Emergent Structures and Training Dynamics in Large Language Models

Ryan Teehan*[1,5], Miruna Clinciu*[2,3,4,5], Oleg Serikov*[5,6,7,8], Eliza Szczechla*[5], Natasha Seelam[5,9], Shachar Mirkin[5,10], Aaron Gokaslan[5,11]

[1]Charles River Analytics [2]Edinburgh Centre for Robotics
[3]Heriot-Watt University [4]University of Edinburgh [5]BigScience [6]AIR Institute
[7]DeepPavlov lab, MIPT [8]HSE University [9]Sherlock Biosciences [10]Lawgeex [11]Cornell University
Contact: rsteehan@gmail.com

Abstract

Large language models have achieved success on a number of downstream tasks, particularly in a few and zero-shot manner. As a consequence, researchers have been investigating both the kind of information these networks learn and how such information can be encoded in the parameters of the model. We survey the literature on changes in the network during training, drawing from work outside of NLP when necessary, and on learned representations of linguistic features in large language models. We note in particular the lack of sufficient research on the emergence of functional units – subsections of the network where related functions are grouped or organized – within large language models, and motivate future work that grounds the study of language models in an analysis of their changing internal structure during training time.

1 Introduction

Recent advances in self-supervised learning, distributed training, and architecture improvements have enabled training massive language models (Devlin et al., 2019; Brown et al., 2020; Radford et al., 2019; Ma et al., 2020; Liu et al., 2019). As these models have grown larger, so has their performance and generalization to new tasks. Furthermore, these techniques have also shown substantial improvements in learning multilingual (Chen et al., 2020) and multimodal representations (Radford et al., 2021). These large language models (LLMs) have advanced the state of the art in few- and zero-shot tasks (Radford et al., 2019; Brown et al., 2020; Radford et al., 2021). However, the size of these models makes them difficult to evaluate, examine, and audit. What structures emerge from training these neural networks? What internal representations do these networks learn?

In part, this opacity is implicit in the models themselves. Many of the fascinating capabilities of LLMs are "implicitly induced, not explicitly constructed" emergent properties (Bommasani et al., 2021). Emergent properties are those that result from the structural relations and interactions between a system's components (Ablowitz, 1939; Callebaut and Rasskin-Gutman, 2005). One way of characterizing the emergence of useful properties from complexity is through self-organization, wherein complex systems come to develop ordered patterns from the interactions of their components (Gershenson et al., 2020). Interactions between the parts of a system can produce complex global behavior, for example in the collective behavior of ants, flocking in birds (Cucker and Smale, 2007), or in the brain and central nervous system (Dresp-Langley, 2020; Brown, 2013). In the context of deep learning models, qualitatively different behavior has been observed during phase transitions in model size or training steps (Steinhardt, 2022). Current research on understanding the generalization abilities of LLMs has largely focused on the degree to which they learn various linguistic features (e.g. syntax) that would support performance on diverse downstream tasks. Our goal instead is to motivate research that grounds the learning of these higher-level representations, and from there, LLMs generalization abilities, in the emergent structures that result from self-organization within the networks.

To analyze LLMs themselves, we survey current research on the following topics and identify gaps in the literature. First, we turn to the development of internal representations of important features of language (e.g. syntax). Second, we look at the structure of the network (neurons, weights, etc.), how it evolves over time, and the emergence of functional units therein. In each case, we include not only research related to trained models, but also the changes that result over training time (termed *training dynamics*). Most research has focused on the aforementioned internal representations and

Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, pages 146 - 159
May 27, 2022 ©2022 Association for Computational Linguistics

their connection to the downstream performance and generalization ability of LLMs, with only limited work on how the network structure changes over time and that change's connection to those representations. We aim to motivate research that not only applies work on the emergent structures within networks from outside of NLP to LLMs, but also develops a language-specific account of useful functional units that emerge in LLMs. Moreover, we identify methods for studying emergence and self-organization in complex systems with potential applications to analyzing LLM training dynamics and behavior. We conclude with a survey of explainability methods that allow researchers to connect structure with function.

2 Internal Representations

Linguistic Structure Representations A significant current area of research is dedicated to interpreting language models from a linguistic point of view. The motivation is to know to what extent models "understand" language, and more specifically, to what extent their generalizations over language agree with the generalizations about language described by linguistics. Following the hierarchy of language levels (morphology, syntax, discourse) (Dalrymple, 2001), experiments in probing studies typically address models' proficiency on a certain level of language. This line of research typically comes down to analyzing how *linguistic structures* are represented in a model's knowledge. Such structures represent syntagmatic/paradigmatic mechanisms of language (how language units combine and alternate, respectively). It is believed (McCoy et al., 2020) that rediscovering these structures would help models get closer to humans performance on a variety of tasks.

Probing Methods to Test for Linguistic Structure Probing tasks measure the linguistic awareness of a model's components, such as layers (Tenney et al., 2019) or groups of neurons (Durrani et al., 2020), by training an auxiliary model, the *probe*, on annotated data. Datasets providing such linguistically annotated data are called *probing datasets*, and cover a wide variety of properties (parts of speech, parse trees, etc.). A high performance of a probe model on a linguistic task implies that the representation tested encodes the property of interest. Several studies using probing methods have reported high accuracy predictions in identifying the underlying linguistic structure (Belinkov et al., 2017a,b; Peters et al., 2018; Tenney et al., 2019; Conneau et al., 2018; Zhang and Bowman, 2018; Alain and Bengio, 2017; Hewitt and Manning, 2019; Hewitt and Liang, 2019).

However, high performance may have confounding factors; there is uncertainty on whether the probing tasks properly test if representations actually encode linguistic structure and on how to interpret the results of probes (Hewitt and Liang, 2019; Zhang and Bowman, 2018; Voita and Titov, 2020; Pimentel et al., 2020b). Toward that end, the following section reviews several probing approaches in the context of language models, and the evaluation criteria used to determine the proficiency of a probe.

Grammatical and Semantic Probing Given the excellent performance of pre-trained representations on numerous linguistic tasks (Kitaev and Klein, 2018; He et al., 2018; Strubell et al., 2018; Lee et al., 2018), several studies have explored how semantic and grammatical knowledge are encoded within language models. *Syntactic* and morphological probing encompasses tasks that identify grammatical structure underlying the vector representations within pre-trained models, whereas *semantic* probing tasks investigate what meaning is conveyed within the representation.

Earlier work using part of speech (POS) and morphological tagging (Belinkov et al., 2017a) indicated that syntactic information may be encoded in layers of neural models. More recently, investigations have considered whether models learn to embed entire parse trees in their representations. In Hewitt and Manning (2019), the authors outline *structural probing* as a method to identify hierarchical, tree-like, structures from vector representations of language via the *syntactic distance* between embeddings. Their results across several large language models suggested that Transformer model encodings possess some hierarchical linguistic structure.

Several studies conducted probing experiments in multilingual settings. Chi et al. (2020) highlighted syntactic generalizations in multilingual language models via structured probing, and Şahin et al. (2020) propose a framework for multilingual morpho-syntactic probing, with 15 probing tasks for multiple languages, showing that, while cross-lingual typological regularities can be found with probing, probing dataset properties strongly impact the results (see Section 2.2 for more details about multilingual models).

Probes have also been used to measure semantic information within language model representations. The authors of Belinkov et al. (2017b) posed a semantic-class labeling task and found that higher layers of a model tend to perform better at semantic tagging. Similarly, semantic labeling tasks have been used to indicate that contextualized representations may encode multiple meanings within a single vector (Yaghoobzadeh et al., 2019). Contrarily, *edge probing*, developed by Tenney et al. (2019), implied that contextualized embeddings show larger gains on syntactic tasks as opposed to semantic ones (with only modest performance gains against non-contextualized baselines). There is no general evidence on how exactly language levels are distributed across model layers (Rogers et al., 2020).

Information Theoretic Probing Information-theoretic probing characterize tasks as a way of estimating the mutual information between an internal representation and the linguistic property of interest (Pimentel et al., 2020b; Pimentel and Cotterell, 2021; Voita and Titov, 2020; Pimentel et al., 2020a). Many of these approaches highlight the need to formalize the "effort" required in encoding a linguistic property, often via some form of a control function (Pimentel et al., 2020b). Counter-intuitively, work from Pimentel et al. (2020b) suggest that the "best" probes are ones that *always* perform highest on the task; their argument is that "learning" the task is equivalent to encoding the linguistic property in the initial representations. They provide approximations to calculate *information gain*, finding that BERT models contain only 12% more information than non-contextualized baselines.

Criticisms of accuracy-based performance metrics have argued that these methods are sensitive to structure, randomization, and hyperparameter selection (Voita and Titov, 2020; Hewitt and Liang, 2019; Zhang and Bowman, 2018; Pimentel et al., 2020b). As an alternative, the minimum description length (MDL) offers an information theoretic view on probe quality (Voita and Titov, 2020). Formally, it describes the "minimum number of bits required to transmit labels, knowing the representations", where better probes are those with smaller codelengths, as they suggest the information available in the representation is sufficiently accessible to solve the task. Prior studies have shown the MDL metric is robust and resilient to randomness (Voita and Titov, 2020). In comparison to the original POS tagging of Hewitt and Liang (2019), the MDL metric consistently distinguishes between the linguistic versus the control tasks across differences in hyperparameters and random seeds. Similarly, following Zhang and Bowman (2018), evaluation using MDL revealed longer codelengths for randomly initialized models as opposed to pre-trained ones.

2.1 Evaluating Probing Performance

Several studies have highlighted the need for interpretable performance scores on probes (Belinkov et al., 2017b; Peters et al., 2018; Tenney et al., 2019; Conneau et al., 2018; Zhang and Bowman, 2018; Alain and Bengio, 2017; Hall Maudslay and Cotterell, 2021). Two common themes have emerged for evaluating the proficiency of a probe: selectivity through *control tasks* and high informatic overlap via *control functions* (Hewitt and Liang, 2019; Pimentel et al., 2020b; Zhu and Rudzicz, 2020). Recent work suggests that both approaches yield comparable results empirically with similar error terms theoretically (Zhu and Rudzicz, 2020).

Control Tasks Selectivity is the trade-off between complexity and performance of the linguistic task. A "good" probe refers to one that performs highly on linguistic tasks, but poorly on control tasks, thus limiting the ability for a probe to "memorize" the task (Hewitt and Liang, 2019).

Arguments preferring "simpler" probes claim that these models should find "accessible" information within the representations (Shi et al., 2016). The simplest probes employ linear functions, yet more complex probes have been commonly used, including multi-layer perceptrons (MLP) or kernel methods (Belinkov et al., 2017a; Conneau et al., 2018; White et al., 2021; Adi et al., 2017), suggesting that some linguistic properties may be encoded non-linearly. Linear functions and MLPs are still commonly in use (Tenney et al., 2019).

Prior works within the probing literature have also explored how the size of training data can influence the performance of the probe (Zhang and Bowman, 2018; Hewitt and Liang, 2019). In an investigation considering probes of pre-trained language models and an untrained baseline on two syntactic tasks: POS tagging and Combinatorial Categorical Grammar (CCG) super-tagging (Hockenmaier and Steedman, 2007), probes with an untrained baseline model could surprisingly attain high performance compared to pre-trained models (Zhang and Bowman, 2018). However, the probe performance decreased dramatically when reducing the amount of available training data when compared to the pre-trained models. This suggested trained encoders captured enough syntactic information, beyond simple word-identities, which enabled these representations to achieve high performance on the linguistic tasks.

An extensive study on selectivity proposed several control tasks for POS tagging and dependency edge prediction (Hewitt and Liang, 2019). Across an array of probe architectures (linear, MLP-1, MLP-2) and hyperparameters, this investigation considered the effect of the hidden state dimensionality (size), number of training examples, regularization, and early stopping. The most effective probes were those with constrained hidden dimensions, yielding the most selective probes.

Control Functions Control functions compare the mutual information against a property of interest and the representation before and after the function is applied. The objective is used to measure the information gain of the representation. In Pimentel et al. (2020b), control functions were used to compare BERT contextualized models against FastText (Bojanowski et al., 2017) and a one-hot encoding on POS tagging. Curiously, their results suggested that BERT models only marginally improved information gain against these simpler baselines.

2.2 Emerging Multilingual Structures

Multilingual large language models, such as multilingual BERT (mBERT) (Devlin et al., 2019; Devlin, 2018) XLM (Conneau and Lample, 2019) or XLM-R (Conneau et al., 2020a) have shown impressive results when used for (zero-shot) cross-lingual transfer; that is, when the pre-trained multilingual language model is used as the basis for a task-specific model that is applied to a language in which it was not trained for. Their efficiency was proven in a wide variety of tasks, such as sentiment analysis, natural language inference, and question answering, to name a few.

Prior to the immense popularity of Transformer-based models, two approaches of using word embeddings for cross-lingual tasks have shown promising results. In the first, representations are learned separately from individual languages and then aligned to a shared space, thus producing *cross-lingual word embeddings* (Ruder et al., 2019), that in turn, are used on the target language. In the second, *multilingual* representations are learned by jointly training over multiple languages. Artetxe and Schwenk (2019), for example, trained a BiLSTM over 93 languages using parallel corpora, producing "universal" embeddings that were successfully used in various tasks.

The same two approaches are being explored with large language models. In Conneau et al. (2020b), monolingual BERT models that were trained separately for different languages produced similar (easily-aligned) representations. Pires et al. (2019) and Vulić et al. (2020) further showed – as expected – that the similarity depends on the typological distance between the languages. Universal language-agnostic embeddings also emerge when training multilingual models, even when no explicit connection (such as parallel corpora or bilingual dictionaries) between the languages is used during training, such as in the case of mBERT.

Multiple works looked into the factors that contribute to the successful transfer. These include domain and language similarity, shared parameters, and perhaps the most straightforward factor: common (sub-) words between the languages (Wu and Dredze, 2019; Conneau et al., 2020b; Pires et al., 2019). Interestingly, Conneau et al. (2020b) and K et al. (2020) showed that the universal representations do not heavily depend on shared vocabulary; instead, multilinguality emerges directly from the fact that parameters are shared in training, from the structure of the network, and is affected by common characteristics of the languages, such as word order (Dufter and Schütze, 2020). Pires et al. (2019) discovered that mBERT can also successfully transfer between languages with different scripts, and that generalization goes beyond the lexical level, and Chi et al. (2020) found that syntactic features representations in mBERT overlap between languages. Still, Ahmad et al. (2021) have shown that augmenting mBERT with syntactic information can improve cross-lingual transfer performance.

The size of each language's corpus in the language model's training set has been shown to be decisive for transfer to that language. Thus, low-resource languages often benefit more from the joint training (Wu and Dredze, 2020), while languages with abundant resources often achieve better performance when trained on their own (Nozza et al., 2020; Lewis et al., 2020).

2.3 Training Dynamics of Internal Representation Development

Training dynamics is an emerging field of research, promising to improve our understanding of knowledge acquisition in neural networks and offering insights into the utility of pre-trained models and embedded representations for downstream tasks. Most studies of Transformers (e.g. RoBERTa (Zhuang et al., 2021)) and LSTMs (Hochreiter and Schmidhuber, 1997) agree that models acquire linguistic knowledge early in the learning process.

Local syntactic information, such as parts of speech, is learned earlier than information encoding long-distance dependencies (e.g. topic) (Liu et al., 2021; Saphra, 2021). Exploration of ALBERT (Lan et al., 2019) and LSTM-based networks reveals different learning patterns for function and content words with more fine-grained distinctions within these categories including part of speech and verb form (Saphra, 2021; Chiang et al., 2020).

Differences in learning trajectory were also observed between layers. In LSTMs, recurrent layers become more task-independent over the course of training, while embeddings become more task-specific (Saphra, 2021). In Transformer-based architectures, i.e.: ALBERT and ELECTRA, Chiang et al. (2020) observe differences in performance patterns between the top and last layers. Similarly to other areas of research in NLP, most of the literature on training dynamics concentrate on English-language models. Another possible direction for future work is extending studies conducted on LSTMs to more widely used Transformers.

2.4 Critique of Testing Methods

Recent research has complicated the picture of grammar learning presented in Sections 2, 2.2, and 2.3. Specifically, there have been two separate but related types of critique leveled at probing and grammar learning. First, specific to probing, researchers question whether probes really identify linguistic representations at all. Secondly, and more fundamentally, it is unclear to what degree language models even learn grammar.

Hall Maudslay and Cotterell (2021) suggest that semantic "cues" may contaminate syntax probes, making it difficult to evaluate their scores. By employing "Jabberwocky probing", where pseudowords with no lexical meaning replace the original components of the sentence in a way that preserves grammar, the authors discovered that performance of syntactic probes considerably dropped for large language models, calling into question whether syntactic probes actually isolate syntactic knowledge withing language models.

A more fundamental issue for syntax learning in language models has been their performance when trained on perturbed or permuted data. Sinha et al. (2021) use a variety of word order permutations that preserve distributional information to isolate whether what language models learn is actually syntax. Word order has been assumed to be important not only for natural language understanding by humans but also by language models, particularly for learning syntax. Surprisingly then, word order appears to have less influence than one would expect on the downstream performance of language models and their performance on probing tasks. In part, the authors note that some syntax information can be acquired during fine-tuning to sufficiently answer tasks that require it. Moreover, in the context of syntax probes, the authors note that "while natural word order is useful for at least some probing tasks, the distributional prior of randomized models alone is enough to achieve a reasonably high accuracy on syntax sensitive probing". Furthermore, the results distinguish between parametric and non-parametric probes, where performance on the latter using randomization models degrades significantly. This degradation provides evidence that non-parametric probes are able to test for syntax learning in ways that parametric probes cannot. Similarly, O'Connor and Andreas (2021) use syntax-level perturbations and ablations to conclude that the information in context windows most useful to language models are local ordering statistics and content words, e.g. nouns, verbs, adverbs, and adjectives. In other words, it does not appear that language models make use of syntactic or other structural information in the context window.

2.5 Further Research

Despite recent probing studies providing a closer look at how linguistic structures are distributed in language models, it is an open question to what extent this knowledge acquisition differs from that of humans. While grammatical structures tend to be learned much faster than downstream knowledge (Conneau et al., 2018), there is still room for the study of more specific questions, such as whether models require more time to acquire the grammar of polysynthetic languages, as has been reported for humans (Kelly et al., 2014).

Another remaining open question is whether linguistic structure knowledge can be transferred between models with the neurons initialization mechanism (Durrani et al., 2021). While rough re-use of neurons is proven to be helpful in model initialization (Sanh et al., 2019), for instance, such neuron "surgery" would potentially lead to even quicker acquisition of grammatical knowledge.

Generally speaking, the performance of multilingual models is inferior to that of monolingual ones, especially when enough resources are available. Yet, high-quality multilingual models remain a desired objective that can particularly benefit low-resource languages. Further understanding the factors that enable learning language-independent representations is key for developing better multilingual training or cross-lingual fine-tuning strategies, especially for transfer between less similar language pairs. A particularly interesting question is whether some tasks require more language-specific adaptation, because, for instance, they depend on linguistic information that is currently not generalized well enough in multilingual LLMs.

3 Self-Organization and the Emergent Structure of Networks

3.1 Network Structure

Inspired by the architecture of biological neural networks (BNNs) and their adaptability to various tasks, where neurons and circuits are capable of self-organization, many researchers have investigated how Artificial Neural Networks (ANNs) can be seen as emergent structures, where interpretability of an ANN's parameters can help us to inspect their functional modularity. Broadly, researchers have approached this by identifying patterns in the weights or neurons especially through subgraphs of the network.

Branch specialization is the organization of branches – or "sequences of layers which temporarily don't have access to 'parallel' information which is still passed to later layers" (Voss et al., 2021) – of the network into functional units, across different architectures and tasks (Zhang et al., 2020; Bunel et al., 2020; Voss et al., 2021; Rössig and Petkovic, 2021). It is somewhat similar to how neurons are connected by synapses, forming small functional units called neural circuits that can be specialized for specific tasks, such as to "mediate reflexes, process sensory information, generate locomotion and mediate learning and memory" (Byrne et al., 2012; Luo, 2021). In their work on AlexNet, Voss et al. (2021) provided initial evidence of self-organization of neurons and circuits (subgraphs) into functional units in a neural network. This self-organized emergent structure is consistent "across different architectures and tasks". A look at evolving neural structures gives another perspective. Inspired by neural architecture search (NAS), So et al. (2019) presented "a first neural architecture search conducted to find improved feed-forward sequence models", where the search space contains five branch-level search fields. Recently, So et al. (2021) introduced Primer (PRIMitives searched TransformER), which can add improvements in the pre-training and one-shot downstream task transfer regime. However, branches are used just for the initialized multi-head attention.

Weight banding is the uniformity in the organization of the weights in a final layer. In neural networks, weights are parameters that can transform the input data between the network's hidden layers. Weight banding resembles another biological phenomenon when a neuron multiplies each input with a synaptic weight, which is represented as a number that highlights the importance assigned to that input. The weighted inputs are summed up in what represents the neuron's output (Iyer et al., 2013). Petrov et al. (2021) note that many vision models display a uniform pattern in their final layer. They investigate the nature of this structural phenomenon, connecting it ultimately to architectural choices in the network and noting that weight banding can serve as a method of preserving spatial information.

Clustering is the grouping of neurons or subnetworks into units that can be used for specific tasks (Hod et al., 2021). Starting from the fact that modular systems allow us to have a better understanding of a system if we can inspect the function of individual modules, different clustering methods for neural networks were proposed. Li et al. (2020) designed a modular neural network based on feature clustering to decompose features into clusters with each module processing different features. These modules work in parallel for a singular task. Filan et al. (2021) proposed a spectral clustering algorithm for decomposition of trained networks into clusters, finding that networks can have some sense of modularity and suggested further work related to clusterability in various domains.

Modularity focuses on the reusability of subnetworks for multiple tasks (Happel and Murre, 1994; Shukla et al., 2010; Csordás et al., 2021). In Csordás et al. (2021) neural networks trained on algorithmic tasks appear to fail to learn general, modular, compositional algorithms, and require specific subset weights to handle a particular combination of the input tokens. With these findings, Csordás et al. (2021) suggest further research about "function dependent weight sharing in the neural networks". Reusable multi-task subnetworks may also be discovered via Neural Architecture Search (NAS) methods (Pham et al., 2018). Pasunuru and Bansal (2019) leverage a technique called multi-task architecture search (MAS) to find multi-task cell structures in RNNs, capable of generalization to unseen tasks.

3.2 Training Dynamics of Network Changes

Understanding the change in network structure over time is equally as important as identifying structure in trained models. Here, the focus is on how the parameters of the model change over the course of training, which can give insight into the types of inductive biases that develop and shed light on the nature of LLMs' abilities to generalize. The most recent work covering this in the context of LLMs focuses on parameter norm growth, which refers to the growth of the ℓ_2 norm during training time. According to Merrill et al. (2021), neural networks learn successfully due to inductive biases introduced during training. Norm growth induces saturation in Transformer models, which reduces the attention heads to "generalized hard attention". The authors find that computations for *argmax* and *mean* are reducible to saturated attention, which partially explains why saturated Transformer models can learn counter languages, a kind of formal language, and may play a broader role in explaining their generalization abilities.

3.3 Further Research

As we have noted, most of the work on network structure is currently outside of NLP, either dealing with general ANNs or specific to Computer Vision with AlexNet and general convolutional networks trained on ImageNet (Voss et al., 2021; Petrov et al., 2021). This work should be replicated in the context of LLMs to test for the existence of language-specific functional units and, more generally, determine whether there are internal network structures that support the learned representations we discuss in Section 2. Likewise, since this research is still in its infancy, it is focused on simple emergent structures. Future research can incorporate higher-order emergent structures (Baas, 2000), new methods of structure detection in networks (Aktas et al., 2019), and even detection of structures whose form is not explicitly specified (Shalizi et al., 2006).

Additionally, by viewing the neural networks in question as time-evolving complex systems we can leverage older research on self-organization that has yet to be applied to understanding LLMs. In particular, Ball et al. (2010) provide a method for quantifying self-organization based on persistent mutual information. Likewise, Shalizi et al. (2004) ground self-organization in information theory and Shalizi (2003) extends this method to a general class of undirected graphs. Methods such as these can be used to identify and quantify self-organization in LLMs and better understand their emergent behavior.

4 Connecting Structure to Function: Explainable AI (XAI)

The rapid increase in the adoption of AI models in recent years and their growing impact on human lives created a need for techniques that offer insight into the models internal operations.

Since attention-based models (Vaswani et al., 2017) have become state-of-the-art tools in NLP, there have been numerous attempts to provide some understanding of their predictions by visualizing the attention layer. However, these approaches have been criticized for their inability to produce meaningful and coherent interpretations (Wiegreffe and Pinter, 2019; Bastings and Filippova, 2020; Serrano and Smith, 2019). To address these limitations, Ghaeini et al. (2018) examine the saliency of attention and LSTM gating signal in the intermediate layers of ESIM models, an architecture designed for natural language inference tasks (Chen et al., 2017). Their results show that visualizing attention saliency allows identifying which parts of the premise and hypothesis contribute most to the final score. Moreover, attention saliency maps compared across different ESIM models reveal differences in focus that reflect the differences in their predictions. According to this study, using saliency is much more effective than using attention alone.

Another approach to revealing how decisions are formed across network layers is *erasure*, where features are deemed irrelevant if their removal has a minor effect on the prediction. De Cao et al. (2020) extend this method to learned masking and adapt it to measure the importance of intermediate states rather than the inputs. They run the proposed DIFF-MASK method on BERT (Devlin et al., 2019) and find that separator tokens play an important role in the input layer for question answering but not for sentiment classification, a task where adjectives and nouns are kept for much longer. Given that separators serve as delimiters between the question and the context, these differences shed light on the connection between the internal latent structure and the task, marking a step toward gaining some understanding of the information flow in the model.

Applying neural models to the NLP domain poses specific challenges. This opens the way for research on the extent to which language-specific characteristics, such as compositionality of meaning, are reflected in the internal representations of neural networks. The work by Li et al. (2016) leverages several methods including variance-based and first-derivative saliency (a technique inspired by back-propagation), to study how models deal with compositionality of meaning, e.g., negation, intensification and combining meaning from different parts of the sentence. The study of recurrent, LSTM and bi-LSTM networks across time steps finds that, as decoding proceeds, the task (language modelling) gradually prevails overbuilding word representations.

An integrated gradients (Sundararajan et al., 2017) based method of finding neurons that encode individual facts has been proposed by Dai et al. (2021). This approach builds on the observation that large pre-trained language models can remember factual knowledge from the training corpus. The authors find that knowledge neurons are located in the feed-forward network of BERT and view these two-layer perceptron modules as knowledge memories in the Transformer architecture. The method allows for explicit editing of specific factual knowledge by manipulating the corresponding knowledge neurons with only a moderate influence on unrelated knowledge. These findings are in line with a work by Meng et al. (2022) that localizes factual knowledge to the feed-forward layer. Further, this approach makes a distinction between the notions of *knowing* and *saying* a fact and concludes that, while the feed-forward layers encode the former, the latter is attended to by the late self-attention.

Other approaches, e.g. SHAP, DeepLift and LIME (Lundberg and Lee, 2017; Shrikumar et al., 2019; Ribeiro et al., 2016) can reveal dependencies missed by the methods discussed here. In NLP, the key challenges include performance and, where applicable, choosing an adequate baseline for word embeddings. The dynamic progress of research in natural language processing has led researchers to review and analyze existing methods of interpreting neural models (Belinkov and Glass, 2019; Danilevsky et al., 2020). While the emerging field of explainable AI (XAI) is seeing faster growth, a path for research and discussion on the desired evaluation criteria of interpretation methods is opening up (Jacovi and Goldberg, 2020).

5 Conclusion and Future Directions

In this paper, we provide an overview of research on network structure, linguistic feature learning, their training dynamics, and explainability research that aims to connect network structure and function. In doing so, we highlight gaps in the literature and opportunities for future research, both in each individual research area and as a broad proposal for grounding research in understanding large language models. We highlight a few areas of future research as particularly important given the gaps in the current literature. For the study of how, and whether, linguistic structures are learned by language models, more work is needed to understand the training dynamics of this learning across a variety of model scales and architectures. More fundamentally, there is disagreement about what it means for a model to "encode" linguistic structures such as syntax, particularly in a multilingual setting.

More broadly, nascent work on the self-organization of neurons and subnetwork structures that emerge during training time has largely not been applied to LLMs, or neural networks in NLP more generally. Research in Computer Vision has shown the existence of emergent functional units with functions that are semantically meaningful to humans. In the context of LLMs, such structures may provide a basis for understanding the nature of linguistic features that LLMs purportedly learn, especially when comparing the development of each during training time. Additional research is needed to not only determine whether such structures emerge in LLMs, but also to apply and extend the literature on self-organization in complex systems. This research can also be used for explainability. Currently, assessment of the quality of interpretations of the information flow in neural models is not straightforward. Identification of modular and emergent structures within networks may be viewed as a way of moving away from the binary definition of *faithfulness* as postulated by Jacovi and Goldberg (2020). Evidence for the existence of structures aligning with human perception of language, if found, can help to enable separate consideration of *plausibility* from a human perspective, as proposed in the same study. More broadly, we propose grounding the study of LLMs properties in the analysis of the self-organization of weights and neurons into emergent structures.

References

Reuben Ablowitz. 1939. The theory of emergence. *Philosophy of Science*, 6(1):1–16.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554, Online. Association for Computational Linguistics.

Mehmet Aktas, Esra Akbas, and Ahmed El Fatmaoui. 2019. Persistence homology of networks: methods and applications. *Applied Network Science*, 4.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Nils A. Baas. 2000. Emergence, hierarchies, and hyperstructures. In Christopher G. Langton, editor, *Artificial Life III*, page 515–537. Addison-Wesley Longman Publishing Co., Inc., USA.

Robin C. Ball, Marina Diakonova, and Robert S. MacKay. 2010. Quantifying emergence in terms of persistent mutual information. *Adv. Complex Syst.*, 13:327–338.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Steven Ravett Brown. 2013. Emergence in the central nervous system. *Cognitive Neurodynamics*, 7(3).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Rudy Bunel, Ilker Turkaslan, Philip H.S. Torr, M. Pawan Kumar, Jingyue Lu, and Pushmeet Kohli. 2020. Branch and bound for piecewise linear neural network verification. *Journal of Machine Learning Research*, 21.

John H Byrne, D Ph, and The Ut. 2012. Introduction to Neurons and Neuronal Networks. *Cellular and Molecular Neurobiology*, 1.

Werner Callebaut and Diego Rasskin-Gutman. 2005. *Modularity: Understanding the Development and Evolution of Natural Complex Systems*. The MIT Press.

Peng-Jen Chen, Ann Lee, Changhan Wang, Naman Goyal, Angela Fan, Mary Williamson, and Jiatao Gu. 2020. Facebook AI's WMT20 news translation task submission. In *Proceedings of the Fifth Conference on Machine Translation*, pages 113–125, Online. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained language model embryology: The birth of ALBERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6813–6828, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6022–6034. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. 2021. Are neural nets modular? inspecting functional modularity through differentiable weight masks. In *International Conference on Learning Representations*.

Felipe Cucker and Steve Smale. 2007. Emergent behavior in flocks. *IEEE Transactions on Automatic Control*, 52(5):852–862.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *ArXiv*, abs/2104.08696.

Mary Dalrymple. 2001. *Lexical functional grammar*. Brill.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online. Association for Computational Linguistics.

Jacob Devlin. 2018. Multilingual bert.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Birgitta Dresp-Langley. 2020. Seven properties of self-organization in the human brain. *Big Data and Cognitive Computing*, 4(2).

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep NLP models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4947–4957, Online. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. *arXiv preprint arXiv:2010.02695*.

Daniel Filan, Stephen Casper, Shlomi Hod, Cody Wild, Andrew Critch, and Stuart Russell. 2021. Clusterability in neural networks. *CoRR*, abs/2103.03386.

Carlos Gershenson, Vito Trianni, Justin Werfel, and Hiroki Sayama. 2020. Self-Organization and Artificial Life. *Artificial Life*, 26(3):391–408.

Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957, Brussels, Belgium. Association for Computational Linguistics.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do syntactic probes probe syntax? experiments with jabberwocky probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131, Online. Association for Computational Linguistics.

Bart L.M. Happel and Jacob M.J. Murre. 1994. Design and evolution of modular neural network architectures. *Neural Networks*, 7(6-7).

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364–369, Melbourne, Australia. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Julia Hockenmaier and Mark Steedman. 2007. Ccgbank: A corpus of ccg derivations and dependency structures extracted from the penn treebank. *Comput. Linguist.*, 33(3):355–396.

Shlomi Hod, Stephen Casper, Daniel Filan, Cody Wild, Andrew Critch, and Stuart J. Russell. 2021. Detecting modularity in deep neural networks. *ArXiv*, abs/2110.08058.

Ramakrishnan Iyer, Vilas Menon, Michael Buice, Christof Koch, and Stefan Mihalas. 2013. The Influence of Synaptic Weight Distribution on Neuronal Population Dynamics. *PLoS Computational Biology*, 9(10).

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Barbara Kelly, Gillian Wigglesworth, Rachel Nordlinger, and Joseph Blythe. 2014. The acquisition of polysynthetic languages. *Language and Linguistics Compass*, 8(2):51–64.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Wenjing Li, Meng Li, Junfei Qiao, and Xin Guo. 2020. A feature clustering-based adaptive modular neural network for nonlinear system modeling. *ISA Transactions*, 100:185–197.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Liqun Luo. 2021. Architectures of neuronal circuits. *Science*, 373(6559):eabg7285.

Shuming Ma, Jian Yang, Haoyang Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla, Alexandre Muzio, Akiko Eriguchi, Saksham Singhal, et al. 2020. Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders. *arXiv preprint arXiv:2012.15547*.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual knowledge in gpt.

William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Effects of parameter norm growth during transformer training: Inductive bias from gradient descent. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1766–1781, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 851–864, Online. Association for Computational Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2019. Continual and multi-task architecture search.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Michael Petrov, Chelsea Voss, Ludwig Schubert, Nick Cammarata, Gabriel Goh, and Chris Olah. 2021. Weight banding. *Distill*. Https://distill.pub/2020/circuits/weight-banding.

Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. 2018. Efficient neural architecture search via parameter sharing.

Tiago Pimentel and Ryan Cotterell. 2021. A bayesian framework for information-theoretic probing.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020a. Pareto probing: Trading off accuracy for complexity. *CoRR*, abs/2010.02180.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020b. Information-theoretic probing for linguistic structure. *CoRR*, abs/2004.03061.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Ansgar Rössig and Milena Petkovic. 2021. Advances in verification of ReLU neural networks. *Journal of Global Optimization*, 81(1).

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. Linspector: Multilingual probing tasks for word representations. *Computational Linguistics*, 46(2):335–385.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Naomi Saphra. 2021. *Training dynamics of neural language models*. Ph.D. thesis, The University of Edinburgh.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Cosma Rohilla Shalizi. 2003. Optimal nonlinear prediction of random fields on networks. In *DMCS*.

Cosma Rohilla Shalizi, Robert Haslinger, Jean-Baptiste Rouquier, Kristina Lisa Klinkner, and Cristopher Moore. 2006. Automatic filters for the detection of coherent structure in spatiotemporal systems. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 73 3 Pt 2:036104.

Cosma Rohilla Shalizi, Kristina Lisa Shalizi, and Robert Haslinger. 2004. Quantifying self-organization with optimal predictors. *Physical review letters*, 93 11:118701.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2019. Learning important features through propagating activation differences.

Anupam Shukla, Ritu Tiwari, and Rahul Kala. 2010. *Modular Neural Networks*, pages 307–335. Springer Berlin Heidelberg, Berlin, Heidelberg.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David So, Quoc Le, and Chen Liang. 2019. The evolved transformer. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5877–5886. PMLR.

David R. So, Wojciech Manke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V. Le. 2021. Primer: Searching for efficient transformers for language modeling. *CoRR*, abs/2109.08668.

Jacob Steinhardt. 2022. Future ml systems will be qualitatively different. Https://bounded-regret.ghost.io/future-ml-systems-will-be-qualitatively-different/.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv e-prints*, page arXiv:1905.06316.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Chelsea Voss, Gabriel Goh, Nick Cammarata, Michael Petrov, Ludwig Schubert, and Chris Olah. 2021. Branch specialization. *Distill*. Https://distill.pub/2020/circuits/branch-specialization.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 132–138, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Katharina Kann, T. J. Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5740–5753, Florence, Italy. Association for Computational Linguistics.

Hongyang Zhang, Junru Shao, and Ruslan Salakhutdinov. 2020. Deep neural networks with multi-branch architectures are intrinsically less non-convex. In *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics*.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

Zining Zhu and Frank Rudzicz. 2020. An information theoretic view on selecting linguistic probes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9251–9262, Online. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.