keynote speech at the 2015 Royal Society *Machine learning: breakthrough science and technologies—Transforming our future conference series*. For full lecture see https://www.youtube.com/watch?v=izrG86jycck/.

of the hidden units [7]. Working with Richard Zemel, they devised a way of training autoencoders using a variation of Minimum Description Length principle by using nonequilibrium Helmholtz free energy as an objective function that simultaneously minimizes the information provided by the activities of the hidden units and the information contained in the reconstruction error [21].

Peter Dayan and Radford Neal joined them and together they generalized that to a multilayer system—**the Helmholtz machine** which is a neural network consisting of multiple layers of binary stochastic units that are connected hierarchically by two sets of weights—bottom-up implement a recognition model that infers a probability distribution over hidden variables given the input, while top-down enforce a generative model that reconstructs the values of the input from the activities of the hidden units [3]. Being an unsupervised neural network, there is no external teaching signal to match so the hidden units need to extract the underlying regularities from the data by learning representations that are economical in their description length but sufficient for an accurate reconstruction of the data, which can be achieved with the **wake-sleep algorithm** [13]. In the wake phase, the bottom-up recognition weights produce a representation of the input in each layer and combine them to determine a conditional probability distribution over total representations; the top-down generative weights are then modified using the delta rule to become better at reconstructing the activities in preceding layers [13]. The sleep phase is driven by top-down generative weights that provide an unbiased sample of the network generative model that is used to train the bottom-up connections again using the delta rule [13].

Another technique Hinton developed for training undirected graphical models was **contrastive divergence** first used on **products of experts** which model high-dimensional data by multiplying and renormalizing several probability distributions of different low-dimensional constraints of that data [8]. Features were learned by following the approximation of the gradient of contrastive divergence, an objective function that removes the mistakes the model generates by minimizing the difference between the two Kullback–Leibler divergences—the one between the data distribution and the equilibrium distribution over the visible variables produced by the Gibbs sampling from the generative model and the other between the reconstructions of the data vectors from the last seen example and the equilibrium distribution—which is equivalent to maximizing the log likelihood of the data [9]. This technique could be applied to **Restricted Boltzmann machines** (RBMs), undirected probabilistic graphical models containing a layer of observable and a layer of latent variables [5], which could be viewed as products of experts with one expert per hidden unit [9]. The learning procedure became much simpler than in the original Boltzmann machines, thus making RBMs practical enough to become not only a staple in the construction of the upcoming deep architectures but also powerful enough on their own to become a part of a million-dollar winning entry to Netflix collaborative filtering competition [35].

8.4 Deep Learning

In 2006 neural networks reemerged under a new name—*deep learning*—and a new era began ushered in by Hinton's, Simon Osindero's and Yee-Whye Teh's, now legendary, paper *A fast learning algorithm for Deep Belief Nets*. In it, they demonstrated that multilayered networks could not only be trained efficiently but could also outperform the models that had dominated the machine learning landscape. This was accomplished by using greedy layer-wise pretraining that corrected the earlier mistake of initializing the weights stupidly—the structure in the input was no longer ignored as with backpropagation alone because by starting off with unsupervised learning, the network could discover latent variables (features) that captured the structure in the training data allowing discriminative learning that followed to model the dependence of the output on the input by fine-tuning those discovered features to discriminate better [11].

The **deep belief network** (DBN) they presented was a multilayer stack of RBMs, the top two hidden layers formed an undirected associative memory and the remaining hidden layers received directed top-down connections and converted the representations in the associative memory into observable variables [16]. By using contrastive divergence learning in RBMs to learn one layer of features at a time, each layer modeling the previous one, the whole learning process was broken down into a sequence of simpler tasks avoiding the inference problems that otherwise appeared in directed generative models [10]. The more layers the network had, the better it performed by exploiting the fact that natural signals are compositional hierarchies [25] and each new hidden layer became an improvement on the variational bound on the log probability of the training data [10]. Generative abilities of the network could be improved by using a contrastive version of the wake-sleep algorithm to fine-tune the learned weights [16], whereas backpropagation could be used to improve the performance on discrimination tasks [18].

The area that was transformed first by these innovations was speech recognition, formally dominated by Hidden Markov Models. In 2009 DBNs used for acoustic modeling by Abdel-rahman Mohamed, George Dahl, and Hinton outperformed other state-of-the-art models on TIMIT [27]. The technology was then offered to RIM,[9] which declined but Google found it interesting enough and by 2012 it was implemented in Android.[8] The academia finally accepted neural networks, at least those who did not subscribe to Max Planck's view on the progress of science,[10] the industry started to take notice but it took some additional tweaking for deep learning to establish itself as the dominant approach to AI.

The problem of overfitting needed to be solved and that was accomplished by taking a leaf out of evolution's book, the chapter on sex, which gave birth to the idea of **dropout**. Sexual reproduction breaks-up sets of complicated coadapted genes, achieving robustness in functionality by forcing each to pull its own weight and mix well with random ones rather than being useful only in tandem with a large

[9] Research in Motion, known as BlackBerry Limited since 2013.

[10] "Science advances one funeral at a time."

number of others already present, thus reducing the probability that small changes in the environment will lead to large decreases in fitness [20]. Dropout, developed by Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov, and Hinton, simulates ensemble learning and leads to similar robustness by adding noise to the states of hidden units in a neural network [37]. The noise temporarily removes a fixed fraction of the feature detectors and their connections on each presentation of each training case, preventing units from co-adapting too much and forcing individual neurons to learn to detect generally helpful features for arriving at the correct solution given the combinatorially large variety of internal contexts in which they must operate [20]. As a result, dropout has improved generalization performance on tasks across many different domains such as vision, speech recognition, document classification, and computational biology with the only trade-off being increased training time [37].

A simple and effective solution to that problem, and the longstanding issue of using the wrong kind of nonlinearity, was in the synergistic effects of using dropout with a special kind of activation function: **Rectified Linear Unit** (ReLU) [2] that performs better than the logistic sigmoid function and would become the default choice in deep neural networks. Thought of by three different groups of researches, ReLUs have become the core of every deep network where they sum the weighted input of a unit into its activation or its output according to g(x) = $\max\{0, x\}$. It is a piecewise linear function with two linear pieces and applying it to linear transformation gives a nonlinear transformation that outputs zero across half its domain which makes its derivatives large whenever it is active and gradients consistent; this mimics biological neurons and their sparse and selective activations that depend on the input [5]. Vinod Nair and Hinton used it in RBMs as an approximation to an infinite set of replicated binary units with tied weights and shifted biases so as to maintain the RBMs probabilistic model and it worked better than binary hidden units for recognizing objects and comparing faces [29]. When combined with dropout, ReLUs no longer overfit quickly in comparison to sigmoid nets and they worked well together leading to improvements in error reduction over other deep neural network models on LVCSR [2].

These two new insights were combined with the efficient use of GPUs and data augmentation, and then applied to a well-developed albeit overlooked deep architecture called a convolutional neural network and **AlexNet** was conceived. Its landslide victory on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) was the tipping point that finally transformed deep learning from a niche interest to an overnight sensation. The competition's goal was to get the lowest top-1 and top-5 error rates in the detection and correct classification of objects and scenes belonging to 1000 different categories with 1000 images in each, totaling approximately 1.2 million training, 50,000 validation and 150,000 testing images. Krizhevsky, Sutskever, and Hinton accomplished exactly that in 2012 and explained how in the seminal *ImageNet Classification with Deep Convolutional Neural Networks* paper.

AlexNet consisted of eight learned layers—the first five were convolutional and the other three fully connected; the output of each was applied to the ReLU allowing much faster learning than with saturating neurons while the output of the last fully connected layer was fed to a 1000-way softmax which produced a distribution over the 1000 class labels [24]. The kernels of the second, fourth, and fifth convolutional layers were connected only to those kernel maps in the previous layer which resided on the same GPU, allowing for an additional trick—GPU communication only in certain layers—along with the parallelization [24]. Response-normalization layers, which aid generalization by creating competition for big activities among neuron outputs that are computed using different kernels thus implementing a variation of lateral inhibition inspired by the one biological neurons use, followed the first and second convolutional layers [24]. Since the architecture contained 60 million parameters, in order to reduce overfitting dropout was used in the first two fully connected layers and the dataset itself was artificially enlarged by generating more training examples by deforming existing ones with image translations, horizontal reflections and altering the intensities of the RGB channels [24]. They achieved the error rate of 15.3%, the runner-up's was 26.2% and it was the depth of AlexNet that enabled that result, with a single convolutional layer removed its performance degraded [24].

Since then, convolutional neural networks have become the bread and butter of recognition and classification tasks and have enabled real-time vision applications in smartphones, cameras, robots, and self-driving cars [25] and deep learning has become the new *it* thing. In 2012, at Andrew Ng's request, Hinton started the first-ever massive online open course on machine learning, *Neural Networks for Machine Learning*,[11] that enabled interested audience from all walks of life to learn about the emerging technology. The industry quickly seized the opportunity and acquired the talent behind the pioneering achievements—Hinton has been working part-time for Google since 2013 and is still trying to discover the learning procedure employed by the brain.

8.5 New Approaches

Having transitioned, in Hinton's own words, from the lunatic fringe to the lunatic core[12] he has gone full circle and hitched his wagon to an idea he feels extremely strongly about but not many people ready to put much stock in[6] **capsules**. A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity, such as an object or an object part; the length of that activity vector is used to represent the probability that the entity exists while its orientation represents the instantiation parameters [34]. They differ from neural networks in the activation—capsules get activated in response to comparison of multiple incoming pose predictions while neural networks compare a single incoming activity vector

[11] At Hinton's request the course has since been removed from Coursera due to being out of date but it can still be accessed from his webpage https://www.cs.toronto.edu/~hinton/coursera_lectures.html or YouTube.

[12] *Meet the man Google hired to make AI a reality*. For full article see https://www.wired.com/2014/01/geoffrey-hinton-deep-learning/.

and a learned weight factor [17] and represent a response to everything that is wrong with neural networks, especially in speech and object recognition where despite great success they still do not work as well as the brain does, which could be due to having fewer levels of structure.

Loosely inspired by minicolumns in the human visual system, capsules seek to rectify the inefficiencies by building an explicit notion of an entity into the architecture itself, thus providing equivariance where changes in the viewpoint, the biggest source of variation in images, lead to corresponding changes in neural activities because they capture the underlying linear structure, making use of the natural linear manifold that efficiently deals with variations in position, orientation, scale, and lightning.[13]

Introduced in 2011 as a simple way of recognizing wholes by recognizing their parts, Hinton, Krizhevsky, and Sida Wang proposed transforming autoencoders into first-level capsules which would have logistic recognition units that compute the outputs sent to higher levels and the probability that the capsule's entity is present in the input, and generation units that compute the capsule's contribution to the transformed image. From pixel intensities, first-level capsules extract explicit pose parameters of recognized fragments of their visual entities, externally supplied transformation matrices that learn to encode intrinsic spatial relationships between the parts and the whole are applied to those fragments and agreement of the poses predicted by active lower level capsules is then used to activate higher level capsules and predict the instantiation parameters of larger, more complex visual entities [14]. Sara Sabour, Nicholas Frosst, and Hinton further refined the system introducing an iterative routing-by-agreement mechanism for connecting layers in the feedforward network—dynamic routing—which enables capsules in the lower levels to send their output to parent capsules in higher levels that compute a prediction vector which, if their scalar product is large, increases the coupling coefficient for the chosen parent thereby increasing the contribution of that capsule and prediction with the parent's output [34]. More effective than max-pooling, dynamic routing, which allows capsules to selectively attend only to some active capsules in the layer bellow, enabled capsules to outperform similarly sized CNNs on affNIST and multiMNIST successfully recognizing multiple overlapping objects in images [34].

Exploring the potential of the new approach, the trio came up with a new matrix capsule system in which each capsule contains a logistic unit that represents the presence of an entity and a 4×4 pose matrix that captures its pose regardless of the viewpoint allowing the system to recognize objects with different azimuths and elevations [17]. A novel iterative routing procedure, based on the EM algorithm, routes the output of children capsules to parent capsules in an adjacent layer so that each active capsule receives a cluster of similar pose votes [17]. With these improvements, capsules outperform CNNs on the smallNORB dataset and display more robustness to white box adversarial attacks [17]. An unsupervised version of

[13] *What's wrong with convolutional nets?* Brain and Cognitive Sciences—Fall Colloquium Series Recorded December 4, 2014. For full lecture see https://techtv.mit.edu/collections/bcs/videos/30698-what-s-wrong-with-convolutional-nets.

capsules—Stacked Capsule Autoencoder—consists of two stages and presents an updated version devised by Adam Kosiorek, Sabour, Teh, and Hinton which no longer needs externally supplied transformation matrices but uses the image as the only input and no longer needs iterative routing because objects now predict parts [23]. In the first stage, part capsules segment the input into parts and poses, and reconstruct each image pixel as a mixture of affine-transforming learned templates while in the second stage object capsules arrange discovered parts and poses into a smaller set of objects thereby discovering the underlying structure [23]. Capsules are a work in progress but their built-in biases and unsupervised way of learning inspired by our own could contribute to the prevailing sentiment that the (next) "*revolution will not be supervised*".[14]

8.6 Five-Year Fog

Unknowable unknowns make long-term predictions about the future a fruitless endeavor. When asked to make them, Hinton prefers to use an analogy with driving a car during a foggy night—the distance that would otherwise be clearly visible and easily navigable becomes almost opaque due to the exponential effects of the fog that absorbs a fraction of photons per unit of distance.[15] Technology follows the same exponential progress and there is simply no telling where AI will end up in a couple of decades. Whatever the potential risks associated with it may be and regardless of the variation of the doomsday scenario one prefers to entertain, Hinton believes the problem is not technology itself—biases in neural networks are easy to fix, people present a greater challenge—but social systems that are rigged to benefit the top 1% at the expense of everybody else and one way to thwart that is through regulation of the use of AI, especially in weaponization, elections, and surveillance [4].

Fixing the system is a better long-term option than demonizing the technology and hindering its progress especially now when deep learning has gone mainstream with troves of researchers flocking to it. Hinton sees universities where young graduate students willing to question the basics and freely pursue truly novel ideas while being well advised by experts with similar views as more likely incubators of future advances than the industry [4]. Having time to read just enough to develop intuitions, ample insight to notice the flaws in the current approach and perseverance to follow those intuitions despite external influences is the only way to further the progress of deep learning,[6] given that this is exactly what Hinton has done it is not a bad example to approximate.

Neural networks started off as an underdog quickly dismissed as a curiosity that would never work and relegated to a fringe interest of especially perseverant individuals with a contrarian streak from various disciplines united under the header

[14] Yann LeCun.

[15] The final lecture of his *Neural Networks for Machine Learning* Coursera course. For full video see https://www.youtube.com/watch?v=IXJhAL6FEj0.

of cognitive science. Geoffrey Hinton is one of those pioneers whose belief in the connectionist approach has never been shaken. Earlier in his career, in the midst of general resistance, he stated: *But sooner or later computational studies of learning in artificial neural networks will converge on the methods discovered by evolution. When that happens a lot of diverse empirical data about the brain will finally make sense, and many new applications of artificial neural networks will become feasible* [7]. Due to his efforts and numerous inventions and improvements that followed from the continuous exchange of ideas between those he considers mentors and those he mentored, what he firmly believed in while many thought it was impossible is starting to happen sooner rather than later. Whatever emerges out of the fog will undoubtedly be influenced and shaped by his numerous and continued contributions.

References

1. Ackley DH, Hinton GE, Sejnowski TJ (1985) A learning algorithm for Boltzmann machines. Cogn Sci 9:147–169
2. Dahl GE, Sainath TN, Hinton GE (2013) Improving deep neural networks for LVCSR using rectified linear units and dropout. In: IEEE international conference on acoustic speech and signal processing (ICASSP 2013), pp 1–5
3. Dayan P, Hinton GE, Neal R, Zemel RS (1995) The Helmholtz machine. Neural Comput 7:1022–1037
4. Ford M (2018) Architects of intelligence: the truth about AI from the people building it. Packt Publishing, Birmingham, UK
5. Goodfellow I, Bengio Y, Courvile A (2016) Deep learning. The MIT Press, Cambridge, MA
6. Hinton GE (1981) Implementing semantic networks in parallel hardware. In: Hinton GE, Anderson JA (eds) Parallel models of associative memory. Lawrence Erlbaum Associates, pp 191–217
7. Hinton GE (1992) How neural networks learn from experience. Sci Am 267(3):145–151
8. Hinton GE (1999) Products of experts. In: Proceedings of the ninth international conference on artificial neural networks (ICANN 99), vol 1, pp 1–6
9. Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14:1771–1800
10. Hinton GE (2007) Learning multiple layers of representation. Trends Cogn Sci 11(10):1527–1554
11. Hinton GE (2007) To recognize shapes, first learn to generate images. In: Drew T, Cisek P, Kalaska J (eds) Computational neuroscience: theoretical insights into brain function. Elsevier, pp 535–548
12. Hinton GE (2014) Where do features come from? Cogn Sci 38(6):1078–1101
13. Hinton GE, Dayan P, Frey BJ, Neal R (1995) The wake-sleep algorithm for unsupervised neural networks. Science 268:1158–1161
14. Hinton GE, Krizhevsky A, Wand SD (2011) Transforming auto-encoders. In: International conference on artificial neural networks systems (ICANN-11), pp 1–8
15. Hinton GE, McClelland JL, Rumelhart DE (1986) Distributed representations. In: Rumelhart DE, McClelland JL (eds) Parallel distributed processing: explorations in the microstructure of cognition. Volume 1: foundations. MIT Press, pp 77–109
16. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18:1527–1554
17. Hinton GE, Sabour S, Frosst N (2018) Matrix capsules with EM routing. In: International conference of learning representations (ICLR 2018), pp 1–15

18. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313:504–507
19. Hinton GE, Sejnowski TJ (1983) Optimal perceptual inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 448–453
20. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580, pp 1–18
21. Hinton GE, Zemel RS (1994) Autoencoders, minimum description length, and Helmholtz free energy. In: Cownan JD, Tesauro G, Alspector G (eds) Advances in neural information processing systems, vol 6. Morgan Kaufman, pp 1–9
22. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1998) An introduction to variational methods for graphical models. In: Jordan MI (ed) Learning in graphical models. Springer Netherlands, pp 105–161
23. Kosiorek AR, Sabour S, The YW, Hinton GE (2019) Stacked capsule autoencoders. arXiv:1906.06818 [stat.ML], pp 1–13
24. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems 25 (NIPS 2012), pp 1–9
25. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444
26. Longuet-Higgins HC (1973) Comments on the Lighthill report and the Sutherland reply. In: Artificial intelligence: a paper symposium. Science Research Council, pp 35–37
27. Mohamed A, Dahl GE, Hinton GE (2009) Deep belief networks for phone recognition. In: NIPS 22 workshop on deep learning for speech recognition, pp 1–9
28. Munakata T (2008) Fundamentals of the new artificial intelligence: neural, evolutionary, fuzzy and more. Springer, London
29. Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on machine learning, pp 1–8
30. Neal RM, Hinton GE (1998) A view of the EM algorithm that justifies incremental, sparse and other variants. In: Jordan MI (ed) Learning in graphical models. Springer Netherlands, pp 355–368
31. Rosenfield E, Hinton GE (2000) In: Anderson JA, Rosenfield E (eds) Talking nets: an oral history of neural networks. The MIT Press, pp 361–386
32. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
33. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL (eds) Parallel distributed processing: explorations in the microstructure of cognition. Volume 1: foundations. MIT Press, pp 318–362
34. Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: Conference on neural information processing systems (NIPS 2017), pp 1–11
35. Salakhutdinov R, Mnih A, Hinton GE (2007) Restricted Boltzmann machines for collaborative filtering. In: International conference on machine learning, Corvallis, Oregon, pp 1–8
36. Sejnowski TJ (2018) The deep learning revolution. The MIT Press, Cambridge, MA
37. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958
38. Waibel A, Hanazawa T, Hinton GE, Shikano K, Lang KJ (1989) Phoneme recognition using time-delay neural networks. IEEE Trans Acoust Speech Signal Process 37(3):147–169