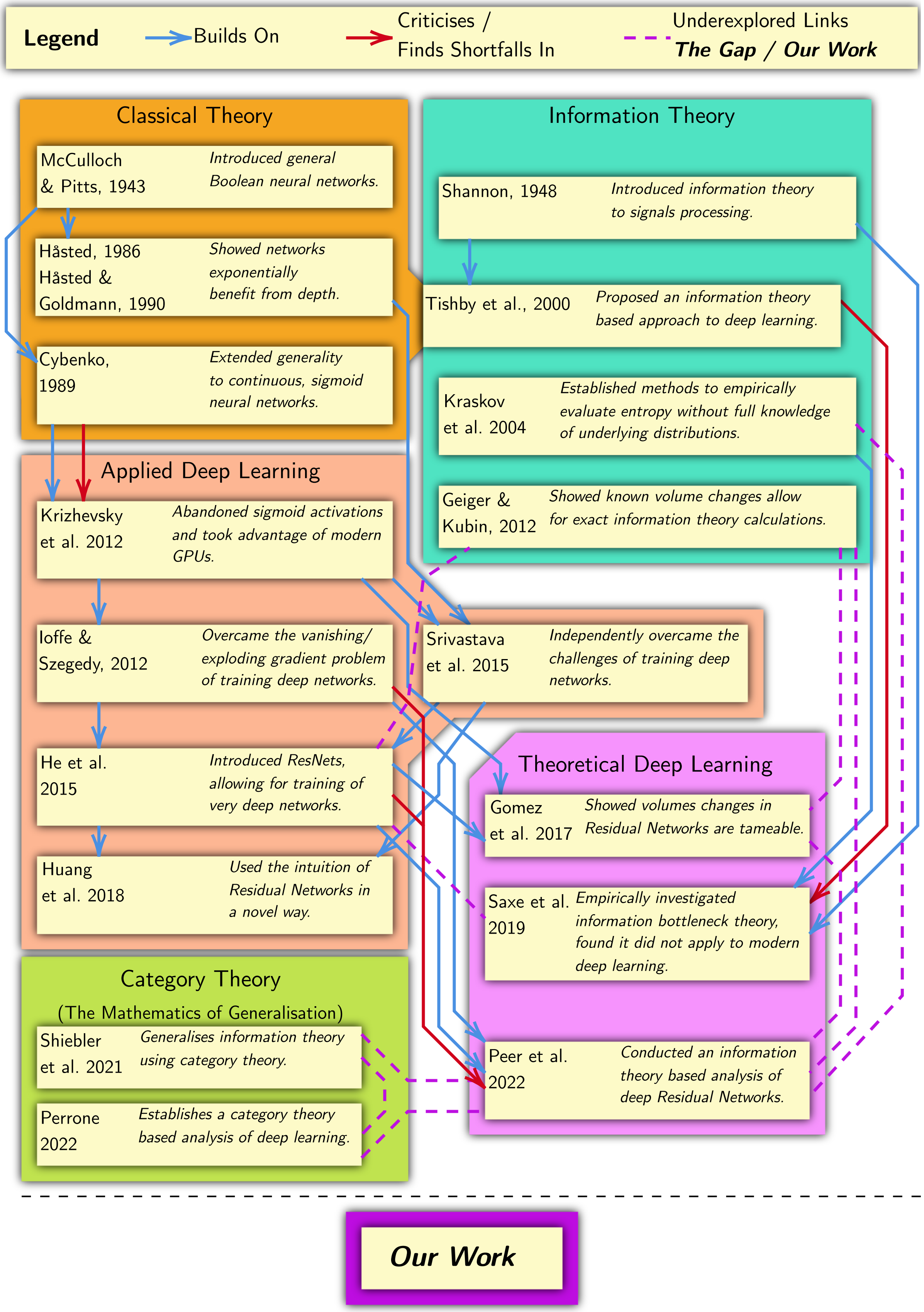
Towards Closing Deep Learning's Theory Gap: Residual Connections and Information Theory

By Vincent Abbott

A Map of The Literature Review



Legend Builds On Criticises / Finds Shortfalls In Underexplored Links *The Gap / Our Work*

Classical Theory

McCulloch & Pitts, 1943 *Introduced general Boolean neural networks.*

Håsted, 1986 Håsted & Goldmann, 1990 *Showed networks exponentially benefit from depth.*

Cybenko, 1989 *Extended generality to continuous, sigmoid neural networks.*

Information Theory

Shannon, 1948 *Introduced information theory to signals processing.*

Tishby et al., 2000 *Proposed an information theory based approach to deep learning.*

Kraskov et al. 2004 *Established methods to empirically evaluate entropy without full knowledge of underlying distributions.*

Geiger & Kubin, 2012 *Showed known volume changes allow for exact information theory calculations.*

Applied Deep Learning

Krizhevsky et al. 2012 *Abandoned sigmoid activations and took advantage of modern GPUs.*

Ioffe & Szegedy, 2012 *Overcame the vanishing/ exploding gradient problem of training deep networks.*

Srivastava et al. 2015 *Independently overcame the challenges of training deep networks.*

He et al. 2015 *Introduced ResNets, allowing for training of very deep networks.*

Huang et al. 2018 *Used the intuition of Residual Networks in a novel way.*

Theoretical Deep Learning

Gomez et al. 2017 *Showed volumes changes in Residual Networks are tameable.*

Saxe et al. 2019 *Empirically investigated information bottleneck theory, found it did not apply to modern deep learning.*

Peer et al. 2022 *Conducted an information theory based analysis of deep Residual Networks.*

Category Theory

(The Mathematics of Generalisation)

Shiebler et al. 2021 *Generalises information theory using category theory.*

Perrone 2022 *Establishes a category theory based analysis of deep learning.*

Our Work

1. Introduction

Deep learning forms the backbone of systems increasingly intertwined with our daily lives. Contemporary deep learning models are immense statistical engines that researchers and engineers have developed by a combination of underlying theory and tinkering of practical implementations. Our current theoretical understanding - however - falls short of fully explaining practical successes, and is particularly ill-equipped to explain Residual Networks (He et al., 2015), a simple design feature introduced in 2015 that led to a paradigm shift in how we constructed and which made modern deep learning possible. Investigating the theoretical success of Residual Networks presents an exciting opportunity to close a gap in our understanding and to develop theoretical and experimental tools that can guide improved model design.

Residual Networks addressed the critical issue of making deep networks trainable. Though arbitrarily wide shallow networks are sufficient to represent arbitrary functions (Cybenko, 1989), since 1986 (Håstad, 1986) we have known that deep networks can be exponentially more effective than wide but shallow networks in some instances. The success of early modern deep learning networks such as 2012's AlexNet (Krizhevsky et al., 2012) supported this theory, as models saw a significant decline in performance if even one layer with 1% of the model parameters was removed. Deep networks, however, were plagued by training difficulties and curiously exhibited declines in performance with depth even when their training was made possible by the introduction of batch normalisation (Ioffe & Szegedy, 2015). Residual Networks made a simple change to models - adding the input of each layer to its output - that made deep networks readily trainable and allowed the theorised performance increases to be observed in practice.

Residual Networks, however, took us until 2015 to discover, despite the theoretical promise of deep networks and the computational difficulty of implementing them being decades old, pointing to a shortfall in our theoretical understanding. A theory should motivate improved model design, justifying simple, effective design choices that improve models. However, our approach to analysing networks meant we only experimented with Residual Networks recently, despite the motivation to do so. This points to a shortfall in our understanding and raises the question of how many other simple, effective design choices we overlook. At the least, the unexpected success of Residual Networks should motivate us to learn from having overlooked them and develop a deep learning perspective that centres residual connections. However, even theoretical approaches to deep learning after 2015 ignored the importance of Residual Networks (Geiger & Kubin, 2012). This presents an exciting opportunity to contribute to theoretical deep learning by rigorously analysing Residual Networks' effectiveness.

Currently, many reasons for Residual Networks' effectiveness are qualitative assessments of improved information management. For example, the qualitative reasons proposed for the success of Residual Networks include preventing information from being washed out and allowing for feature reuse. This qualitative intuition has successfully motivated improved architecture design (Huang et al., 2018), motivating us to reinforce this intuition with testable theories with clear implications.

Information theory is uniquely positioned to back this intuition with quantitative experiments and robust theories. Information theory provides quantitative tools to analyse the mechanics of information. Almost every deep learning model already uses it to derive loss functions to guide the training of models. It has produced valuable empirical results and insights into the operation of deep learning architectures (Geiger & Kubin, 2012), and has many theoretical tools that open a promising realm of further insights. By having the output of each layer be closely related to the layer inputs, Residual Networks offer potentially mathematically tractable methods of analysis that are worth investigating.

This literature review integrates into a research proposal to close the gap of our insufficient understanding of Residual Networks with quantitative information theory methods. First, the review covers the history of deep learning theory, showing the original theories that are still widely used (Section 2.), covers the beginning of modern deep learning, which saw divergences from these theories (Section 3.1.), and then goes over the development of Residual Networks that made deep learning possible (Section 3.2.). Next, we show the details and success of our qualitative understanding (Section 3.3.), motivating us to improve our understanding with quantitative tools. Finally, we present information theory as a promising framework to develop empirical tools (Section 4.1.) and theoretical analyses (Section 4.2.) that may close this gap. We also present how focusing on the relationship between information theory and deep learning may make contributions to information theory, providing real examples to motivate grounded interpretation of deep learning theorems (Section 4.3.). Throughout the review, and especially in the final sections, the promising opportunities of applying information theory tools to the experimental investigation and theoretical analysis of deep learning are highlighted, motivating the importance of our approach with clear evidence from the literature.

2. Classical Neural Network Theory

Deep learning begins with neural networks – animal neural networks, that is. In 1943 McCulloch and Pitts presented the first logical analysis of networked Boolean statements – arranged to mimic what was understood of animal neural networks at the time (McCulloch & Pitts, 1943). These boolean neural networks provided a basis for representing computation abstractly, similar to Turing machines, but with a more direct link between how programs are theorised and implemented. These techniques enabled logic networks to be effectively analysed with various theorems and evolved into the deep learning networks we use today.

These networks implemented arbitrary boolean relationships, and in 1986 Håstad used the discrete models proposed by McCulloch and Pitts to show that deep networks can be exponentially more effective at representing certain operations (J. Håstad & Goldmann, 1990; Johan Håstad, 1986). We can collapse layered boolean logic in a way that eliminates many of the components while keeping expressive power. This work showed that deep networks could be far more effective than shallow but wide ones, laying the ground for deep learning. Despite the motivation for deep networks, they would only become practically trainable with the introduction of Residual Networks in 2015 (He et al., 2015).

Extending these theoretical, Boolean models to a more readily trainable continuous form lead to artificial neural networks operating on the sigmoid activation. Instead of zero or one inputs and outputs, a sigmoid neuron sums its inputs according to trainable weights. It produces an output between 0 and 1, saturated at the extremes to mimic a Boolean activation. Significantly, the universal approximation theorem introduced by Cybenko in 1989 showed that finite combinations of sigmoid activations could approximate continuous multivariate functions in a specific range with only one hidden layer (Cybenko, 1989). Though this theory showed that sigmoid activations are sufficiently powerful to represent arbitrary functions even with just one layer, it requires an immense width that we can not extend to practice. Håstad's theorem regarding the exponential advantages of deep networks would be required for artificial neural networks to be practical.

An information theoretic approach to deep learning took root in the late 90s. Information theory originates in Shannon's analysis of communication (Shannon, 1948), who introduced the thermodynamic term *entropy* to computer science and statistics after noticing an equivalence in many equations. An information theory-motivated signals-based analysis of deep learning networks was introduced with the information bottleneck method (Tishby et al., 2000), which proposed that networks initially compress their inputs into an abstract codeword before interpreting these codewords into a final output. This approach used information theory to quantify the mechanics of information, presenting a theory that could (and was) rigorously tested through experiments.

So far, we have covered a collection of classical deep-learning theory papers. The relationship between these papers has the discrete neural networks of McCulloch and Pitts (McCulloch & Pitts, 1943) developed by Håstad (J. Håstad & Goldmann, 1990; Johan Håstad, 1986) to prove the exponential advantages of deep networks. These general discrete networks were extended to the continuous case by Cybenko's universal approximation theorem (Cybenko, 1989) for a network with sigmoid activations. Furthermore, the information bottleneck method (Tishby et al., 2000) integrated continuous networks using sigmoid activations with the information theory approach of Shannon (Shannon, 1948). So far, ideas have been combined and extended to new situations, but the models which use and criticise these theorems would have to wait for the improved GPUs of the 2010s.

3. Applied Deep Learning

3.1. AlexNet: The Beginning of Modern Deep Learning

The 2010s marked a pivotal moment in deep learning, ignited by the launch of the ImageNet competition. Powerful GPUs became available, allowing theories to translate into practice and catalysing techniques like batch normalisation and Residual Networks. This era, heralded by the release of AlexNet, significantly advanced our understanding of deep learning. By analysing the relationship between papers and characterising their technical contributions, we develop a critical view of the state of our current understanding of deep learning.

AlexNet (Krizhevsky et al., 2012), introduced in 2012, revolutionised the field, reducing the top-5 error on ImageNet image recognition tasks from 26.2% to 15.3%. This leap in performance was made possible by taking advantage of contemporary GPUs and employing ReLU activations instead of sigmoids. ReLU activations prevent extreme value saturation, representing a shift from perceiving neural networks as discrete, probabilistic feature extractors to versatile data manipulators. While this divergence from classical theories was notable, ReLU activations upheld the concept of universal approximation, thereby building on Cybenko's work (Cybenko, 1989).

Having 60 million parameters across 650,000 neurons, AlexNet represented a leap in network complexity. Its design emphasised depth, with five convolutional layers and three fully connected ones. This structure upheld Håstad's theories (J. Håstad & Goldmann, 1990; Johan Håstad, 1986), as removing even a single convolutional layer, comprising just 1% of the model's parameters, led to inferior performance. The authors acknowledged that with faster GPUs and larger datasets, performance could be further improved.

AlexNet marked a departure from traditional theories towards practical solutions that harnessed powerful ReLU activations and cutting-edge processors. Despite its success, the challenge of training deep networks remained. Residual Networks would later address this issue, setting the stage for the next wave of advancements in deep learning.

3.2. Deep Learning Made Possible

By 2014, deep learning had started veering into ever deeper networks. That year's ImageNet competition was won by a "very deep" 16 layer deep network using small 3x3 convolution filters, further showing how deep yet thin networks can have impressive performance (Simonyan & Zisserman, 2015). It achieved a 7.3% top-5 error, a vast improvement over the 15.3% error of AlexNet.

Around this time, the size of networks would quickly explode. First came batch normalisation (Ioffe & Szegedy, 2015). Batch normalisation emerged as a powerful solution to a persistent issue in training deep neural networks—_internal covariance shift_. The changing distribution of inputs and outputs during training had long impeded model performance. Batch normalisation addressed this by adjusting the mean and variance of hidden features, significantly enhancing training efficiency. Batch normalisation achieved the same accuracy as state-of-the-art image classification models in *14 times* fewer steps. When fully trained, batch normalised networks exceeded the accuracy of human raters, getting top-5 error rates as low as 4.8%.

Then came Residual Networks (He et al., 2015). Achieving first place in the 2015 image net classification competition with only 3.57% top-5 test error, Residual Networks effectively solved the decades-long problem of training deep networks. While 2014's winning architecture exploited 16 layers and batch normalisation could get as deep as 30, deep Residual Networks could go as deep as *152 layers*. Residual Networks are so critical that the original paper - as of 21 May 2023 - has over 164 thousand citations.

Networks without residual connections, plain networks, completely alter data with every layer. The reasoning goes that a combination of highly general layers allows for an extremely general overarching algorithm, in line with Cybenko's theory that the generality of models allows for universal approximation (Cybenko, 1989). Generality at the level of every layer, however, raises issues. For example, intermediate layers can delete critical information when altering data, which models can never recover. This effect causes information to become "washed out" while being processed.

Instead of completely altering data at every layer, Residual Networks learn to *add* some value to the data. Plain networks teleport data around its possible values, while Residual Networks gently push it. Layers with residual connections are far more stable and less prone to deleting input information. Before residual connections, the performance of models would *decrease* with depth, while residual connections saw models improve monotonically to one hundred layers and beyond. By noticing this issue and resolving it, the Residual Networks paper showed a clear limitation of earlier work (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015) and presented an elegant solution to overcome it.

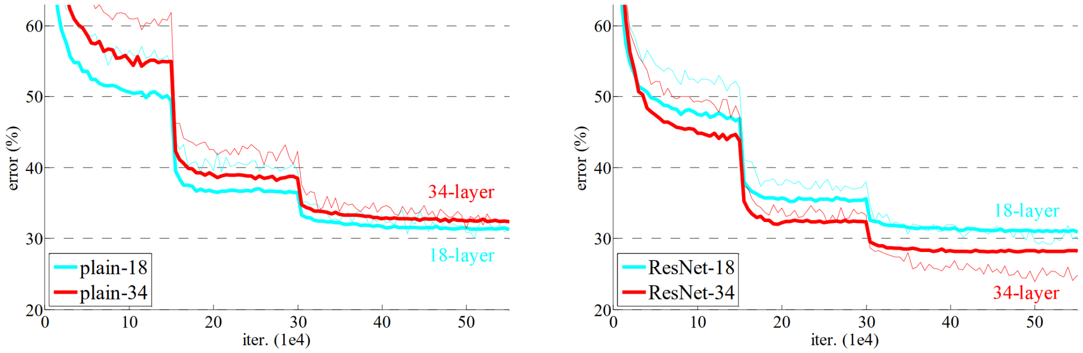


Figure 4. Training on **ImageNet**. Thin curves denote training error, and bold curves denote validation error of the center crops. Left: plain networks of 18 and 34 layers. Right: ResNets of 18 and 34 layers. In this plot, the residual networks have no extra parameter compared to their plain counterparts.

Figure 1: This figure from the Residual Networks paper shows how, unlike plain networks, they exhibited performance increases with depth.

The concept of data "washing out" in networks without residual connections, and the contrasting capability of Residual Networks to preserve data, was a key insight from the original Residual Networks paper. However, this insight was qualitative and possibly ad-hoc, leaving the true mechanics of information preservation and manipulation in Residual Networks uncertain. A lingering question is whether the success of Residual Networks relies on first resolving internal covariance shift and gradient issues with batch normalisation.

Just before the advent of Residual Networks, Highway Networks (Srivastava et al., 2015) independently tackled the issue of training deep networks. While similar in approach to Residual Networks in adding the input of layers to the output, Highway Networks were slightly more complex, as the extent of input preservation was an additional variable that layers had to learn to generate.

However, that work gives valuable insight into the role that residual connections - or in its case, functionally similar highway connections - play. The Highway Networks paper does not mention batch normalisation (Ioffe & Szegedy, 2015)
, and was developed to tackle the same problem that batch normalisation partially addressed: vanishing and exploding gradients with deep networks. It overcame these issues, efficiently training one hundred depth networks *without* batch normalisation. Therefore, it highlights that Residual Networks are not dependent on batch normalisation for their effectiveness and could have been discovered independently. Furthermore, unlike the Residual Networks paper, it mentions the theories of Håstad (J. Håstad & Goldmann, 1990; Johan Håstad, 1986) and others, which justifies the theoretical effectiveness of deep networks.

The relationship of the Highway Networks paper to the rest of the literature shows how batch normalisation is not a prerequisite for the success of Residual Networks. It highlights how old theory could have provided the insight for their development. Therefore, it emphasises the theoretical shortfalls we are trying to address. The Residual Networks paper frames itself as building on the work of batch normalisation, yet it appears to be successful independent of their use. The mechanics of deep networks, therefore, seem only partially understood even by the authors of the Residual Networks paper. It is important that our theoretical framework makes simple and valuable improvements clear. In comparison, our old deep learning theory took us an excruciatingly long time to reveal the simple improvements that have become essential to all modern architectures.

3.3. Successes and Shortfalls of Our Current Intuition

Densely Connected Convolutional Networks (Huang et al., 2018) is a paper that generalises the intuition presented in the Residual Networks paper (He et al., 2015) and Highway Networks paper (Srivastava et al., 2015) to produce another successful architecture for image processing, thereby emphasising the value of grounding our theoretical understanding. Densely Connected Convolutional Networks take feature preservation to the extreme - preserving all prior extracted features for use in every subsequent layer, with the number of input channels to each layer growing linearly. Despite an increased number of inputs, the number of required parameters was lower as resources did not have to be spent by the network relearning essential features that had already been extracted. As a result, they achieved competitive performance on object recognition benchmark tasks while requiring less computation than other models, even in the hyper-competitive deep learning landscape that had emerged by 2018.

The relationship of the Densely Connected Convolutional Networks paper to the rest of the literature provides insight into how our current qualitative assessments of skip connections relate to improved architecture design. It makes explicit mention of the Residual Networks (He et al., 2015) and Highway Networks (Srivastava et al., 2015) papers, stating that it distilled their insights to produce its results. Their design prioritised preventing early information from being washed out and maximising the flow of information and gradients throughout a network. Crucially, their work diverges from those papers in preserving old information by concatenation rather than summation. It criticises Residual Networks for having an excessive number of layers that contribute little and that can even be randomly dropped out, with the concatenation of features fed to highly specialised convolutional layers requiring far fewer overall parameters. Therefore, its comments on the literature highlight both the promise of the insights of residual connections while showing how our understanding is incomplete and alternative methods of preserving early information for later use may also be successful. It motivates our proposed avenue of research, a rigorous investigation of Residual Networks to uncover what truly underpins their success.

At this point, several reasons for the success of residual and skip connections more broadly are apparent from our analysis of the literature and the relationship between papers:

1. They allow for improved information flow, a qualitative assessment that offers practical guidance for improved model design.
2. They greatly stabilise training and regulate the behaviour of gradients, even independent of batch normalisation.
3. They allow for feature reuse, preventing resources such as parameters from being wasted on discovering already extracted features.

These are all qualitative assessments, however, motivating us to think of the valuable insights that may lay in a more in-depth scientific and quantitative understanding of their role.

So far, our literature review has focused on identifying the current literature on residual connections and the successes and shortfalls of our current intuition. The ImageNet papers' clique was assessed and related to our initial presentation of classical deep learning theories. The relationship between papers, their development and their criticism of each other's ideas has identified a rough qualitative understanding of the role of Residual Networks and the need to formalise these ideas with empirical and quantitative methods.

4. Information Theory and Deep Learning

Contemporary machine learning needs to back its qualitative intuition with quantitative tools and scientific experimentation. Information theory, which provides quantitative techniques to analyse information mechanics, is a promising approach. It has already found immense success in many fields; it made effective signal processing possible (Shannon, 1948), is why CDs can be scratched and still played back, and even explains part of why humans age (Sinclair & LaPlante, 2019). Compared to these successful applications, its relevance to deep learning is far more direct. Deep learning architectures are large statistical machines, with the information shared between a model's input, intermediate, output and target data being particularly interesting to us.

In constructing an overall picture of the literature, we will identify various independent papers taking different empirical and theoretical approaches to combining deep learning with information theory. The relationship of these works to our investigation of Residual Networks and deep learning theory will often lay in the lack of relatedness. Hence, we show the opportunity to consolidate various ideas developed by the overall literature, motivating our research proposal that seeks to fill this gap.

4.1. Information Theory's Empirical Tools

A paper worth high consideration is Saxe et al.'s empirical analysis of the information bottleneck theory (Saxe et al., 2019). It uses information theory to test the information bottleneck theory with various experiments that find little evidence for its predictions translating to modern deep learning. The paper leaves us with many tools to conduct information theory-based deep learning analysis and is related to other papers that provide additional tools. However, despite being released at a top deep learning conference in 2018, it does not consider the paradigm shift that Residual Networks offer, meaning its relationship (or rather, lack therefore) to other parts of the literature leaves a gap that motivates our proposal to understand residual connections through the lens of information theory.

That paper starts by noting the gap identified by this literature review: "the practical successes of deep neural networks have not been matched by theoretical progress that satisfyingly explains their behavior." Already, we are in good company. It addresses the information bottleneck theory. The hypothesis is that during training networks first compress their inputs into abstract representations before extracting the critical features required by the task (Tishby et al., 2000). Admirably, it is a scientific hypothesis that can be quantitatively assessed. It has various components, which Saxe et al. found did not extend to the general case.

So far, this review has emphasised the role of Residual Networks and the particular emphasis they deserve in a contemporary theoretical framework for deep learning architectures. The empirical study by Saxe et al. into the information bottleneck theory showed that the prediction of a compression followed by an extraction phase resulted from saturating non-linearities, meaning it failed to translate to the ReLU activations that researchers had started using instead. We have reason to believe that Residual Networks would not experience this compression and extraction behaviour, even if saturating non-linearities were employed. When adding the input back to the output, the possible values multiple layers working together can achieve are unrestricted, not experiencing saturation. The study did not investigate this avenue, thus overlooking the role of Residual Networks, further indicating that opportunities lay in extending information theory tools to Residual Networks.

Their empirical approach develops and collects numerous tools for studying the mechanics of information throughout a network. These have promising applications for studying residual connections and link our investigation to the information theory literature. Information theory is the most elegant when the true distributions are known, a luxury rarely available in practice. Nonetheless, they successfully use various approximations to study mutual information at various phases in a model. These techniques come from thermodynamics (Kraskov et al., 2004), joining diffusion models in enhancing deep learning with physical insights.

They extended their study to *quantitatively* show how irrelevant information added to the input is compressed throughout training while relevant information is largely maintained. This is a perfect perspective from which to quantify the data degradation problem - we can run experiments to compare the mutual information throughout residual and plain networks and see what empirical evidence we find for the hypothesis that Residual Networks are more capable at maintaining critical information throughout training, especially as we increase depth.

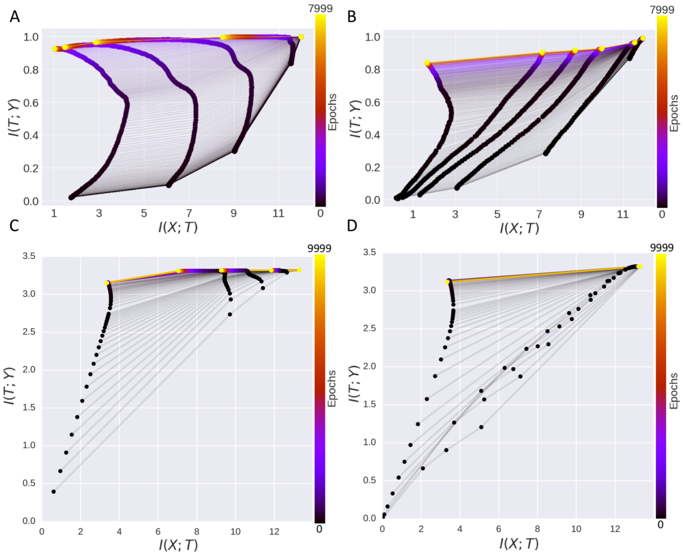


Figure 2: This information plane diagram produced by Saxe et al. shows the relationship between the input data X and intermediate data T along the x-axis. As data is processed, information from the input is reduced. This leads to the points veering left. The y-axis shows the shared information between the target data Y and the intermediate data T. After training, it stays constant, as the model avoids deleting relevant information to their task, but some information is invariably lost.

Throughout training, different algorithms exhibit different patterns. For example, the top two diagrams have saturating activation functions, which first have an up-and-right compression phase that attempts to maintain information relevant to the input before generalising to the target. Saxe et al. showed that this pattern was an artifact of saturating activation functions. In the bottom two diagrams, ReLU activations were used, which did not exhibit this pattern.

These diagrams are part of the insights a quantitative information theory-based approach to Residual Networks may produce, and it would be valuable to see how they differ from plain networks. For example, does their training start at a higher curve? Are their training curves higher up for an equal number of training steps?

A gap in their empirical approach is the focus on shallow networks with less than ten layers. In contrast, Residual Networks shine with very deep networks. Furthermore, they exhibit limiting behaviour as depth increases, which would be fascinating to explore. The deeper a network becomes, the adjustment each layer imposes may become ever smaller as a series of fine adjustments become preferable to course changes. We could compare the information curves that ever deeper networks form. It is easy to hypothesise that deeper networks will provide curves located above the more shallow networks in the information plane diagrams, being able to maintain more information related to the final output. Observations such as these would provide empirical evidence for the hypothesis that depth aids Residual Networks because they can make ever-finer adjustments to data. Simultaneously collecting data on plain networks would enhance the insights provided by such experiments, providing further information about residual connections' role.

4.2. Information Theory and Analysing Deep Residual Networks

In addition to providing a wealth of empirical tools, information theory provides many theoretical tools to enhance our understanding of deep learning. Already, information theory is necessary to derive the loss functions that guide the training of models. Previously, information theory-based approaches have been limited by the mathematical intractability of deep learning architectures. As a result, restrictions are required to make the mathematics work, which separates the proposed models from those used in practice. However, because Residual Networks' input and outputs are closely related, they present a potentially mathematically tractable framework to investigate information flows. Our approach can address both shortfalls in understanding deep learning networks and gaps in how information theory concepts relate to real-world cases.

This section aims to show the promise of consolidating various approaches by showing how the technical characteristics of Residual Networks may overcome their limitations. However, given the breadth of the information theory literature and its lack of consolidation for deep learning purposes, the relationships between these papers are looser, with our motivation stemming from unifying various ideas rather than building on a developed line of research as with the ImageNet papers and their successful application of our existing intuition.

Previously, exact mathematical analyses of deep learning has required models to be heavily restricted in a way that no longer reflects practice. Residual Networks stabilise the mathematics of networks in a manner that opens up exciting avenues of exploration, potentially closing the gap between analytical and practical solutions. An exact analysis of the dynamics of deep networks was proposed by Saxe et al. (Saxe et al., 2014) but requires orthogonalised inputs and is restricted to linear networks, and does not make the link to information theory and its many tools. Similarly, other analytical approaches (Kawaguchi et al., 2019) require orthogonalising weights. A version of these analytical but restricted models - volume-preserving Residual Networks (Gomez et al., 2017) - is of interest to us. It starts by restricting Residual Networks to preserve volume, allowing training without storing prior weights. This method reduces the prohibitive memory loads of large models, a result with clear application to training large models. The success of their approach shows that Residual Networks offer a foundation for tractable analysis, and under certain conditions, we can extend their method to allow for the volume-changing maps we see in practice.

Residual Networks may present mathematical analysis that is exact, practical, and closely related to information theory. Even if volume changes are left unregulated, they may remain manageable and allow tractable information theory quantities to be found (Geiger & Kubin, 2012). We can find the conditional entropy and mutual information from the volume changes induced at each network layer. These volume changes can be derived from the gradients already derived during backpropagation, making these theorems easily applicable without storing additional values. Furthermore, suppose we are investigating the limiting case where the residual modifications become ever smaller. In that case, the average logarithm of the determinant of each layer's gradient - the logarithmic change in volume it induces - approaches the trace of the residual component's gradient. This elegant mathematical simplification warrants further exploration. The previous literature on exactly analysing the information flow of models may be made mathematically tractable by using Residual Networks, motivating our research approach.

A recent work by Peer et al. has examined the relationship between information theory and deep learning with a focus on data degradation (Peer et al., 2022). They derive a regulariser that motivates model stability during training derived from information theory. It allows exceptionally deep models to be trained - 500 layers - *without* residual connections. However, they analyse networks at the level of individual neurons, losing information regarding the correlations between neuron activations. Two tools of interest to us, mutual information and the elegant relationship between volume changes and the change in entropy, still need to be investigated. These methods allow us to examine the relatedness of information while considering the cross-correlation of data within layers. Furthermore, the volume change has a limit in the case of deep Residual Networks that is worth investigating. Nonetheless, that study encourages us that information theory is a worthwhile perspective from which to investigate data degradation.

4.3. Using Deep Learning to Understand Information Theory

Furthermore, an information theory-based approach to deep learning lets us address the gaps in our understanding of information theory. Information theory is much more well-behaved for discrete distributions, with the continuous distributions we observe in practice having promising but unfinished theoretical development. The generalisation of information theory to Markov categories (Perrone, 2022) already makes strides in generalising entropy, the data processing inequality and other concepts in information theory but leave open how metric-dependent continuous distributions should be analysed. Deep neural networks can be viewed as Markov categories (Shiebler et al., 2021), meaning our approach would provide insights into these open questions.

Continuous entropy is considered to be a crude generalisation of discrete entropy, as the continuous limit is technically infinite (Marsh, 2013). Nonetheless, continuous entropy can be meaningfully evaluated in some instances, including in well-behaved Residual Networks (Geiger & Kubin, 2012). If our approach successfully evaluates continuous entropy in a real-world case, that would aid in providing an interpretation for continuous entropy, helping close a gap in the information theory literature.

Unlike previous approaches (Peer et al., 2022; Xu & Raginsky, 2017), we could use the exact volume changes induced by every layer to minimise the number of rough approximations of entropy that we make. Entropy is closely related to other properties of distributions (Gibbs & Su, 2002), and much previous work uses the logarithm of the standard deviation as an upper bound. This has had success in deriving loss functions (Ho et al., 2020) and in previous theoretical analyses of deep networks (Peer et al., 2022; Xu & Raginsky, 2017). However, this upper bound "forgets" the difference between our distribution and an associated uncorrelated normal distribution, losing much potential insight in the process.

5. Conclusion

This literature review has analysed present work to identify notable gaps in our understanding of deep learning, notably the underappreciated relevance of Residual Networks to our theoretical understanding and the exciting opportunities that an information theory-based approach offers. Residual Networks present a paradigm shift to practical model design and how we should theoretically conceptualise deep learning architectures. The need to meet our practical success with theoretical understanding motivates a quantitative, empirical, and scientific approach that information theory has particular promise in addressing. It provides us with clear empirical tools to develop and evaluate different hypotheses and ties deep learning in the era of Residual Networks to the formal mathematics of information. Furthermore, our approach allows information theory to be applied to real-world cases with limited approximations, allowing us to simultaneously assist in closing gaps in the information theory literature.

6. Bibliography

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.

Geiger, B. C., & Kubin, G. (2012). *On the Information Loss in Memoryless Systems: The Multivariate Case*.

Gibbs, A. L., & Su, F. E. (2002). On Choosing and Bounding Probability Metrics. *International Statistical Review*, *70*(3), 419–435.

Gomez, A. N., Ren, M., Urtasun, R., & Grosse, R. B. (2017). The Reversible Residual Network: Backpropagation Without Storing Activations. *CoRR*, *abs/1707.04585*.

Håstad, J., & Goldmann, M. (1990). On the power of small-depth threshold circuits. *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, 610–618 vol.2.

Håstad, Johan. (1986). *Computational limitations for small-depth circuits*. MIT Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. arXiv.

Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. arXiv.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). *Densely Connected Convolutional Networks*. arXiv.

Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. arXiv.

Kawaguchi, K., Bengio, Y., Verma, V., & Kaelbling, L. P. (2019). *Generalization in Machine Learning via Analytical Learning Theory*. arXiv.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, *69*(6), 066138.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*.

Marsh, C. (2013). *Introduction to Continuous Entropy*.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Peer, D., Keulen, B., Stabinger, S., Piater, J., & Rodríguez-Sánchez, A. (2022). *Improving the Trainability of Deep Neural Networks through Layerwise Batch-Entropy Regularization*. arXiv.

Perrone, P. (2022). *Markov Categories and Entropy*. arXiv.

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., & Cox, D. D. (2019). On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, *2019*(12), 124020.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2014). *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*. arXiv.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shiebler, D., Gavranović, B., & Wilson, P. (2021). *Category Theory in Machine Learning*. arXiv.

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv.

Sinclair, D. A., & LaPlante, M. D. (2019). *Lifespan: the revolutionary science of why we age–and why we don't have to* (First Atria Books hardcover edition). Atria Books.

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). *Highway Networks*. arXiv.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). *The information bottleneck method*. arXiv.

Xu, A., & Raginsky, M. (2017). *Information-theoretic analysis of generalization capability of learning algorithms*. arXiv.