Depth in Deep Learning: Knowledgeable, Layered, and Impenetrable

Taylor Arnold and Lauren Tilton

Introduction

Over the past decade, research in machine learning has made remarkable progress in the processing of text and image data. Computational models are now able to outperform human experts in certain classes of well-defined tasks. These improvements come in part through advances in computer hardware, as well as access to larger public datasets for training models. Perhaps the largest contributing factor, however, has been the refinement and application of so-called "deep learning" models. The incredible accuracy achieved through these models has led directly to read-world applications, including automated medical imaging diagnoses, prototypes for self-driving cars, and machine translation software. Of course, not all applications have resulted in a net-positive. Deep learning models have also been employed in troubling ways as tools for the security state, as justification for heavily policing minorities, and in the resurgence of a computational study of eugenics.[1] The predictive power of deep learning demands that we take seriously the study of their structure and impact in society. The unmatched predictive power of these models in certain critical domains guarantee that deep learning will continue to inject algorithmic logic into critical decisions affecting everyday lives.

Deep learning models are a class of algorithms that find latent hierarchical structures within large datasets. They are constructed my chaining together layers of smaller transformations. Taken together, these layers transform input data — raw text, images, sound files, and other unstructured formats — into predictive outputs that capture semantic features detected in the original data. While the internal structures of deep learning models make them ideal for certain tasks, these benefits come at a significant cost. State-of-the-art models contain dozens of layered models and billions of numerical parameters. This complexity makes their inner workings impossible to fully understand, even by experts in the field.

In order to understand what exactly is meant by "deep learning", we argue that the "depth" explicit in the term has a triple meaning: *knowledgeable*, the accuracy displayed in the model's ability to excel in certain image process tasks, *layered*, a visualization of the learned hierarchical structures, and *impenetrable*, the inherent lack of interpretability and understanding (such as in the "deep sea" or "deep space") of their algorithmic operations. In this chapter, we interrogate these three meanings and then argue that all three are intricately linked to each other. There is no way to achieve the observed levels of accuracy without constructing layered models and introducing

[1] Wu, Xiaolin, and Xi Zhang, "Responses to Critiques on Machine Learning of Criminality Perceptions," *arXiv preprint arXiv:1611.04135* (2016): 1-14. Wang, Yilun, and Michal Kosinski. "Deep Neural Networks are More Accurate than Humans at Detecting Sexual Orientation from Facial Images," *Journal of personality and social psychology* 114, no. 2 (2018): 246.

black-box methods. Further, there is an intrinsic depth to the tasks in domains where deep learning models are applied. That is, depth is not just an instrumental feature of working with text and image data. The deepness is inherent in the tasks themselves. Meaningful computational results in these domains require a deep learning approach. Finally, we relate the essential deep nature of certain computational tasks to implications for future study in the humanities and social sciences and to the proliferation of deep learning models throughout society. We will limit our analysis to the task of processing image and text data, as they are particularly well suited for deep learning and also are a primary object of study for humanists and social scientists.

Knowledgeable

Deep learning techniques are now used in nearly all subfields of machine learning but are most well known for their application within predictive modelling. Predictive models make use of tagged datasets to algorithmically discover patterns that can be used to predict tags for new objects outside of the original collection. A classic example consists of starting with a collection of emails tagged as being "spam" or "not spam" and finding patterns that can be used to automatic spam detection on new messages. There are several classes of powerful, general purpose predictive models that are frequently used in machine learning. Linear regression, support vector machines, and gradient boosted trees have all been shown to produce reliably predictions within a wide range of applications.[2] These models struggle, however, on some important classes of problems, most notably when applied to tasks involving the processing of text and image data. Processing unstructured inputs such as raw text and images are precisely the types of problems where deep learning models excel.

The inherent difficulty of building predictive models with text and image data can be understood from two related perspectives. First, the way that text and image data are stored digitally does not directly capture semantic meaning. Consider, as a point of contrast, the task of predicting the sale price of a work of art. Features that may be available to determine the price include: who created the work of art, the medium of the object, the original date of creation, and its overall size. Each of these values measures real-world quantities that directly impact the value of the work. Compare this to the task of building a model that detects sarcasm in a corpus of text or determines the identities of people depicted in a collection of photographs. What features will be available for these tasks? Machine readable text is stored as a stream of characters. Digital images are represented as three rectangular grid of pixel intensities (red, green, and blue). Unlike the semantic variables describing the sale price of works of art, individual characters and pixels are essentially meaningless in isolation. It is only in context that we comprehend the significance of the textual or visual message. Further, there is no obvious alternative representation that would map directly into a semantic meaning.[3]

[2] Hastie, Trevor, Robert Tibshirani and Jerome Friedman, *The Elements of Statistical Learning*, *2nd ed.* (New York: Springer, 2009).

[3] This is not entirely true for textual data. The characters could be grouped together into words in a process known as *tokenization* and words do have some intrinsic meaning. Many text processing tasks, such as spam detection and authorship attribute, can be accurately modelled with general-purpose algorithms using only word counts. However, more advanced tasks such as document summarization and automatic language translation require taking word order and high-level grammatical features into account. There is no such analogue for image data and the problem remains for even relatively simple tasks.

A second, closely related, challenge of working with text and image data concerns the machine learning concept known as *dimensionality*. Textual data are represented as a stream of characters; however, the converse does not hold: many streams of characters are not (understandable) textual documents. In fact, only a very small proportion of randomly constructed streams of characters will result in readable text. Similarly, almost no random constructed rectangular grids of pixels will resemble a recognizable image. Most will look like static noise. This creates a challenge for predictive models because the majority of possible inputs, random streams of characters or grids of pixels, fail to be sensible objects for consideration in the first place. Therefore, a predictive model must simultaneously detect the hidden structures within text and image data while also predicting the specific tag of interest. This task turns out to be very difficult but well-suited to deep learning approaches, the specifics of which we discuss in the next section.

Text and image processing, in addition to being difficult objects of study in machine learning, share another common feature: the human brain seems particularly well-designed for both tasks.[4] The power of billions of interconnected neurons firing signals to one another inspired the neurophysiologist Warren McCulloch and logician Walter Pitts to design a computational model in which signals are passed between independent nodes using a threshold potential similar to the biochemical functioning of neurons.[5] The approach of McCulloch and Pitts, applied to predictive modelling tasks, is considered the genesis of the class of models known as *neural networks*, the earliest example of a deep learning algorithm.[6] Early work on neural networks was heavily integrated with neurophysiology. Modern developments have diverged sharply from biological motivations to the point where "state-of-the-art deep learning algorithms rely on mechanisms that seem biologically implausible."[7] Despite this disconnect, the language of neurology — *neurons*, *neural* networks, *potential*, long-term *memory, activation* functions, *developmental* networks — remains dominant within the machine learning community. Partially this is a result of momentum from the earliest research, but today also serves as a strong cultural signal that deep learning represents, more than alternatives, "real" human-like intelligence.[8]

Interest in neural networks has varied over time. Early excitement was dampened by the negative results of Minsky and Papert, and the inability to train large networks with the computational resources available at the time.[9] Advances in the 1980's and 1990's addressed some of these

[4] While interesting in its own right, the contentious evolutionary details behind these two strengths, and the specific pathways for language acquisition such as Chomsky's language acquisition device (LAD), are not important for the discussion here. For more details on both, see: Thorpe, Simon, Denis Fize, and Catherine Marlot, "Speed of Processing in the Human Visual System," *Nature* 381, no. 6582 (1996): 520. Chomsky, Noam, *Aspects of the Theory of Syntax* (Boston: MIT Press, 1965).

[5] McCulloch, Warren S., and Walter Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *The Bulletin of Mathematical Biophysics* 5, no. 4 (1943): 115-133.

[6] In fact, neural networks are arguably the *only* practical example of a deep learning algorithm. Often the terms are used interchangeably. In this article, "deep learning" is used to describe the general modelling approach and the term "neural networks" is used only to refer to specific applications of deep learning.

[7] Bengio, Yoshua, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard, and Zhouhan Lin. "Towards Biologically Plausible Deep Learning," *arXiv preprint arXiv:1502.04156* (2015), 1.

[8] In an effort to reclaim the physiological legitimacy of neural networks, some have called for a renewed partnership with neuroscience, with minimal recent success. See: Marblestone, Adam H., Greg Wayne, and Konrad P. Kording. "Toward an Integration of Deep Learning and Neuroscience," *Frontiers in Computational Neuroscience* 10 (2016): 94.

[9] Minsky, Marvin; Papert, Seymour, *Perceptrons: An Introduction to Computational Geometry* (Boston: MIT Press, 1969).

concerns and led to several well-known examples, including LeCun's classification of hand-written digits.[10] However, continued computational challenges and lack of strong empirical motivations for neural network models held off general interest until very recently.

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an image classification contest held annually since 2010 in which teams compete to build algorithms that classify images into one of one-thousand categories.[11] In 2012, building off of nearly a decade of work refining neural network architectures, Alex Krizhevsky, Ilya Sutskever and Geoffry Hinton produced a winning neural network model — commonly known as AlexNet — that had a top-5 error rate of only 16%, compared to the 25% top-5 error rate achieved by the second place team's model.[12] In 2017, for example, a neural network model achieved an error rate of only 2% on the ILSVRC dataset.[13] Early and continued success on ILSVRC is largely seen to have launched the deep learning "revolution in computer vision", which continues with no sign of slowing down anytime soon.[14] Today, the vast majority of research in predictive models for computer vision is built on neural networks. Text analysis had at first been slower to adopt deep learning, but neural networks have recently become popular in the processing of text too. Neural networks have produced state-of-the-art results in machine translation, sentiment analysis, and topic classification.[15]

The popularity of deep learning models is a direct result of their unmatched power to produce predictive models for difficult tasks such as text and image processing. In other words, their ability to build off of existing knowledge to predict new knowledge. It is for this same reason that deep learning is an important object for humanistic study. Neural network applications are not contained to relatively obscure academic competitions; rather, they are already being employed today behind the scenes in a wide variety of applications. Some of these applications directly serve the public good, such as advances in the automated detection and classification of brain tumors from MRI scans.[16] Others play directly into the needs of mass-surveillance.[17] The power of deep learning allows for the automation of wide-scale privacy invasions for national and capitalistic motivations, without the limiting cost of human annotation. It is likely that many of the technological advances

[10] LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation* 1, no. 4 (1989): 541-551.

[11] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang. "ImageNet: Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* 115, no. 3 (2015): 211-252.

[12] For each image, the model predicts the 5 most likely categories the image corresponds to and is "correct" if the true category is one of these five. Leeway is provided because some categories are incredibly difficult even for expert human labellers, such as two closely related dog breeds or edge-cases over the distinctions between a generic "cup" and a "coffee cup". Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* (2012), 1097-1105.

[13] Hu, Jie, Li Shen, and Gang Sun, "Squeeze-and-Excitation Networks," *arXiv preprint arXiv:1709.01507 7* (2017), 1-11.

[14] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature* 521, no. 7553 (2015): 440.

[15] Howard, Jeremy, and Sebastian Ruder, "Universal Language Model Fine-Tuning for Text Classification," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* no. 1 (2018): 328-339.

[16] Havaei, Mohammad, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle, "Brain Tumor Segmentation with Deep Neural Networks," *Medical Image Analysis* 35 (2017): 18-31.

[17] Levi, Gil, and Tal Hassner, "Age and Gender Classification Using Convolutional Neural Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2015): 34-42.

of the near future, such as self-driving cars, will be built on top of deep learning models.[18] In order to understand how these extant and future applications affect society, it is necessary to also understand the deep learning models themselves. The next section proceeds to explain the internal architecture of deep learning models and the relationship of this structure to their observed predictive strengths.

Layered

The focus of the discussion so far has been on the impressive predictive power of deep learning models in the difficult domains of text and image processing. Other than the original connection of neural networks to neurophysiology, which has largely been lost in modern developments, we have not explained why deep learning is particularly well-adapted to these applications. It is this task that we now address. As a starting point, a suitable definition of deep learning is required.

Deep learning models apply a sequence of successive transformations to an input object of study and ultimately produce a modified output value. In the predictive modelling context, the final outputs are the predicted tags and the transformations are adaptively learned by a training algorithm applied to a large collection of pre-tagged objects. Each transformation should, at least in theory, assist in moving from “raw” input formats, such as pixel intensities or character streams, towards meaningful features that capture semantic meaning within an image or textual document. Typically, the first few transformations only consider interactions between small groups of nearby pixels or characters. Successive transformations are applied to larger “windows” of the object, with the final layers applied to the entire image.[19] An example of a particular, highly idealized, deep learning model is useful to further explain the concept.

Consider the example image in Figure 1 and the selected boxes of interest. The first few layers of a neural network may only look at nearby swatches of the image and convert the raw pixels into numbers that at first describe their overall color and shading. A slightly larger view provides information about the texture of the grass and edges that make up the nose of the man. Subsequent layers reveal small objects (nose), larger objects (faces), and finally objects within their context. A final layer, not shown within the boxes, could be applied to the entire image to aggregate information about the individual objects. This layer would capture features about the scene as a whole. Figure 2 shows a similar linguistic example. Subsequent layers of the neural network look at larger windows of the text by grouping characters into words, words into phrases, phrases into sentences, and sentences in entire documents.[20] The layered nature of deep learning models, likely the original motivation behind the term “deep”, is the fundamental feature differentiating them from other approaches and directly addresses the representational issues presented as the primary challenges to working with text and image data.

It is not an easy task to build deep learning models from scratch. Dozens of complex transformations operating seamlessly together must be created and iteratively modified in an

[18] Ramos, Sebastian, Stefan Gehrig, Peter Pinggera, Uwe Franke, and Carsten Rother, “Detecting Unexpected Obstacles for Self-Driving Cars: Fusing Deep Learning and Geometric Modelling,” *Intelligent Vehicles Symposium,* (2017): 1025-1032.

[19] Goodfellow, Ian, and Yoshua Bengio, *Deep Learning* (Boston: MIT Press, 2016).

[20] The models described here are, more precisely, examples of convolutional neural networks, or CNNs, which are frequently used in both text and image analysis.

attempt to address the most challenging problems in predictive modelling. Very large datasets are required in order to estimate the hundreds of millions of parameters that describe state-of-the-art neural networks. Public datasets for training image models, such as the ILSVRC challenge and VGGFace2 dataset, typically provide millions of tagged images to produce accurate face-detection models.[21] Once a large dataset is assembled, special hardware in the form of expensive and powerful graphical processing units (GPUs) are required in order to process such large datasets through the complex architectures of modern neural networks.[22] Even with good training data and the required hardware, the actual construction of neural networks is still a significant challenge. Adjusting the millions of training parameters is known to be an incredibly fraught task; subtle changes to the structure of the network can drastically alter the output of the model.[23] It would appear that the power of neural networks may be restricted to well-funded companies or research groups and available only for a small set of high-impact tasks for which the payoff in time and money is worthwhile. In practice, this is far from the case due to the special layered structure of deep learning models.

When trained on sufficiently large image datasets, the initial transformations described by large neural networks tend to be generalizable to new problems unrelated to the original prediction task. Recall that the first layers in a neural network only work locally over small regions of an image. These initial layers detect general features such as shading, color, and texture. Even layers in the middle of the network correspond to rough shapes and the formation of larger objects. It is only the last few layers that are directly related to the specific predictive modelling tasks of interest. As a result, predictive neural networks can be adapted through the process of *transfer learning* to predict new outputs by re-using the trained values in the interior layers and only learning the form of the final 1-3 transformations. This drastically reduces the amount of data, hardware, and expertise required to construct a new model. For example, recently a research group built a highly-predictive image processing neural network using a set of only 443 frontal chest x-ray images through transfer learning applied to the AlexNet model.[24] They copied all but the final layer of the network and trained the relatively small set of final weights with their own data. The ability to perform transfer learning, which drastically increases the number of feasible applications of deep learning, is another direct feature of the layered nature of the models.

The underlying idea of transfer learning — that interior layers in neural networks code generic features that can be adapted to new problems — can also be used to motivate the related concept of *embeddings*. An embedding applies a selection of lower level transformations from a neural network to an object of interest, the output of which can be viewed as a sequence of numeric values. In transfer learning a predictive model is built on top of these embedded values, but there is also intrinsic value in the embedding itself. Embeddings have, for example, recently received

[21] Cao, Qiong, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman, “Vggface2: A dataset for Recognising Faces Across Pose and Age,” *Automatic Face & Gesture* (2018): 67-74.

[22] Li, Haoxiang, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua, “A Convolutional Neural Network Cascade for Face Detection,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015): 5325-5334.

[23] Goldberg, Yoav, “A Primer on Neural Network Models for Natural Language Processing,” *Journal of Artificial Intelligence Research* 57 (2016): 345-420.

[24] Bar, Yaniv, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, and Hayit Greenspan, “Chest Pathology Detection Using Deep Learning with Non-Medical Training,” *International Symposium on Biomedical Imaging* (2015): 294-297.

attention in applications as diverse as cognitive psychology and the digital humanities.[25] The numbers described by the embedding captures, according to our description of how neural networks function, important semantic features present within the input image or textual document. Each number in the embedding does not directly correspond to a meaningful quantity. Rather, the spatial relationships of the objects within the embedding capture various semantic meanings. Inputs with similar features, in particular, will have similar sequences of numbers. Connecting objects that have similar embeddings to one another has been shown to provide accurate image and document similarity metrics that require no manual tagging or re-training.[26] As an example, Figure 3 illustrates a two-dimensional word embedding for a small collection of words. The full embedding these are taken from were trained on the English-language text from Wikipedia containing 300,000 words arrange in 300-dimensional space.[27] Food items and verbs / professions are visibly separated in the embedding space. Pairs of closely related terms, such as journalist-writer, believe-understand, and read-write, are embedding next to one another. Also, the profession "chef" is situated closer to the food items than any other As with transfer learning, the feasibility of embeddings are directly tied to the layered nature of deep learning models.

The layered structure of deep learning models is a direct consequence of the challenges posed by the processing of image and textual data. Without the sequential application of transformations, deep learning would offer no immediate benefit to predictive modelling to these difficult classes of important machine learning problems. The layers also immediately make way for the important application of transfer learning and embeddings, without which deep learning models would be inaccessible to all but a small number of applications. Unfortunately, the layered nature also comes at the cost of interpretability. It is notoriously difficult, if not outright impossible, to comprehend how neural networks achieve their amazing predictive results. As we witness a proliferation of neural network applications in society, our inability to understand exactly what they are doing poses a number of concerns.

Impenetrable

Modern neural networks for text and image processing typically consist of dozens of layered transformations and hundreds of millions of learned parameters. Our description above of how neural networks transform images, by successively detecting larger and larger regions of interest and stitching them together, is a highly idealized version of how networks actually function. The general concepts have been validated through the efficacy of transfer learning and visualizations of embedding spaces, but the specific meaning of any given internal representation is generally impossible to discern. Because of the complex dependencies present within the layers of a neural network, classic approaches to interrogating a particular model, such as applying small perturbations to a single parameter and watching the result, are rarely very enlightening. The general lack of interpretability in deep learning is a well-known problem; several recent workshops

[25] Patwardhan, Siddharth, and Ted Pedersen, "Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts," *Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together* (2006): 1-8.

[26] Lau, Jey Han, and Timothy Baldwin, "An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation," *arXiv preprint arXiv:1607.05368* (2016): 1-11.

[27] Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou, "Word Translation Without Parallel Data," *arXiv preprint arXiv:1710.04087* (2017): 1-12.

were specifically dedicated to papers on the interpretability of neural network models.[28] A collection of approaches have been proposed in an attempt to build an understanding of how neural networks function.

One approach for understanding the inner workings of neural networks is to focus on the objects in a predictive modelling task that are incorrectly tagged. One simple approach is to investigate those categories that have particularly high error rates. Does the model have trouble with a specific category or does it struggle to differentiate between a certain set of objects? Analysis of the results can be insightful in understanding the internal mechanisms of the neural network. For example, the ILSVRC challenge revealed that animals with distinctive furs (e.g., foxes, porcupines, and tigers) are particularly easy to classify. On the other hand, long slender objects (e.g., letter openers, flagpoles, and water bottles) are typically the hardest to detect. Abstract concepts such as "restaurant" and "grocery store" are also amongst the most difficult categories for algorithms to distinguish.

Taken together, this evidence shows that neural networks are best at understanding localized features and struggle the most on categories that require putting together contextual knowledge across the entire image. Looking at specific objects that are misclassified by a model, the *negative examples*, is another method of understanding the behavior of neural networks. For example, an investigation of the negative examples from the GoogLeNet model — the 2014 winner of ILSVRC — showed particular difficulty with "images that contain multiple objects, images of extreme closeups and uncharacteristic views, images with filters, images that significantly benefit from the ability to read text [a salt shaker], images that contain very small and thin objects [fishing reel], images with abstract representations."[29] These highlight challenges in the existing model and suggest significant gaps between the way the model understands images and human-like processing of visual data. Ongoing research in computer vision is often motivated by understanding where, and ideally why, current models fail on certain tasks.

Alternatively, another approach to understanding neural networks is to focus on objects where the model performs well. The motivation behind the use of neural networks is their incredible predictive power. It seems reasonable that if we want to understand how neural networks function some attention should be paid to the many objects that are correctly tagged. A clever approach to studying these *positive examples* is to occlude part of the object and observe the extent to which these occlusions effect the predicted categories. For an image, this involves replacing a region of the image with a monochromatic box, effectively hiding a region of the image from the neural network. [30] In text analysis, a similar approach removes one or more words or phrases.[31] Visualizing the regions that most directly impact the predicted values, and quantifying how much of the text or image can be removed without significantly impacting the results, provides an

[28] For example, the NIPS 2017 workshop on "Interpreting, Explaining and Visualizing Deep Learning... Now What?" and the EMNLP 2018 workshop "Analyzing and Interpreting Neural Networks for NLP."

[29] Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang. "ImageNet Large Scale Visual Recognition Challenge," 243.

[30] Zeiler, Matthew D., and Rob Fergus, "Visualizing and Understanding Convolutional Networks," *European Conference on Computer Vision* (2014): 827.

[31] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016): 1135-1144.

additional understanding of how the neural network represents knowledge. Similarly, the embeddings of correctly classified categories can be investigated for each layer of a neural network. Identifying which layers separate specific categories provides a window into the specific role of each layer in the overall prediction task.

Despite the existence of techniques for understanding neural network architectures, there remains a fundamental inability to understand how the network performs the task of transforming inputs into reliable predictions. Negative examples elucidate those abstract features that are generally missed by the network. Positive examples, along with occlusion, hint at those regions and features that are captured by the model. How these features are captured, however, remains a mystery. The fundamental trouble is the depth of the model. Each layer is co-dependent on all of the others and understanding the network therefore requires understanding the entire network all at once, which is impossible given the size and depth of modern neural networks. And the problem is only getting worse.

Over time, neural networks have grown deeper and more complex. The wining ILSVRC model "ResNet" from 2015 had a total of 152 layers (for comparison, AlexNet has only 8 layers).[32] By 2016, the ResNet model had expanded to a total of 1000 layers.[33] Popular models for text analysis now commonly employ recurrent neural networks, which contain complex architectures for storing "memories" as the network cycles through characters and words within a document. As further evidence to our inability to understand neural network models, recent research has revealed strange and unintuitive results from seemingly powerful predictive models. Carefully constructed perturbations can be applied to an image that is imperceptible to the human eye but cause arbitrarily large changes in the predicted tags associated with the image.[34] Conversely, images that appear to be pure noise can be found that are confidently categorized with an extremely high probability on one particular tag. These examples point to a significant gap in the way that neural networks process data compared to the human processing of images and text.

As neural networks become integrated into systems that directly affect people, it becomes increasingly important to understand how deep learning models function. Most of the work on interpretability has so far focused on understanding neural network in order to modify their architecture and improve the predictive power of future models. However, it is arguably even more important to understand the models from a social perspective. How can we be convinced that an algorithm for tracking passengers at airport security are not motivated by racial profiling? What checks exist to detect when models employed by the medical industry are being optimized for insurance money rather than patient health? Or, what confidence do we have that autonomous vehicles trained in sunny California will accurately deal with snowy New England winters? All of these questions can be addressed on a macroscopic scale through external validation and regulatory

[32] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016): 770-778.

[33] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity Mappings in Deep Residual Networks," *European Conference on Computer Vision* (2016): 630-645.

[34] Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard, "Deepfool: A Simple and Accurate Method to Fool Deep Neural Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016): 2574-2582.

transparency, but it becomes difficult for an individual to trust or challenge the results of a specific model that alludes any direct ability to understand its internal mechanisms.

While many machine learning algorithms have been characterized as being uninterpretable black boxes, our characterization of neural networks as impenetrably difficult to understand draws on features unique to deep learning methods. Figure 4 displays a schematic representation of a “shallow” — a model that is not deep — predictive model. A collection of transformations is applied to the raw input data and then combined together again to produce the output classifications. For comparison, Figure 5 provides a diagram of a deep learning model in which transformations are sequentially applied to the input data in order to yield the output categories. Shallow models may become quite complex if there are a large number of transformations. For example, models such as boosted trees also often involve millions of parameters and it can be difficult to understand how these parameters come together to produce a final set of predicted tags.[35] However, shallow models can be decomposed into individual elements that each act independently on the input variables and produce distinct contributions to the output classification values. On a local level, at least, there is a possibility for understanding how regression and tree-based models construct predictions from their inputs. In contrast, the layered structure of deep learning models makes even this level of understanding impossible. The lack of a local understanding makes it difficult to assess the structure of a neural network and determine whether a specific application is (algorithmically) reasonable or advisable. The layered structure, then, is a fundamental cause of the impenetrable nature of deep learning models.

Unavoidable trichotomy

We have shown that deep learning models exhibit their depth along three alternative meanings of the term “deep”. They exhibit a deep knowledge in understanding image and textual data by producing accurate labels for a range of predictive modelling tasks. This predictive power is achieved through structures that consist of a deep succession of transformations that gradually push the input objects towards the predicted output tags. Finally, these chains of interrelated transformations hide the parameters of the model with an impenetrable depth that obscures exactly how they arrive at their results. Crucially, we have seen that these three elements are related by far more than the polysemous nature of the English word “deep”. The layered nature of deep learning models is a necessary feature for their ability to make predictions for hard tasks such as text and image processing, and these layers in turn are fundamentally difficult to interpret. From the unavoidable interdependence between these elements of deep learning, we conclude here with implications for continued study of deep learning as an object of humanistic inquiry.

First, it is important to start describing the concept of a *deep problem* in addition to deep learning. The way in which text and image data are stored, as streams of characters and pixel intensities, necessitate the use of layered models that modify the original data and represent objects within a new space. In other words, the nature of working with text and image data requires deep learning models in order to achieve high levels of accuracy. Analysis of these objects is an intrinsically *deep problem*, irrespective of the specific models used to study them.

[35] Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad, “Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission.” *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015): 1721.

The classification of tasks, rather than the algorithms for performing them, as "deep" is important. It signals that many of the challenges underlying modern machine learning are actually problems of how knowledge is transmitted and represented. In turn, this directly draws connections from machine learning into well-trodden areas of humanistic inquiry such as epistemology, semiotics, and communication theory. Saussure's assertion that words draw meaning in their "simultaneous coexistence of all the others" to other words maps directly into the model of a word embedding.[36] Autoencoders, a particular class of neural networks, provide a mathematical formulation directly related to Stuart Hall's encoding/decoding model of social communication.[37] Panofsky's tripartite series of levels to understanding works of art, which starts with the literal subject matter and proceeds up through iconography and iconology, mirrors the layered hierarchy of levels in modern convolutional neural networks and hints at the application of transfer learning. [38] The first level of interpretation is, in theory, universal; cultural considerations become more explicit higher up the chain of understanding. In short, by focusing on the tasks and not there explicit solutions we find numerous points of contact between predictive modelling tasks and ongoing questions in a wide range of other fields. Providing points of connection across fields allows for more productive critiques and fruitful interdisciplinary interactions.

As a second implication, embeddings — the output of a particular sequence of transformations within a deep learning model — should be considered as an object of study in its own right. We have argued that the intermediate representations offered by the internal layers of a neural network do encode generalizable semantic information that can be utilized in new tasks through transfer learning. The internal representations, unlike the raw inputs, more directly contain useful semantic information that have the potential to allow for the utilization of shallow models, even for many complex tasks.[39] If we are able to find good general purpose embeddings that work with shallow models, this would help alleviate some concerns about the lack of interpretability in deep learning models. While the process of converting raw inputs into the embedding space may remain opaque, with time and analysis, a direct characterization of the embedding space could be achieved. Several results suggest that a universal embedding, or a close approximation to one, could be attainable. In image processing, it has been shown that significant sequences of layers in neural networks can be *inverted* to recreate "photographically accurate information" about the image, establishing that relatively information is being lost in (at least the lower level) transformations.[40] For text analysis, where transferable embeddings initially proved more difficult to detect, recent work has produced several candidates that appear to generalized very well to a variety of tasks.[41] The culture in deep learning research of making research, code, and datasets openly available is a great start for making it possible to offer meaningful studies of embedding spaces. We now need more scholars actively engaged in treating these embeddings as an important object of study.

[36] Saussure, Ferdinand de, *Cours de Linguistique Générale,* trans. Roy Harris (Chicago: Open Court, 1998): 159.

[37] Hall, Stuart, "Encoding/decoding," *Culture, Media, Language* (1980): 128-138.

[38] Panofsky, Erwin, *Studies in Iconology. Humanistic Themes in the Art of the Renaissance*, trans. Gerda S. Panofsky (New York: Routledge, 1972).

[39] That is, models that are shallow in respect to the embedding layer. They are still deep relative to the raw text or image input data.

[40] Mahendran, Aravindh, and Andrea Vedaldi, "Understanding Deep Image Representations by Inverting Them," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015): 5188-5196.

[41] Howard, Jeremy, and Sebastian Ruder, "Universal Language Model Fine-Tuning for Text Classification." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* no. 1 (2018): 328-339.

Finally, it is also important to train scholars from a wide range of fields in the technology of deep learning, specifically neural networks. Layered models that successively reparametrize raw inputs are, as we have seen, necessary for sufficiently predictive models. Deep learning techniques will likely remain popular for a considerable amount of time and are poised to become even more integrated into important real-world systems. We need domain experts from fields such as medicine, biology, public policy, law, economics, and across the humanities to understand this technology. To do so opens up avenues for both important innovations as well as meaningful critiques of current practices. As we have shown, deep learning models are difficult enough to comprehend even for those working in the field of machine learning who have been working with them for decades. Due to this complexity, meaningful collaborations between domain experts and researchers in deep learning require a working understanding of the power and challenges of neural networks across disciplinary boundaries.

Deep learning approaches are here to stay. They offer amazing predictive accuracy and a plethora of exciting technological advances, but also make way for a wide range of troubling applications. As deep learning becomes increasingly ubiquitous in real-world systems, the unavoidable trichotomy between knowledge, layers, and a lack of interpretability has important implications for anyone concerned with the use and proliferation of algorithmic logic in society.[42] Direct humanistic inquiry into the algorithms behind deep learning is needed as we grapple with their cultural and social implications.

[42] Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. (New York: NYU Press, 2018).

Bibliography

Bar, Yaniv, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, and Hayit Greenspan, "Chest Pathology Detection Using Deep Learning with Non-Medical Training," International Symposium on Biomedical Imaging (2015): 294-297.

Bengio, Yoshua, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard, and Zhouhan Lin. "Towards Biologically Plausible Deep Learning," arXiv preprint arXiv:1502.04156 (2015), 1.

Cao, Qiong, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman, "Vggface2: A dataset for Recognising Faces Across Pose and Age," Automatic Face & Gesture (2018): 67-74.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad, "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015): 1721.

Chomsky, Noam, Aspects of the Theory of Syntax. Boston: MIT Press, 1965.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou, "Word Translation Without Parallel Data," *arXiv preprint arXiv:1710.04087* (2017): 1-12.

Goodfellow, Ian, and Yoshua Bengio, Deep Learning. Boston: MIT Press, 2016.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman, The Elements of Statistical Learning, 2nd ed. New York: Springer, 2009.

Havaei, Mohammad, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle, "Brain Tumor Segmentation with Deep Neural Networks," Medical Image Analysis 35 (2017): 18-31.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016): 770-778.

———, "Identity Mappings in Deep Residual Networks," European Conference on Computer Vision (2016): 630-645.

Howard, Jeremy, and Sebastian Ruder, "Universal Language Model Fine-Tuning for Text Classification," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, no. 1 (2018): 328-339.

Goldberg, Yoav, "A Primer on Neural Network Models for Natural Language Processing," Journal of Artificial Intelligence Research 57 (2016): 345-420.

Hall, Stuart, "Encoding/decoding," Culture, Media, Language (1980): 128-138.

Howard, Jeremy, and Sebastian Ruder, "Universal Language Model Fine-Tuning for Text Classification." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, no. 1 (2018): 328-339.

Hu, Jie, Li Shen, and Gang Sun, "Squeeze-and-Excitation Networks," arXiv preprint arXiv:1709.01507 7 (2017), 1-11.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems (2012), 1097-1105.

Lau, Jey Han, and Timothy Baldwin, "An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation," arXiv preprint arXiv:1607.05368 (2016): 1-11.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," Nature 521, no. 7553 (2015): 436-444.

LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," Neural Computation 1, no. 4 (1989): 541-551.

Levi, Gil, and Tal Hassner, "Age and Gender Classification Using Convolutional Neural Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2015): 34-42.

Li, Haoxiang, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua, "A Convolutional Neural Network Cascade for Face Detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015): 5325-5334.

Mahendran, Aravindh, and Andrea Vedaldi, "Understanding Deep Image Representations by Inverting Them," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015): 5188-5196.

Marblestone, Adam H., Greg Wayne, and Konrad P. Kording. "Toward an Integration of Deep Learning and Neuroscience," Frontiers in Computational Neuroscience 10 (2016): 94.

McCulloch, Warren S., and Walter Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," The Bulletin of Mathematical Biophysics 5, no. 4 (1943): 115-133.

Minsky, Marvin; Papert, Seymour, Perceptrons: An Introduction to Computational Geometry. Boston: MIT Press, 1969.

Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard, "Deepfool: A Simple and Accurate Method to Fool Deep Neural Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016): 2574-2582.

Noble, Safiya Umoja. *Algorithms of Oppression: How search engines reinforce racism*. NYU Press, 2018.

Panofsky, Erwin, Studies in Iconology. Humanistic Themes in the Art of the Renaissance, trans. Gerda S. Panofsky (New York: Routledge, 1972).

Patwardhan, Siddharth, and Ted Pedersen, "Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts," Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together (2006): 1-8.

Ramos, Sebastian, Stefan Gehrig, Peter Pinggera, Uwe Franke, and Carsten Rother, "Detecting Unexpected Obstacles for Self-Driving Cars: Fusing Deep Learning and Geometric Modelling," Intelligent Vehicles Symposium, (2017): 1025-1032.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016): 1135-1144.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang. "ImageNet: Large Scale Visual Recognition Challenge," International Journal of Computer Vision 115, no. 3 (2015): 211-252.

Saussure, Ferdinand de, Cours de Linguistique Générale, trans. Roy Harris. Chicago: Open Court, 1998.

Thorpe, Simon, Denis Fize, and Catherine Marlot, "Speed of Processing in the Human Visual System," Nature 381, no. 6582 (1996): 520-522.

Wang, Yilun, and Michal Kosinski. "Deep Neural Networks are More Accurate than Humans at Detecting Sexual Orientation from Facial Images," Journal of personality and social psychology 114, no. 2 (2018): 246-257.

Wu, Xiaolin, and Xi Zhang, "Responses to Critiques on Machine Learning of Criminality Perceptions," arXiv preprint arXiv:1611.04135 (2016): 1-14.

Zeiler, Matthew D., and Rob Fergus, “Visualizing and Understanding Convolutional Networks,” European Conference on Computer Vision (2014): 827-838.

Figure 1

Conceptual depiction of features detected by subsequent layers in a neural network when applied to a photograph image (Photo by Greg Parish, "NYC Central Park," 2015; Licensed under CC-BY 4.0).

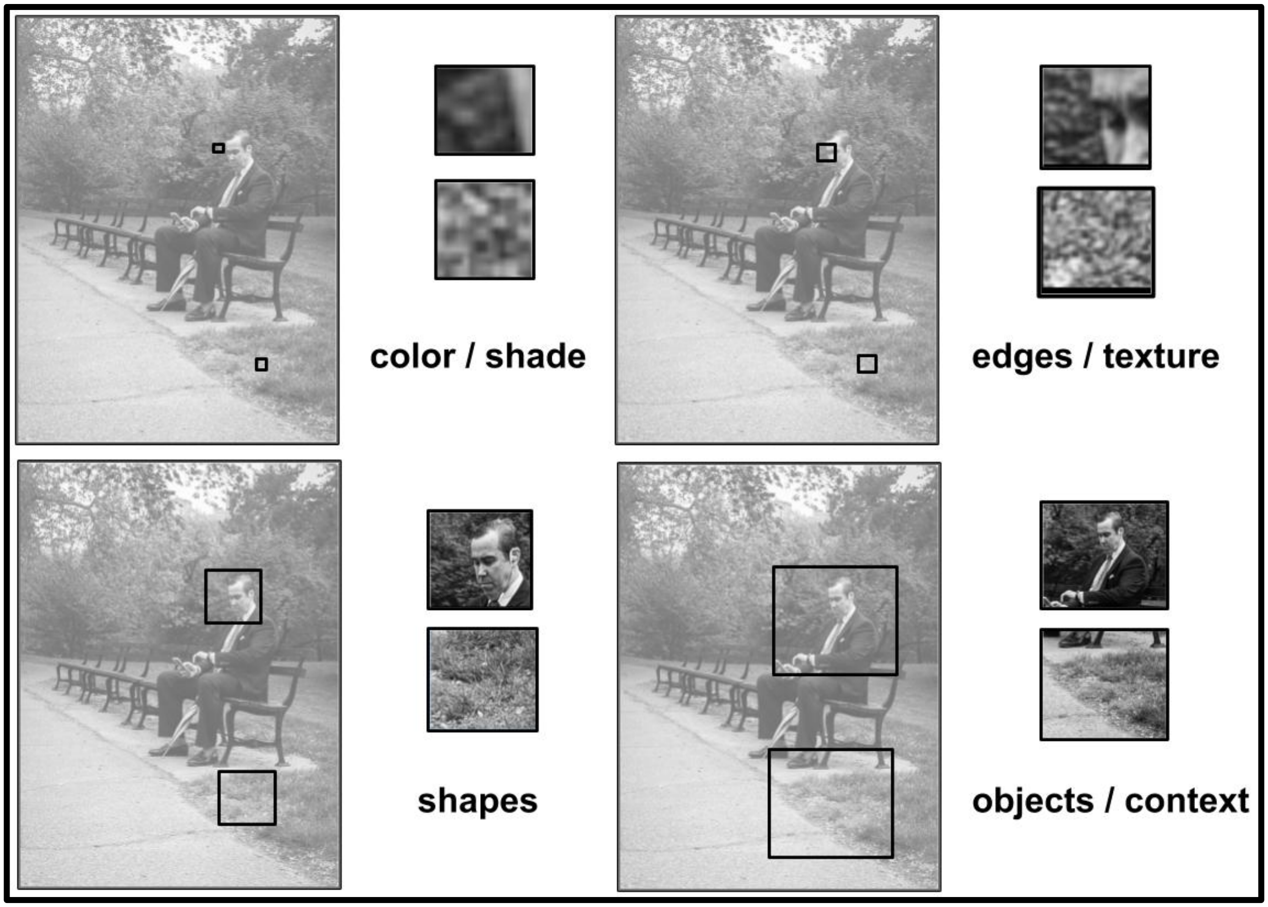


Figure 2

Depiction of hierarchical features detected applying a neural network algorithm to a sentence of textual data.

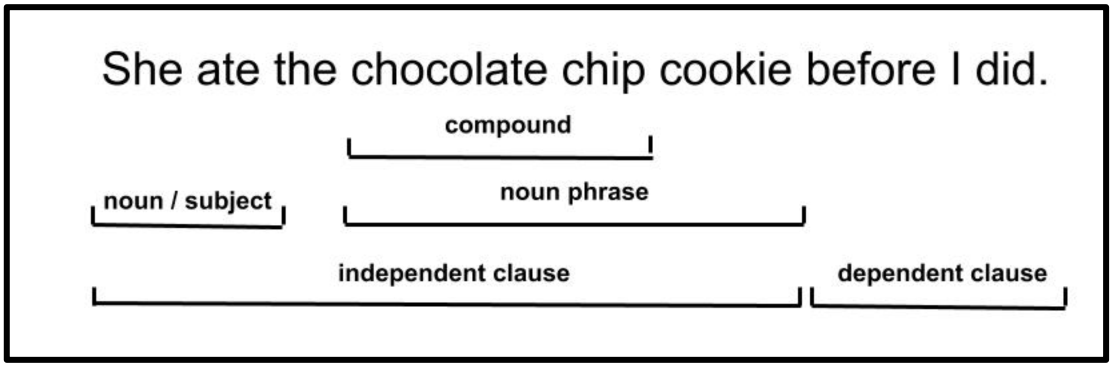


Figure 3

A two-dimensional word embedding of various fruits, vegetables, occupations, and verbs.

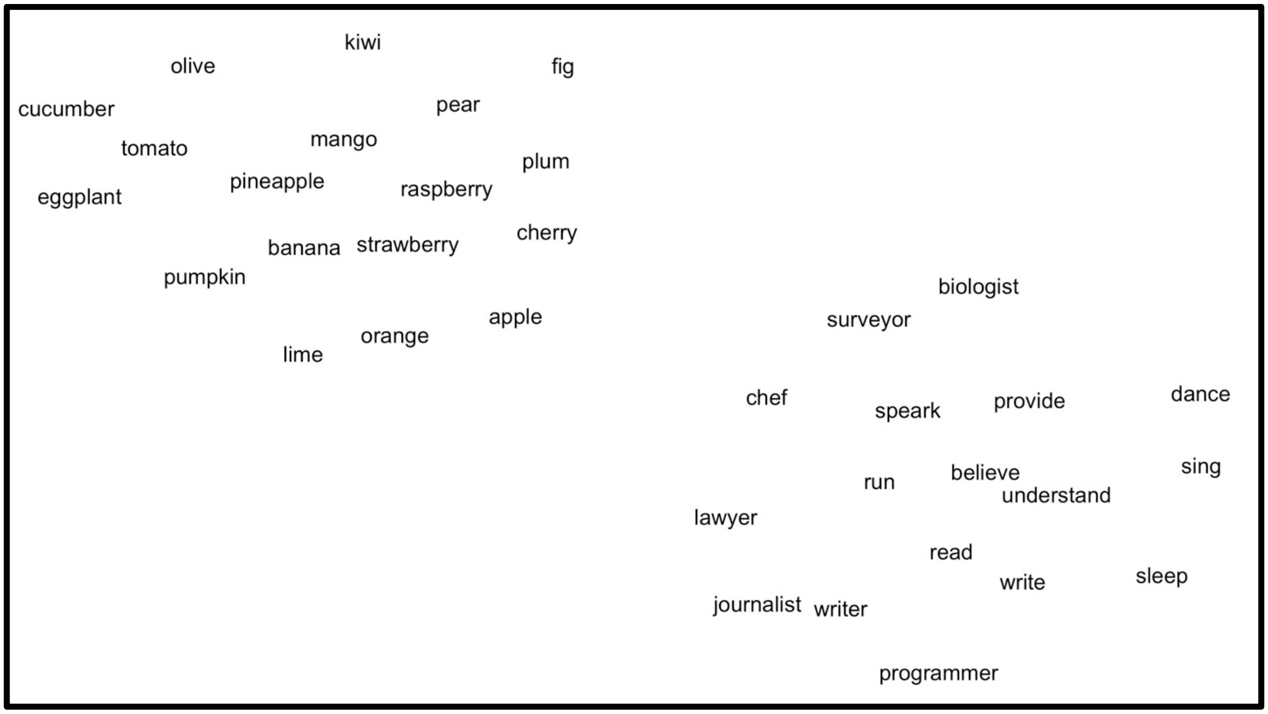


Figure 4

Schematic visualization of a “shallow learning” predictive model.

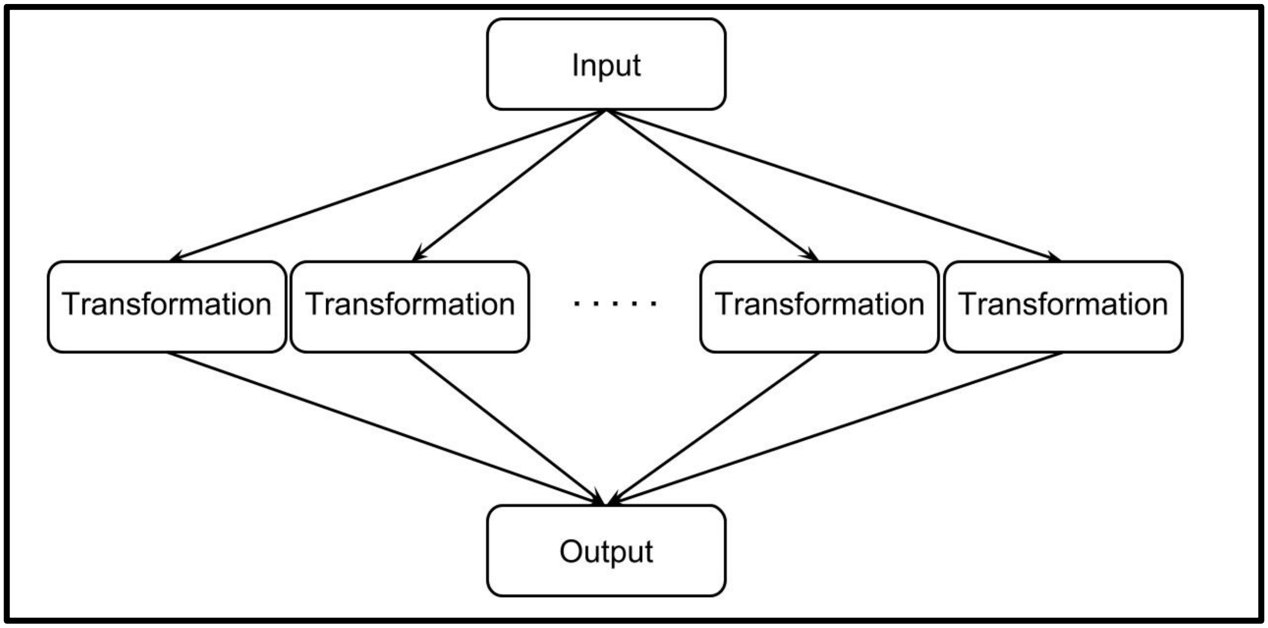
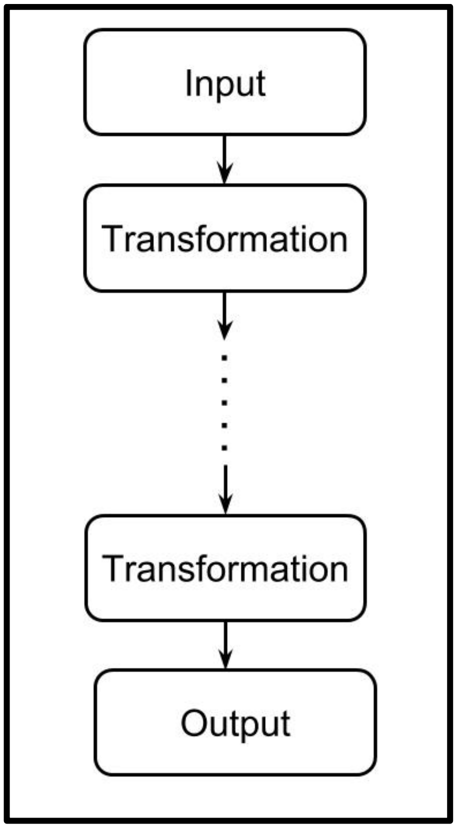


Figure 5

Schematic visualization of a “deep learning” predictive model.


Input
Transformation
Transformation
Output