A Primer in BERTology: What We Know About How BERT Works

Anna Rogers
Center for Social Data Science
University of Copenhagen
arogers@sodas.ku.dk

Olga Kovaleva
Dept. of Computer Science
University of
Massachusetts Lowell
okovalev@cs.uml.edu

Anna Rumshisky
Dept. of Computer Science
University of
Massachusetts Lowell
arum@cs.uml.edu

Abstract

Transformer-based models have pushed state of the art in many areas of NLP, but our understanding of what is behind their success is still limited. This paper is the first survey of over 150 studies of the popular BERT model. We review the current state of knowledge about how BERT works, what kind of information it learns and how it is represented, common modifications to its training objectives and architecture, the overparameterization issue, and approaches to compression. We then outline directions for future research.

1 Introduction

Since their introduction in 2017, Transformers (Vaswani et al., 2017) have taken NLP by storm, offering enhanced parallelization and better modeling of long-range dependencies. The best known Transformer-based model is BERT (Devlin et al., 2019); it obtained state-of-the-art results in numerous benchmarks and is still a must-have baseline.

Although it is clear that BERT works remarkably well, it is less clear *why*, which limits further hypothesis-driven improvement of the architecture. Unlike CNNs, the Transformers have little cognitive motivation, and the size of these models limits our ability to experiment with pre-training and perform ablation studies. This explains a large number of studies over the past year that attempted to understand the reasons behind BERT's performance.

In this paper, we provide an overview of what has been learned to date, highlighting the questions that are still unresolved. We first consider the linguistic aspects of it, namely, the current evidence regarding the types of linguistic and world knowledge learned by BERT, as well as where and how this knowledge may be stored in the model. We then turn to the technical aspects of the model and provide an overview of the current proposals to improve BERT's architecture, pre-training, and fine-tuning. We conclude by discussing the issue of overparameterization, the approaches to compressing BERT, and the nascent area of pruning as a model analysis technique.

2 Overview of BERT Architecture

Fundamentally, BERT is a stack of Transformer encoder layers (Vaswani et al., 2017) that consist of multiple self-attention ''heads''. For every input token in a sequence, each head computes key, value, and query vectors, used to create a weighted representation. The outputs of all heads in the same layer are combined and run through a fully connected layer. Each layer is wrapped with a skip connection and followed by layer normalization.

The conventional workflow for BERT consists of two stages: pre-training and fine-tuning. Pre-training uses two self-supervised tasks: masked language modeling (MLM, prediction of randomly masked input tokens) and next sentence prediction (NSP, predicting if two input sentences are adjacent to each other). In fine-tuning for downstream applications, one or more fully connected layers are typically added on top of the final encoder layer.

The input representations are computed as follows: Each word in the input is first tokenized into wordpieces (Wu et al., 2016), and then three embedding layers (token, position, and segment) are combined to obtain a fixed-length vector. Special token `[CLS]` is used for classification predictions, and `[SEP]` separates input segments.

Google[1] and HuggingFace (Wolf et al., 2020) provide many variants of BERT, including the original ''base'' and ''large'' versions. They vary in the number of heads, layers, and hidden state size.

[1] https://github.com/google-research/bert.

Transactions of the Association for Computational Linguistics, vol. 8, pp. 842–866, 2020. https://doi.org/10.1162/tacl_a_00349
Action Editor: Dipanjas Das. Submission batch: 4/2020; Revision batch: 8/2020; Published 12/2020.

3 What Knowledge Does BERT Have?

A number of studies have looked at the knowledge encoded in BERT weights. The popular approaches include fill-in-the-gap probes of MLM, analysis of self-attention weights, and probing classifiers with different BERT representations as inputs.

3.1 Syntactic Knowledge

Lin et al. (2019) showed that **BERT representations are hierarchical rather than linear**, that is, there is something akin to syntactic tree structure in addition to the word order information. Tenney et al. (2019b) and Liu et al. (2019a) also showed that **BERT embeddings encode information about parts of speech, syntactic chunks, and roles**. Enough syntactic information seems to be captured in the token embeddings themselves to recover syntactic trees (Vilares et al., 2020; Kim et al., 2020; Rosa and Mareček, 2019), although probing classifiers could not recover the labels of distant parent nodes in the syntactic tree (Liu et al., 2019a). Warstadt and Bowman (2020) report evidence of hierarchical structure in three out of four probing tasks.

As far as *how* syntax is represented, it seems that **syntactic structure is not directly encoded in self-attention weights**. Htut et al. (2019) were unable to extract full parse trees from BERT heads even with the gold annotations for the root. Jawahar et al. (2019) include a brief illustration of a dependency tree extracted directly from self-attention weights, but provide no quantitative evaluation.

However, **syntactic information can be recovered from BERT token representations**. Hewitt and Manning (2019) were able to learn transformation matrices that successfully recovered syntactic dependencies in PennTreebank data from BERT's token embeddings (see also Manning et al., 2020). Jawahar et al. (2019) experimented with transformations of the [CLS] token using Tensor Product Decomposition Networks (McCoy et al., 2019a), concluding that dependency trees are the best match among five decomposition schemes (although the reported MSE differences are very small). Miaschi and Dell'Orletta (2020) perform a range of syntactic probing experiments with concatenated token representations as input.

Note that all these approaches look for the evidence of gold-standard linguistic structures,

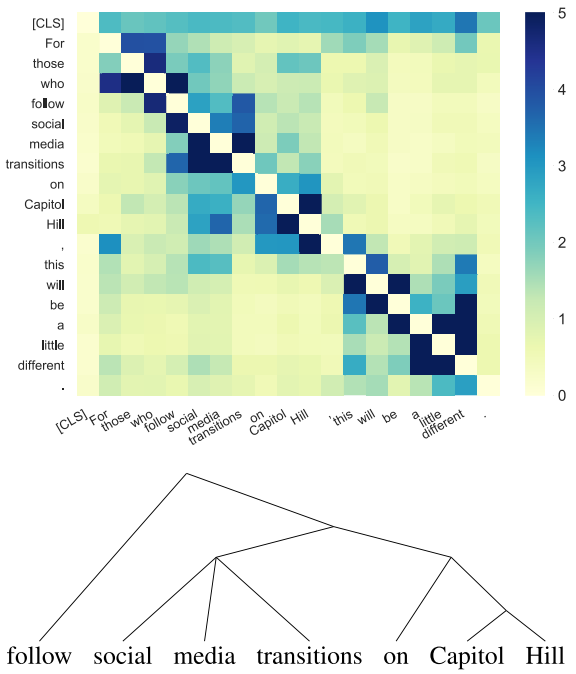

Figure 1: Parameter-free probe for syntactic knowledge: words sharing syntactic subtrees have larger impact on each other in the MLM prediction (Wu et al., 2020).

and add some amount of extra knowledge to the probe. Most recently, Wu et al. (2020) proposed a parameter-free approach based on measuring the impact that one word has on predicting another word within a sequence in the MLM task (Figure 1). They concluded that **BERT ''naturally'' learns some syntactic information, although it is not very similar to linguistic annotated resources**.

The fill-in-the-gap probes of MLM showed that **BERT takes subject-predicate agreement into account when performing the cloze task** (Goldberg, 2019; van Schijndel et al., 2019), even for meaningless sentences and sentences with distractor clauses between the subject and the verb (Goldberg, 2019). A study of negative polarity items (NPIs) by Warstadt et al. (2019) showed that **BERT is better able to detect the presence of NPIs** (e.g., ''ever'') **and the words that allow their use** (e.g., ''whether'') **than scope violations.**

The above claims of syntactic knowledge are belied by the evidence that **BERT does not ''understand'' negation and is insensitive to malformed input**. In particular, its predictions were not altered[2] even with shuffled word order,

[2]See also the recent findings on adversarial triggers, which get the model to produce a certain output even though they

truncated sentences, removed subjects and objects (Ettinger, 2019). This could mean that **either BERT's syntactic knowledge is incomplete, or it does not need to rely on it for solving its tasks**. The latter seems more likely, since Glavaš and Vulić (2020) report that an intermediate fine-tuning step with supervised parsing does not make much difference for downstream task performance.

3.2 Semantic Knowledge

To date, more studies have been devoted to BERT's knowledge of syntactic rather than semantic phenomena. However, we do have evidence from an MLM probing study that **BERT has some knowledge of semantic roles** (Ettinger, 2019). BERT even displays some preference for the incorrect fillers for semantic roles that are semantically related to the correct ones, as opposed to those that are unrelated (e.g., ''to tip a chef'' is better than ''to tip a robin'', but worse than ''to tip a waiter'').

Tenney et al. (2019b) showed that **BERT encodes information about entity types, relations, semantic roles, and proto-roles**, since this information can be detected with probing classifiers.

BERT struggles with representations of numbers. Addition and number decoding tasks showed that BERT does not form good representations for floating point numbers and fails to generalize away from the training data (Wallace et al., 2019b). A part of the problem is BERT's wordpiece tokenization, since numbers of similar values can be divided up into substantially different word chunks.

Out-of-the-box **BERT is surprisingly brittle to named entity replacements**: For example, replacing names in the coreference task changes 85% of predictions (Balasubramanian et al., 2020). This suggests that the model does not actually form a generic idea of named entities, although its F1 scores on NER probing tasks are high (Tenney et al., 2019a). Broscheit (2019) finds that fine-tuning BERT on Wikipedia entity linking ''teaches'' it additional entity knowledge, which would suggest that it did not absorb all the relevant entity information during pre-training on Wikipedia.

are not well-formed from the point of view of a human reader (Wallace et al., 2019a).

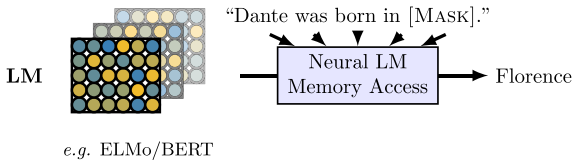

Figure 2: BERT world knowledge (Petroni et al., 2019).

3.3 World Knowledge

The bulk of evidence about commonsense knowledge captured in BERT comes from practitioners using it to extract such knowledge. One direct probing study of BERT reports that **BERT struggles with pragmatic inference and role-based event knowledge** (Ettinger, 2019). BERT also struggles with abstract attributes of objects, as well as visual and perceptual properties that are likely to be assumed rather than mentioned (Da and Kasai, 2019).

The MLM component of BERT is easy to adapt for knowledge induction by filling in the blanks (e.g., ''Cats like to chase [__]''). Petroni et al. (2019) showed that, **for some relation types, vanilla BERT is competitive with methods relying on knowledge bases** (Figure 2), and Roberts et al. (2020) show the same for open-domain QA using the T5 model (Raffel et al., 2019). Davison et al. (2019) suggest that it generalizes better to unseen data. In order to retrieve BERT's knowledge, we need good template sentences, and there is work on their automatic extraction and augmentation (Bouraoui et al., 2019; Jiang et al., 2019b).

However, **BERT cannot reason based on its world knowledge**. Forbes et al. (2019) show that BERT can ''guess'' the affordances and properties of many objects, but cannot reason about the relationship between properties and affordances. For example, it ''knows'' that people can walk into houses, and that houses are big, but it cannot infer that houses are bigger than people. Zhou et al. (2020) and Richardson and Sabharwal (2019) also show that the performance drops with the number of necessary inference steps. Some of BERT's world knowledge success comes from learning stereotypical associations (Poerner et al., 2019), for example, a person with an Italian-sounding name is predicted to be Italian, even when it is incorrect.

3.4 Limitations

Multiple probing studies in section 3 and section 4 report that BERT possesses a surprising amount of syntactic, semantic, and world knowledge. However, Tenney et al. (2019a) remark, ''the fact that a linguistic pattern is not observed by our probing classifier does not guarantee that it is not there, and the observation of a pattern does not tell us how it is used.'' There is also the issue of how complex a probe should be allowed to be (Liu et al., 2019a). If a more complex probe recovers more information, to what extent are we still relying on the original model?

Furthermore, different probing methods may lead to complementary or even contradictory conclusions, which makes a single test (as in most studies) insufficient (Warstadt et al., 2019). A given method might also favor one model over another, for example, RoBERTa trails BERT with one tree extraction method, but leads with another (Htut et al., 2019). The choice of linguistic formalism also matters (Kuznetsov and Gurevych, 2020).

In view of all that, the alternative is to focus on identifying what BERT actually relies on at inference time. This direction is currently pursued both at the level of architecture blocks (to be discussed in detail in subsection 6.3), and at the level of information encoded in model weights. Amnesic probing (Elazar et al., 2020) aims to specifically remove certain information from the model and see how it changes performance, finding, for example, that language modeling does rely on part-of-speech information.

Another direction is information-theoretic probing. Pimentel et al. (2020) operationalize probing as estimating mutual information between the learned representation and a given linguistic property, which highlights that the focus should be not on the amount of information contained in a representation, but rather on how easily it can be extracted from it. Voita and Titov (2020) quantify the amount of effort needed to extract information from a given representation as minimum description length needed to communicate both the probe size and the amount of data required for it to do well on a task.

4 Localizing Linguistic Knowledge

4.1 BERT Embeddings

In studies of BERT, the term ''embedding'' refers to the output of a Transformer layer (typically, the final one). Both conventional static embeddings (Mikolov et al., 2013) and BERT-style embeddings can be viewed in terms of mutual information maximization (Kong et al., 2019), but the latter are **contextualized**. Every token is represented by a vector dependent on the particular context of occurrence, and contains at least some information about that context (Miaschi and Dell'Orletta, 2020).

Several studies reported that **distilled contextualized embeddings better encode lexical semantic information** (i.e., they are better at traditional word-level tasks such as word similarity). The methods to distill a contextualized representation into static include aggregating the information across multiple contexts (Akbik et al., 2019; Bommasani et al., 2020), encoding ''semantically bleached'' sentences that rely almost exclusively on the meaning of a given word (e.g., "This is <>") (May et al., 2019), and even using contextualized embeddings to train static embeddings (Wang et al., 2020d).

But this is not to say that there is no room for improvement. Ethayarajh (2019) measure how similar the embeddings for identical words are in every layer, reporting that later BERT layers produce more context-specific representations.[3] They also find that BERT embeddings occupy a narrow cone in the vector space, and this effect increases from the earlier to later layers. That is, **two random words will on average have a much higher cosine similarity than expected if embeddings were directionally uniform (isotropic)**. Because isotropy was shown to be beneficial for static word embeddings (Mu and Viswanath, 2018), this might be a fruitful direction to explore for BERT.

Because BERT embeddings are contextualized, an interesting question is to what extent they capture phenomena like polysemy and homonymy. There is indeed evidence that **BERT's contextualized embeddings form distinct clusters corresponding to word senses** (Wiedemann et al., 2019; Schmidt and Hofmann, 2020), making BERT successful at word sense disambiguation task. However, Mickus et al. (2019) note that **the representations of the same word depend**

[3] Voita et al. (2019a) look at the evolution of token embeddings, showing that in the earlier Transformer layers, MLM forces the acquisition of contextual information at the expense of the token identity, which gets recreated in later layers.

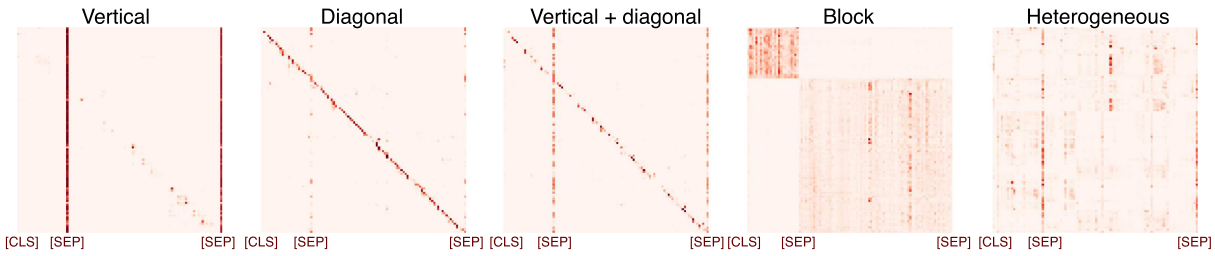

Figure 3: Attention patterns in BERT (Kovaleva et al., 2019).

on the position of the sentence in which it occurs, likely due to the NSP objective. This is not desirable from the linguistic point of view, and could be a promising avenue for future work.

The above discussion concerns token embeddings, but BERT is typically used as a sentence or text encoder. The standard way to generate sentence or text representations for classification is to use the [CLS] token, but alternatives are also being discussed, including concatenation of token representations (Tanaka et al., 2020), normalized mean (Tanaka et al., 2020), and layer activations (Ma et al., 2019). See Toshniwal et al. (2020) for a systematic comparison of several methods across tasks and sentence encoders.

4.2 Self-attention Heads

Several studies proposed classification of attention head types. Raganato and Tiedemann (2018) discuss attending to the token itself, previous/next tokens, and the sentence end. Clark et al. (2019) distinguish between attending to previous/next tokens, `[CLS]`, `[SEP]`, punctuation, and ''attending broadly'' over the sequence. Kovaleva et al. (2019) propose five patterns, shown in Figure 3.

4.2.1 Heads With Linguistic Functions

The ''heterogeneous'' attention pattern shown in Figure 3 *could* potentially be linguistically interpretable, and a number of studies focused on identifying the functions of self-attention heads. In particular, **some BERT heads seem to specialize in certain types of syntactic relations.** Htut et al. (2019) and Clark et al. (2019) report that there are BERT heads that attended significantly more than a random baseline to words in certain syntactic positions. The datasets and methods used in these studies differ, but they both find that there are heads that attend to words in `obj` role more than the positional baseline. The evidence for `nsubj`, `advmod`, and `amod` varies between these two studies. The overall conclusion is also supported by Voita et al.'s (2019b) study of the base Transformer in machine translation context. Hoover et al. (2019) hypothesize that even complex dependencies like `dobj` are encoded by a combination of heads rather than a single head, but this work is limited to qualitative analysis. Zhao and Bethard (2020) looked specifically for the heads encoding negation scope.

Both Clark et al. (2019) and Htut et al. (2019) conclude that **no single head has the complete syntactic tree information**, in line with evidence of partial knowledge of syntax (cf. subsection 3.1). However, Clark et al. (2019) identify a BERT head that can be directly used as a classifier to perform coreference resolution on par with a rule-based system, which by itself would seem to require quite a lot of syntactic knowledge.

Lin et al. (2019) present evidence that **attention weights are weak indicators of subject-verb agreement and reflexive anaphora.** Instead of serving as strong pointers between tokens that should be related, BERT's self-attention weights were close to a uniform attention baseline, but there was some sensitivity to different types of distractors coherent with psycholinguistic data. This is consistent with conclusions by Ettinger (2019).

To our knowledge, morphological information in BERT heads has not been addressed, but with the sparse attention variant by Correia et al. (2019) in the base Transformer, some attention heads appear to merge BPE-tokenized words. For semantic relations, there are reports of self-attention heads encoding core frame-semantic relations (Kovaleva et al., 2019), as well as lexicographic and commonsense relations (Cui et al., 2020).

The overall popularity of self-attention as an interpretability mechanism is due to the idea that ''attention weight has a clear meaning: how much

a particular word will be weighted when computing the next representation for the current word'' (Clark et al., 2019). This view is currently debated (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Brunner et al., 2020), and in a multilayer model where attention is followed by nonlinear transformations, the patterns in individual heads do not provide a full picture. Also, although many current papers are accompanied by attention visualizations, and there is a growing number of visualization tools (Vig, 2019; Hoover et al., 2019), the visualization is typically limited to qualitative analysis (often with cherry-picked examples) (Belinkov and Glass, 2019), and should not be interpreted as definitive evidence.

4.2.2 Attention to Special Tokens

Kovaleva et al. (2019) show that **most self-attention heads do not directly encode any non-trivial linguistic information**, at least when fine-tuned on GLUE (Wang et al., 2018), since only fewer than 50% of heads exhibit the ''heterogeneous'' pattern. Much of the model produced the vertical pattern (attention to `[CLS]`, `[SEP]`, and punctuation tokens), consistent with the observations by Clark et al. (2019). This redundancy is likely related to the overparameterization issue (see section 6).

More recently, Kobayashi et al. (2020) showed that the norms of attention-weighted input vectors, which yield a more intuitive interpretation of self-attention, reduce the attention to special tokens. However, even when the attention weights are normed, it is still not the case that most heads that do the ''heavy lifting'' are even potentially interpretable (Prasanna et al., 2020).

One methodological choice in in many studies of attention is to focus on inter-word attention and simply exclude special tokens (e.g., Lin et al. [2019] and Htut et al. [2019]). However, if attention to special tokens actually matters at inference time, drawing conclusions purely from inter-word attention patterns does not seem warranted.

The functions of special tokens are not yet well understood. `[CLS]` is typically viewed as an aggregated sentence-level representation (although all token representations also contain at least some sentence-level information, as discussed in subsection 4.1); in that case, we may not see, for example, full syntactic trees in inter-word attention because part of that information is actually packed in `[CLS]`.

Clark et al. (2019) experiment with encoding Wikipedia paragraphs with base BERT to consider specifically the attention to special tokens, noting that heads in early layers attend more to `[CLS]`, in middle layers to `[SEP]`, and in final layers to periods and commas. They hypothesize that its function might be one of ''no-op'', a signal to ignore the head if its pattern is not applicable to the current case. As a result, for example, `[SEP]` gets increased attention starting in layer 5, but its importance for prediction drops. However, after fine-tuning both `[SEP]` and `[CLS]` get a lot of attention, depending on the task (Kovaleva et al., 2019). Interestingly, BERT also pays a lot of attention to punctuation, which Clark et al. (2019) explain by the fact that periods and commas are simply almost as frequent as the special tokens, and so the model might learn to rely on them for the same reasons.

4.3 BERT Layers

The first layer of BERT receives as input a combination of token, segment, and positional embeddings.

It stands to reason that **the lower layers have the most information about linear word order.** Lin et al. (2019) report a decrease in the knowledge of linear word order around layer 4 in BERT-base. This is accompanied by an increased knowledge of hierarchical sentence structure, as detected by the probing tasks of predicting the token index, the main auxiliary verb and the sentence subject.

There is a wide consensus in studies with different tasks, datasets, and methodologies that **syntactic information is most prominent in the middle layers of BERT.**[4] Hewitt and Manning (2019) had the most success reconstructing syntactic tree depth from the middle BERT layers (6-9 for base-BERT, 14-19 for BERT-large). Goldberg (2019) reports the best subject-verb agreement around layers 8-9, and the performance on syntactic probing tasks used by Jawahar et al. (2019) also seems to peak around the middle of the model. The prominence of syntactic information in the middle BERT layers is related to Liu et al.'s

[4]These BERT results are also compatible with findings by Vig and Belinkov (2019), who report the highest attention to tokens in dependency relations in the middle layers of GPT-2.

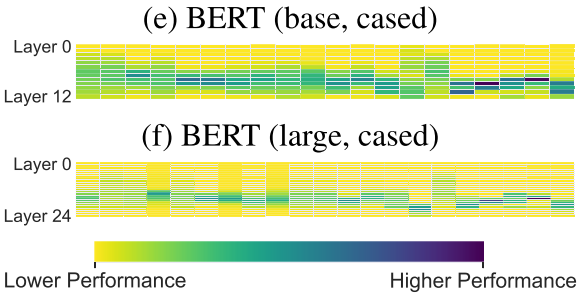

Figure 4: BERT layer transferability (columns correspond to probing tasks, Liu et al. (2019a).

(2019a) observation that the middle layers of Transformers are best-performing overall and the most transferable across tasks (see Figure 4).

There is **conflicting evidence about syntactic chunks**. Tenney et al. (2019a) conclude that ''the basic syntactic information appears earlier in the network while high-level semantic features appear at the higher layers'', drawing parallels between this order and the order of components in a typical NLP pipeline—from POS-tagging to dependency parsing to semantic role labeling. Jawahar et al. (2019) also report that the lower layers were more useful for chunking, while middle layers were more useful for parsing. At the same time, the probing experiments by Liu et al. (2019a) find the opposite: Both POS-tagging and chunking were performed best at the middle layers, in both BERT-base and BERT-large. However, all three studies use different suites of probing tasks.

The final layers of BERT are the most task-specific. In pre-training, this means specificity to the MLM task, which explains why the middle layers are more transferable (Liu et al., 2019a). In fine-tuning, it explains why the final layers change the most (Kovaleva et al., 2019), and why restoring the weights of lower layers of fine-tuned BERT to their original values does not dramatically hurt the model performance (Hao et al., 2019).

Tenney et al. (2019a) suggest that whereas syntactic information appears early in the model and can be localized, **semantics is spread across the entire model**, which explains why certain non-trivial examples get solved incorrectly at first but correctly at the later layers. This is rather to be expected: Semantics permeates all language, and linguists debate whether meaningless structures can exist at all (Goldberg, 2006, p.166–182). But this raises the question of what stacking more Transformer layers in BERT actually achieves in terms of the spread of semantic knowledge, and whether that is beneficial. Tenney et al. compared BERT-base and BERT-large, and found that the overall pattern of cumulative score gains is the same, only more spread out in the larger model.

Note that Tenney et al.'s (2019a) experiments concern sentence-level semantic relations; Cui et al. (2020) report that the encoding of ConceptNet semantic relations is the worst in the early layers and increases towards the top. Jawahar et al. (2019) place ''surface features in lower layers, syntactic features in middle layers and semantic features in higher layers'', but their conclusion is surprising, given that only one semantic task in this study actually topped at the last layer, and three others peaked around the middle and then considerably degraded by the final layers.

5 Training BERT

This section reviews the proposals to optimize the training and architecture of the original BERT.

5.1 Model Architecture Choices

To date, the most systematic study of BERT architecture was performed by Wang et al. (2019b), who experimented with the number of layers, heads, and model parameters, varying one option and freezing the others. They concluded that **the number of heads was not as significant as the number of layers**. That is consistent with the findings of Voita et al. (2019b) and Michel et al. (2019) (section 6), and also the observation by Liu et al. (2019a) that the middle layers were the most transferable. Larger hidden representation size was consistently better, but the gains varied by setting.

All in all, **changes in the number of heads and layers appear to perform different functions**. The issue of model depth must be related to the information flow from the most task-specific layers closer to the classifier (Liu et al., 2019a), to the initial layers which appear to be the most task-invariant (Hao et al., 2019), and where the tokens resemble the input tokens the most (Brunner et al., 2020) (see subsection 4.3). If that is the case, a deeper model has more capacity to encode information that is not task-specific.

On the other hand, many self-attention heads in vanilla BERT seem to naturally learn the same patterns (Kovaleva et al., 2019). This explains

why pruning them does not have too much impact. The question that arises from this is how far we could get with intentionally encouraging diverse self-attention patterns: Theoretically, this would mean increasing the amount of information in the model with the same number of weights. Raganato et al. (2020) show for Transformer-based machine translation we can simply pre-set the patterns that we already know the model would learn, instead of learning them from scratch.

Vanilla BERT is symmetric and balanced in terms of self-attention and feed-forward layers, but it may not have to be. For the base Transformer, Press et al. (2020) report benefits from more self-attention sublayers at the bottom and more feedforward sublayers at the top.

5.2 Improvements to the Training Regime

Liu et al. (2019b) demonstrate **the benefits of large-batch training**: With 8k examples, both the language model perplexity and downstream task performance are improved. They also publish their recommendations for other parameters. You et al. (2019) report that with a batch size of 32k BERT's training time can be significantly reduced with no degradation in performance. Zhou et al. (2019) observe that the normalization of the trained `[CLS]` token stabilizes the training and slightly improves performance on text classification tasks.

Gong et al. (2019) note that, because self-attention patterns in higher and lower layers are similar, **the model training can be done in a recursive manner**, where the shallower version is trained first and then the trained parameters are copied to deeper layers. Such a ''warm-start'' can lead to a 25% faster training without sacrificing performance.

5.3 Pre-training BERT

The original BERT is a bidirectional Transformer pre-trained on two tasks: NSP and MLM (section 2). Multiple studies have come up with **alternative training objectives** to improve on BERT, and these could be categorized as follows:

- **How to mask.** Raffel et al. (2019) systematically experiment with corruption rate and corrupted span length. Liu et al. (2019b) propose diverse masks for training examples within an epoch, while Baevski et al. (2019) mask every token in a sequence instead of a random selection. Clinchant et al. (2019) replace the MASK token with `[UNK]` token, to help the model learn a representation for unknowns that could be useful for translation. Song et al. (2020) maximize the amount of information available to the model by conditioning on both masked and unmasked tokens, and letting the model see how many tokens are missing.

- **What to mask.** Masks can be applied to full words instead of word-pieces (Devlin et al., 2019; Cui et al., 2019). Similarly, we can mask spans rather than single tokens (Joshi et al., 2020), predicting how many are missing (Lewis et al., 2019). Masking phrases and named entities (Sun et al., 2019b) improves representation of structured knowledge.

- **Where to mask.** Lample and Conneau (2019) use arbitrary text streams instead of sentence pairs and subsample frequent outputs similar to Mikolov et al. (2013). Bao et al. (2020) combine the standard autoencoding MLM with partially autoregressive LM objective using special pseudo mask tokens.

- **Alternatives to masking.** Raffel et al. (2019) experiment with replacing and dropping spans; Lewis et al. (2019) explore deletion, infilling, sentence permutation and document rotation; and Sun et al. (2019c) predict whether a token is capitalized and whether it occurs in other segments of the same document. Yang et al. (2019) train on different permutations of word order in the input sequence, maximizing the probability of the original word order (cf. the n-gram word order reconstruction task (Wang et al., 2019a)). Clark et al. (2020) detects tokens that were replaced by a generator network rather than masked.

- **NSP alternatives.** Removing NSP does not hurt or slightly improves performance (Liu et al., 2019b; Joshi et al., 2020; Clinchant et al., 2019). Wang et al. (2019a) and Cheng et al. (2019) replace NSP with the task of predicting both the next and the previous sentences. Lan et al. (2020) replace the negative NSP examples by swapped

sentences from positive examples, rather than sentences from different documents. ERNIE 2.0 includes sentence reordering and sentence distance prediction. Bai et al. (2020) replace both NSP and token position embeddings by a combination of paragraph, sentence, and token index embeddings. Li and Choi (2020) experiment with utterance order prediction task for multiparty dialogue (and also MLM at the level of utterances and the whole dialogue).

- **Other tasks.** Sun et al. (2019c) propose simultaneous learning of seven tasks, including discourse relation classification and predicting whether a segment is relevant for IR. Guu et al. (2020) include a latent knowledge retriever in language model pretraining. Wang et al. (2020c) combine MLM with a knowledge base completion objective. Glass et al. (2020) replace MLM with span prediction task (as in extractive question answering), where the model is expected to provide the answer not from its own weights, but from a *different* passage containing the correct answer (a relevant search engine query snippet).

Another obvious source of improvement is pre-training data. Several studies explored the benefits of increasing the corpus volume (Liu et al., 2019b; Conneau et al., 2019; Baevski et al., 2019) and longer training (Liu et al., 2019b). The data also does not have to be raw text: There is a number efforts to **incorporate explicit linguistic information**, both syntactic (Sundararaman et al., 2019) and semantic (Zhang et al., 2020). Wu et al. (2019b) and Kumar et al. (2020) include the label for a given sequence from an annotated task dataset. Schick and Schütze (2020) separately learn representations for rare words.

Although BERT is already actively used as a source of world knowledge (see subsection 3.3), there is also work on **explicitly supplying structured knowledge**. One approach is entity-enhanced models. For example, Peters et al. (2019a); Zhang et al. (2019) include entity embeddings as input for training BERT, while Poerner et al. (2019) adapt entity vectors to BERT representations. As mentioned above, Wang et al. (2020c) integrate knowledge not through entity

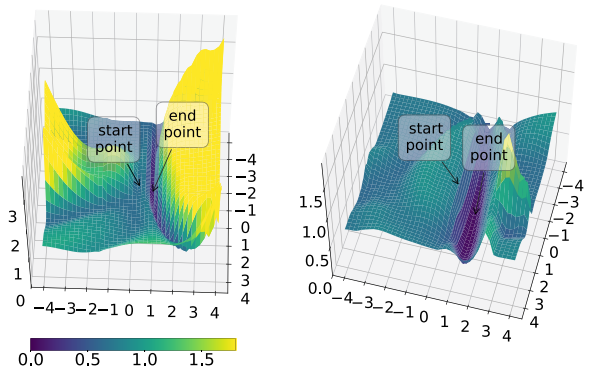

Figure 5: Pre-trained weights help BERT find wider optima in fine-tuning on MRPC (right) than training from scratch (left) (Hao et al., 2019).

embeddings, but through the additional pre-training objective of knowledge base completion. Sun et al. (2019b,c) modify the standard MLM task to mask named entities rather than random words, and Yin et al. (2020) train with MLM objective over both text and linearized table data. Wang et al. (2020a) enhance RoBERTa with both linguistic and factual knowledge with task-specific adapters.

Pre-training is the most expensive part of training BERT, and it would be informative to know how much benefit it provides. On some tasks, a randomly initialized and fine-tuned BERT obtains competitive or higher results than the pre-trained BERT with the task classifier and frozen weights (Kovaleva et al., 2019). The consensus in the community is that pre-training does help in most situations, but the degree and its exact contribution requires further investigation. Prasanna et al. (2020) found that *most* weights of pre-trained BERT are useful in fine-tuning, although there are ‘‘better’’ and ‘‘worse’’ subnetworks. One explanation is that pre-trained weights help the fine-tuned BERT find wider and flatter areas with smaller generalization error, which makes the model more robust to overfitting (see Figure 5 from Hao et al. [2019]).

Given the large number and variety of proposed modifications, one would wish to know how much impact each of them has. However, due to the overall trend towards large model sizes, systematic ablations have become expensive. Most new models claim superiority on standard benchmarks, but gains are often marginal, and estimates of model stability and significance testing are very rare.

5.4 Fine-tuning BERT

Pre-training + fine-tuning workflow is a crucial part of BERT. The former is supposed to provide task-independent knowledge, and the latter would presumably teach the model to rely more on the representations useful for the task at hand.

Kovaleva et al. (2019) did not find that to be the case for BERT fine-tuned on GLUE tasks:[5] during fine-tuning, the most changes for three epochs occurred in the last two layers of the models, but those changes caused self-attention to focus on `[SEP]` rather than on linguistically interpretable patterns. It is understandable why fine-tuning would increase the attention to `[CLS]`, but not `[SEP]`. If Clark et al. (2019) are correct that `[SEP]` serves as ''no-op'' indicator, fine-tuning basically tells BERT what to ignore.

Several studies explored the possibilities of improving the fine-tuning of BERT:

- **Taking more layers into account**: learning a complementary representation of the information in deep and output layers (Yang and Zhao, 2019), using a weighted combination of all layers instead of the final one (Su and Cheng, 2019; Kondratyuk and Straka, 2019), and layer dropout (Kondratyuk and Straka, 2019).
- **Two-stage fine-tuning** introduces an intermediate supervised training stage between pre-training and fine-tuning (Phang et al., 2019; Garg et al., 2020; Arase and Tsujii, 2019; Pruksachatkun et al., 2020; Glavaš and Vulić, 2020). Ben-David et al. (2020) propose a pivot-based variant of MLM to fine-tune BERT for domain adaptation.
- **Adversarial token perturbations** improve the robustness of the model (Zhu et al., 2019).
- **Adversarial regularization** in combination with *Bregman Proximal Point Optimization* helps alleviate pre-trained knowledge forgetting and therefore prevents BERT from overfitting to downstream tasks (Jiang et al., 2019a).
- **Mixout regularization** improves the stability of BERT fine-tuning even for a small number of training examples (Lee et al., 2019).

With large models, even fine-tuning becomes expensive, but Houlsby et al. (2019) show that it can be successfully approximated with adapter modules. They achieve competitive performance on 26 classification tasks at a fraction of the computational cost. Adapters in BERT were also used for multitask learning (Stickland and Murray, 2019) and cross-lingual transfer (Artetxe et al., 2019). An alternative to fine-tuning is extracting features from frozen representations, but fine-tuning works better for BERT (Peters et al., 2019b).

A big methodological challenge in the current NLP is that the reported performance improvements of new models may well be within variation induced by environment factors (Crane, 2018). BERT is not an exception. Dodge et al. (2020) report significant variation for BERT fine-tuned on GLUE tasks due to both weight initialization and training data order. They also propose early stopping on the less-promising seeds.

Although we hope that the above observations may be useful for the practitioners, this section does not exhaust the current research on fine-tuning and its alternatives. For example, we do not cover such topics as Siamese architectures, policy gradient training, automated curriculum learning, and others.

6 How Big Should BERT Be?

6.1 Overparameterization

Transformer-based models keep growing by orders of magnitude: The 110M parameters of base BERT are now dwarfed by 17B parameters of Turing-NLG (Microsoft, 2020), which is dwarfed by 175B of GPT-3 (Brown et al., 2020). This trend raises concerns about computational complexity of self-attention (Wu et al., 2019a), environmental issues (Strubell et al., 2019; Schwartz et al., 2019), fair comparison of architectures (Aßenmacher and Heumann, 2020), and reproducibility.

Human language is incredibly complex, and would perhaps take many more parameters to describe fully, but the current models do not make good use of the parameters they already have. Voita et al. (2019b) showed that **all but a few Transformer heads could be pruned without**

[5] Kondratyuk and Straka (2019) suggest that fine-tuning on Universal Dependencies does result in syntactically meaningful attention patterns, but there was no quantitative evaluation.

		Compression	Performance	Speedup	Model	Evaluation
	BERT-base (Devlin et al., 2019)	×1	100%	×1	$BERT_{12}$	All GLUE tasks, SQuAD
	BERT-small	×3.8	91%	–	$BERT_4$†	All GLUE tasks
Distillation	DistilBERT (Sanh et al., 2019)	×1.5	90%§	×1.6	$BERT_6$	All GLUE tasks, SQuAD
	$BERT_6$-PKD (Sun et al., 2019a)	×1.6	98%	×1.9	$BERT_6$	No WNLI, CoLA, STS-B; RACE
	$BERT_3$-PKD (Sun et al., 2019a)	×2.4	92%	×3.7	$BERT_3$	No WNLI, CoLA, STS-B; RACE
	Aguilar et al. (2019), Exp. 3	×1.6	93%	–	$BERT_6$	CoLA, MRPC, QQP, RTE
	BERT-48 (Zhao et al., 2019)	×62	87%	×77	$BERT_{12}$*†	MNLI, MRPC, SST-2
	BERT-192 (Zhao et al., 2019)	×5.7	93%	×22	$BERT_{12}$*†	MNLI, MRPC, SST-2
	TinyBERT (Jiao et al., 2019)	×7.5	96%	×9.4	$BERT_4$†	No WNLI; SQuAD
	MobileBERT (Sun et al., 2020)	×4.3	100%	×4	$BERT_{24}$†	No WNLI; SQuAD
	PD (Turc et al., 2019)	×1.6	98%	×2.5‡	$BERT_6$†	No WNLI, CoLA and STS-B
	WaLDORf (Tian et al., 2019)	×4.4	93%	×9	$BERT_8$†‖	SQuAD
	MiniLM (Wang et al., 2020b)	×1.65	99%	×2	$BERT_6$	No WNLI, STS-B, $MNLI_{mm}$; SQuAD
	MiniBERT(Tsai et al., 2019)	×6**	98%	×27**	$mBERT_3$†	CoNLL-18 POS and morphology
	BiLSTM-soft (Tang et al., 2019)	×110	91%	×434‡	$BiLSTM_1$	MNLI, QQP, SST-2
Quantization	Q-BERT-MP (Shen et al., 2019)	×13	98%¶	–	$BERT_{12}$	MNLI, SST-2, CoNLL-03, SQuAD
	BERT-QAT (Zafrir et al., 2019)	×4	99%	–	$BERT_{12}$	No WNLI, MNLI; SQuAD
	GOBO (Zadeh and Moshovos, 2020)	×9.8	99%	–	$BERT_{12}$	MNLI
Pruning	McCarley et al. (2020), ff2	×2.2‡	98%‡	×1.9‡	$BERT_{24}$	SQuAD, Natural Questions
	RPP (Guo et al., 2019)	×1.7‡	99%‡	–	$BERT_{24}$	No WNLI, STS-B; SQuAD
	Soft MvP (Sanh et al., 2020)	×33	94%¶	–	$BERT_{12}$	MNLI, QQP, SQuAD
	IMP (Chen et al., 2020), rewind 50%	×1.4–2.5	94–100%	–	$BERT_{12}$	No MNLI-mm; SQuAD
Other	ALBERT-base (Lan et al., 2020)	×9	97%	–	$BERT_{12}$†	MNLI, SST-2
	ALBERT-xxlarge (Lan et al., 2020)	×0.47	107%	–	$BERT_{12}$†	MNLI, SST-2
	BERT-of-Theseus (Xu et al., 2020)	×1.6	98%	×1.9	$BERT_6$	No WNLI
	PoWER-BERT (Goyal et al., 2020)	N/A	99%	×2–4.5	$BERT_{12}$	No WNLI; RACE

Table 1: Comparison of BERT compression studies. Compression, performance retention, and inference time speedup figures are given with respect to $BERT_{base}$, unless indicated otherwise. Performance retention is measured as a ratio of average scores achieved by a given model and by $BERT_{base}$. The subscript in the model description reflects the number of layers used. *Smaller vocabulary used. †The dimensionality of the hidden layers is reduced. ‖Convolutional layers used. ‡Compared to $BERT_{large}$. **Compared to mBERT. §As reported in Jiao et al. (2019).¶In comparison to the dev set.

significant losses in performance. For BERT, Clark et al. (2019) observe that most heads in the same layer show similar self-attention patterns (perhaps related to the fact that the output of all self-attention heads in a layer is passed through the same MLP), which explains why Michel et al. (2019) were able to reduce most layers to a single head.

Depending on the task, some BERT heads/layers are not only redundant (Kao et al., 2020), but also harmful to the downstream task performance. **Positive effect from head disabling** was reported for machine translation (Michel et al., 2019), abstractive summarization (Baan et al., 2019), and GLUE tasks (Kovaleva et al., 2019). Additionally, Tenney et al. (2019a) examine the cumulative gains of their structural probing classifier, observing that in 5 out of 8 probing tasks some layers cause a drop in scores (typically in the final layers). Gordon et al. (2020) find that 30%–40% of the weights can be pruned without impact on downstream tasks.

In general, larger BERT models perform better (Liu et al., 2019a; Roberts et al., 2020), but not always: BERT-base outperformed BERT-large on subject-verb agreement (Goldberg, 2019) and sentence subject detection (Lin et al., 2019). Given the complexity of language, and amounts of pre-training data, it is not clear why BERT ends up with redundant heads and layers. Clark et al. (2019) suggest that one possible reason is the use of attention dropouts, which causes some attention weights to be zeroed-out during training.

6.2 Compression Techniques

Given the above evidence of overparameterization, it does not come as a surprise that **BERT can be efficiently compressed with minimal accuracy loss**, which would be highly desirable for real-world applications. Such efforts to date are summarized in Table 1. The main approaches are knowledge distillation, quantization, and pruning.

The studies in the **knowledge distillation framework** (Hinton et al., 2014) use a smaller

student-network trained to mimic the behavior of a larger teacher-network. For BERT, this has been achieved through experiments with loss functions (Sanh et al., 2019; Jiao et al., 2019), mimicking the activation patterns of individual portions of the teacher network (Sun et al., 2019a), and knowledge transfer at the pre-training (Turc et al., 2019; Jiao et al., 2019; Sun et al., 2020) or fine-tuning stage (Jiao et al., 2019). McCarley et al. (2020) suggest that distillation has so far worked better for GLUE than for reading comprehension, and report good results for QA from a combination of structured pruning and task-specific distillation.

Quantization decreases BERT's memory footprint through lowering the precision of its weights (Shen et al., 2019; Zafrir et al., 2019). Note that this strategy often requires compatible hardware.

As discussed in section 6, individual self-attention heads and BERT layers can be disabled without significant drop in performance (Michel et al., 2019; Kovaleva et al., 2019; Baan et al., 2019). **Pruning** is a compression technique that takes advantage of that fact, typically reducing the amount of computation via zeroing out of certain parts of the large model. In structured pruning, architecture blocks are dropped, as in LayerDrop (Fan et al., 2019). In unstructured, the weights in the entire model are pruned irrespective of their location, as in magnitude pruning (Chen et al., 2020) or movement pruning (Sanh et al., 2020).

Prasanna et al. (2020) and Chen et al. (2020) explore BERT from the perspective of the lottery ticket hypothesis (Frankle and Carbin, 2019), looking specifically at the ''winning'' subnetworks in pre-trained BERT. They independently find that such subnetworks do exist, and that transferability between subnetworks for different tasks varies.

If the ultimate goal of training BERT is compression, Li et al. (2020) recommend training larger models and compressing them heavily rather than compressing smaller models lightly.

Other techniques include decomposing BERT's embedding matrix into smaller matrices (Lan et al., 2020), progressive module replacing (Xu et al., 2020), and dynamic elimination of intermediate encoder outputs (Goyal et al., 2020). See Ganesh et al. (2020) for a more detailed discussion of compression methods.

6.3 Pruning and Model Analysis

There is a nascent discussion around pruning as a model analysis technique. The basic idea is that a compressed model a priori consists of elements that are useful for prediction; therefore by finding out what they do we may find out what the whole network does. For instance, BERT has heads that seem to encode frame-semantic relations, but disabling them might not hurt downstream task performance (Kovaleva et al., 2019); this suggests that this knowledge is not actually used.

For the base Transformer, Voita et al. (2019b) identify the functions of self-attention heads and then check which of them survive the pruning, finding that the syntactic and positional heads are the last ones to go. For BERT, Prasanna et al. (2020) go in the opposite direction: pruning on the basis of importance scores, and interpreting the remaining ''good'' subnetwork. With respect to self-attention heads specifically, it does not seem to be the case that only the heads that potentially encode non-trivial linguistic patterns survive the pruning.

The models and methodology in these studies differ, so the evidence is inconclusive. In particular, Voita et al. (2019b) find that before pruning the majority of heads are syntactic, and Prasanna et al. (2020) find that the majority of heads do not have potentially non-trivial attention patterns.

An important limitation of the current head and layer ablation studies (Michel et al., 2019; Kovaleva et al., 2019) is that they inherently assume that certain knowledge is contained in heads/layers. However, there is evidence of more diffuse representations spread across the full network, such as the gradual increase in accuracy on difficult semantic parsing tasks (Tenney et al., 2019a) or the absence of heads that would perform parsing ''in general'' (Clark et al., 2019; Htut et al., 2019). If so, ablating individual components harms the weight-sharing mechanism. Conclusions from component ablations are also problematic if the same information is duplicated elsewhere in the network.

7 Directions for Further Research

BERTology has clearly come a long way, but it is fair to say we still have more questions than answers about how BERT works. In this section,

we list what we believe to be the most promising directions for further research.

Benchmarks that require verbal reasoning. Although BERT enabled breakthroughs on many NLP benchmarks, a growing list of analysis papers are showing that its language skills are not as impressive as they seem. In particular, they were shown to rely on shallow heuristics in natural language inference (McCoy et al., 2019b; Zellers et al., 2019; Jin et al., 2020), reading comprehension (Si et al., 2019; Rogers et al., 2020; Sugawara et al., 2020; Yogatama et al., 2019), argument reasoning comprehension (Niven and Kao, 2019), and text classification (Jin et al., 2020). Such heuristics can even be used to reconstruct a non–publicly available model (Krishna et al., 2020). As with any optimization method, if there is a shortcut in the data, we have no reason to expect BERT to not learn it. But harder datasets that cannot be resolved with shallow heuristics are unlikely to emerge if their development is not as valued as modeling work.

Benchmarks for the full range of linguistic competence. Although the language models seem to acquire a great deal of knowledge about language, we do not currently have comprehensive stress tests for different aspects of linguistic knowledge. A step in this direction is the ''Checklist'' behavioral testing (Ribeiro et al., 2020), the best paper at ACL 2020. Ideally, such tests would measure not only errors, but also sensitivity (Ettinger, 2019).

Developing methods to ''teach'' reasoning. While large pre-trained models have a lot of knowledge, they often fail if any reasoning needs to be performed on top of the facts they possess (Talmor et al., 2019, see also subsection 3.3). For instance, Richardson et al. (2020) propose a method to ''teach'' BERT quantification, conditionals, comparatives, and Boolean coordination.

Learning what happens at inference time. Most BERT analysis papers focus on different probes of the model, with the goal to find what the language model ''knows''. However, probing studies have limitations (subsection 3.4), and to this point, far fewer papers have focused on discovering what knowledge actually gets used. Several promising directions are the ''amnesic probing'' (Elazar et al., 2020), identifying features important for prediction for a given task (Arkhangelskaia and Dutta, 2019), and pruning the model to remove the non-important components (Voita et al., 2019b; Michel et al., 2019; Prasanna et al., 2020).

8 Conclusion

In a little over a year, BERT has become a ubiquitous baseline in NLP experiments and inspired numerous studies analyzing the model and proposing various improvements. The stream of papers seems to be accelerating rather than slowing down, and we hope that this survey helps the community to focus on the biggest unresolved questions.

9 Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work is funded in part by NSF award number IIS-1844740 to Anna Rumshisky.

References

Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Edward Guo. 2019. Knowledge Distillation from Internal Representations. *arXiv preprint arXiv:1910.03723*.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled Contextualized Embeddings for Named Entity Recognition. 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 724–728, Minneapolis, Minnesota. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N19-1078

Yuki Arase and Jun'ichi Tsujii. 2019. Transfer Fine-Tuning: A BERT Case Study. 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 5393–5404, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1542

Ekaterina Arkhangelskaia and Sourav Dutta. 2019. Whatcha lookin'at? DeepLIFTing BERT's Attention in Question Answering. *arXiv preprint arXiv:1910.06431*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the Cross-lingual Transferability of Monolingual Representations. *arXiv:1911.03310 [cs]*. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.421

Matthias Aßenmacher and Christian Heumann. 2020. On the comparability of Pre-Trained Language Models. *arXiv:2001.00781 [cs, stat]*.

Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. 2019. Understanding Multi-Head Attention in Abstractive Summarization. *arXiv preprint arXiv:1911.03898*.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. Cloze-driven Pretraining of Self-Attention Networks. 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 5360–5369, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1539

He Bai, Peng Shi, Jimmy Lin, Luchen Tan, Kun Xiong, Wen Gao, and Ming Li. 2020. Sega BERT: Pre-training of Segment-aware BERT for Language Understanding. *arXiv:2004.14996 [cs]*.

Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi. 2020. What's in a Name? Are BERT Named Entity Representations just as Good for any other Name? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 205–214, Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.repl4nlp-1.24

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2020. UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. *arXiv:2002.12804 [cs]*.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72. **DOI:** https://doi.org/10.1162/tacl_a_00254

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. PERL: Pivot-based Domain Adaptation for Pre-trained Deep Contextualized Embedding Models. *arXiv:2006.09075 [cs]*. **DOI:** https://doi.org/10.1162/tacl_a_00328

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.431

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2019. Inducing Relational Knowledge from BERT. *arXiv:1911.12753 [cs]*. **DOI:** https://doi.org/10.1609/aaai.v34i05.6242

Samuel Broscheit. 2019. Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/K19-1063

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *arXiv:2005.14165 [cs]*.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On Identifiability in Transformers. In *International Conference on Learning Representations*.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The Lottery Ticket Hypothesis for Pre-trained BERT Networks. *arXiv:2007.12223 [cs, stat]*.

Xingyi Cheng, Weidi Xu, Kunlong Chen, Wei Wang, Bin Bi, Ming Yan, Chen Wu, Luo Si, Wei Chu, and Taifeng Wang. 2019. Symmetric Regularization based BERT for Pair-Wise Semantic Reasoning. *arXiv:1909.03405 [cs]*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W19-4828, **PMID:** 31709923

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of BERT for Neural Machine Translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117, Hong Kong. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-5611

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-Lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.747

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. 2019. Adaptively Sparse Transformers. 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 2174–2184, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1223

Matt Crane. 2018. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Transactions of the Association for Computational Linguistics*, 6:241–252. **DOI:** https://doi.org/10.1162/tacl_a_00018

Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2020. Does BERT Solve Commonsense Task via Commonsense Knowledge? *arXiv:2008.03945 [cs]*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-Training with Whole Word Masking for Chinese BERT. *arXiv:1906.08101 [cs]*.

Jeff Da and Jungo Kasai. 2019. Cracking the Contextual Commonsense Code: Understanding Commonsense Reasoning Aptitude of Deep Contextual Representations. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 1–12, Hong Kong, China. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense Knowledge Mining from Pretrained Models. 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 1173–1178, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1109

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.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv:2002.06305 [cs]*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2020. When Bert Forgets How To POS: Amnesic Probing of Linguistic Properties and MLM Predictions. *arXiv:2006.00995 [cs]*.

Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1006

Allyson Ettinger. 2019. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *arXiv:1907.13528 [cs]*. **DOI:** https://doi.org/10.1162/tacl_a_00298

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing Transformer Depth on Demand with Structured Dropout. In *International Conference on Learning Representations*.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do Neural Language Representations Learn Physical Commonsense? In *Proceedings of the 41st Annual Conference of the Cognitive Science Society (CogSci 2019)*, page 7.

Jonathan Frankle and Michael Carbin. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*.

Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Deming Chen, Marianne Winslett, Hassan Sajjad, and Preslav Nakov. 2020. Compressing large-scale transformer-based models: A case study on BERT. *arXiv preprint arXiv:2002.11985*.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection. In *AAAI*. **DOI:** https://doi.org/10.1609/aaai.v34i05.6282

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G.P. Shrivatsa Bhargav, Dinesh Garg, and Avi Sil. 2020. Span Selection Pre-training for Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782, Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.247

Goran Glavaš and Ivan Vulić. 2020. Is Supervised Syntactic Parsing Beneficial for Language Understanding? An Empirical Investigation. *arXiv:2008.06788 [cs]*.

Adele Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*, Oxford University Press, USA.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Linyuan Gong, Di He, Zhuohan Li, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Efficient training of BERT by progressively stacking. In *International Conference on Machine Learning*, pages 2337–2346.

Mitchell A. Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the effects of weight pruning on transfer learning. *arXiv preprint arXiv:2002.08307*.

Saurabh Goyal, Anamitra Roy Choudhary, Venkatesan Chakaravarthy, Saurabh ManishRaje, Yogish Sabharwal, and Ashish Verma. 2020. Power-bert: Accelerating BERT inference for classification tasks. *arXiv preprint arXiv:2001.08950*.

Fu-Ming Guo, Sijia Liu, Finlay S. Mungall, Xue Lin, and Yanzhi Wang. 2019. Reweighted Proximal Pruning for Large-Scale Language Representation. *arXiv:1909.12486 [cs, stat]*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. *arXiv:2002.08909 [cs]*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and Understanding the 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 4143–4152, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1424

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the Knowledge in a Neural Network. In *Deep Learning and Representation Learning Workshop: NIPS 2014*.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2019. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformers Models. *arXiv:1910.05276 [cs]*. **DOI:** https://doi.org/10.18653/v1/2020.acl-demos.22

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. *arXiv:1902.00751 [cs, stat]*.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do attention heads in BERT track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556.

Ganesh Jawahar, Benoît Sagot, Djamé Seddah, Samuel Unicomb, Gerardo Iñiguez, Márton Karsai, Yannick Léo, Márton Karsai, Carlos Sarraute, Éric Fleury, et al. 2019. What does BERT learn about the structure of language? In *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*. **DOI:** https://doi.org/10.18653/v1/P19-1356

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019a. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. *arXiv preprint arXiv:1911.03437*. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.197, **PMID:** 33121726, **PMCID:** PMC7218724

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2019b. How Can We Know What Language Models Know? *arXiv:1911.12543 [cs]*. **DOI:** https://doi.org/10.1162/tacl_a_00324

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. TinyBERT: Distilling BERT for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. In *AAAI 2020*. **DOI:** https://doi.org/10.1609/aaai.v34i05.6311

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-Training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. **DOI:** https://doi.org/10.1162/tacl_a_00300

Wei-Tsung Kao, Tsung-Han Wu, Po-Han Chi, Chun-Cheng Hsieh, and Hung-Yi Lee. 2020. Further boosting BERT-based models by duplicating existing layers: Some intriguing

phenomena inside BERT. *arXiv preprint arXiv:2001.09309*.

Taeuk Kim, Jihun Choi, Daniel Edmiston, and Sang-goo Lee. 2020. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. In *ICLR 2020*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention Module is Not Only a Weight: Analyzing Transformers with Vector Norms. *arXiv:2004.10102 [cs]*.

Dan Kondratyuk and Milan Straka. 2019. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1279

Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. 2019. A mutual information maximization perspective of language representation learning. In *International Conference on Learning Representations*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets 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 4356–4365, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1445

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. 2020. Thieves on Sesame Street! Model Extraction of BERT-Based APIs. In *ICLR 2020*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data Augmentation using Pre-Trained Transformer Models. *arXiv:2003.02245 [cs]*.

Ilia Kuznetsov and Iryna Gurevych. 2020. A Matter of Framing: The Impact of Linguistic Formalism on Probing Results. *arXiv:2004.14999 [cs]*.

Guillaume Lample and Alexis Conneau. 2019. Cross-Lingual Language Model Pretraining. *arXiv:1901.07291 [cs]*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020a. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. In *ICLR*.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. 2019. Mixout: Effective regularization to finetune large-scale pretrained language models. *arXiv preprint arXiv:1909.11299*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]*. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.703

Changmao Li and Jinho D. Choi. 2020. Transformers to Learn Hierarchical Contexts in Multiparty Dialogue for Span-based Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5709–5714, Online. Association for Computational Linguistics.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joseph E. Gonzalez. 2020. Train large, then compress: Rethinking model size for efficient training and inference of transformers. *arXiv preprint arXiv:2002.11794*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting inside BERT's Linguistic Knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. 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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv: 1907.11692 [cs]*.

Xiaofei Ma, Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2019. Universal Text Representation from BERT: An Empirical Study. *arXiv:1910.07973 [cs]*.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, page 201907367. **DOI:** https://doi.org/10.1073/pnas.1907367117, **PMID:** 32493748

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

J. S. McCarley, Rishav Chakravarti, and Avirup Sil. 2020. Structured Pruning of a BERT-based Question Answering Model. *arXiv:1910.06360 [cs]*.

R. Thomas McCoy, Tal Linzen, Ewan Dunbar, and Paul Smolensky. 2019a. RNNs implicitly implement tensor-product representations. In *International Conference on Learning Representations*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019b. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1334

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and Non-Contextual Word Embeddings: An in-depth Linguistic Investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119. **DOI:** https://doi.org/10.18653/v1/2020.repl4nlp-1.15

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are Sixteen Heads Really Better than One? *Advances in Neural Information Processing Systems 32 (NIPS 2019)*.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemeter. 2019. What do you mean, BERT? assessing BERT as a distributional semantics model. *arXiv preprint arXiv:1911.05758*.

Microsoft. 2020. Turing-NLG: A 17-billion-parameter language model by microsoft.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1459

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and

Noah A. Smith. 2019a. Knowledge Enhanced Contextual Word Representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1005, **PMID:** 31383442

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019b. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W19-4302, **PMCID:** PMC6351953

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1250

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-Data Tasks. *arXiv:1811.01088 [cs]*.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-Theoretic Probing for Linguistic Structure. *arXiv:2004.03061 [cs]*. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.420

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. BERT is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised qa. *arXiv preprint arXiv:1911.03681*.

Sai Prasanna, Anna Rogers, and Anna Rumshisky. 2020. When BERT Plays the Lottery, All Tickets Are Winning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Online. Association for Computational Linguistics.

Ofir Press, Noah A. Smith, and Omer Levy. 2020. Improving Transformer Models by Reordering their Sublayers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2996–3005, Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.270

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.467

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]*.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. Fixed Encoder Self-Attention Patterns in Transformer-Based Machine Translation. *arXiv:2002.10260 [cs]*.

Alessandro Raganato and Jörg Tiedemann. 2018. An Analysis of Encoder Representations in Transformer-Based Machine Translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W18-5431

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912,

Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.442

Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2020. Probing Natural Language Inference Models through Semantic Fragments. In *AAAI 2020*. **DOI:** https://doi.org/10.1609/aaai.v34i05.6397

Kyle Richardson and Ashish Sabharwal. 2019. What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge. *arXiv:1912.13337 [cs]*. **DOI:** https://doi.org/10.1162/tacl_a_00331

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? *arXiv preprint arXiv:2002.08910*.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks. In *AAAI*, page 11. **DOI:** https://doi.org/10.1609/aaai.v34i05.6398

Rudolf Rosa and David Mareček. 2019. Inducing syntactic trees from BERT representations. *arXiv preprint arXiv:1906.11511*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*.

Victor Sanh, Thomas Wolf, and Alexander M. Rush. 2020. Movement Pruning: Adaptive Sparsity by Fine-Tuning. *arXiv:2005.07683 [cs]*.

Timo Schick and Hinrich Schütze. 2020. BERTRAM: Improved Word Embeddings Have Big Impact on Contextualized Model Performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3996–4007, Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.368

Florian Schmidt and Thomas Hofmann. 2020. BERT as a Teacher: Contextual Embeddings for Sequence-Level Reward. *arXiv preprint arXiv:2003.02738*.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green AI. *arXiv: 1907.10597 [cs, stat]*.

Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable? *arXiv:1906.03731 [cs]*. **DOI:** https://doi.org/10.18653/v1/P19-1282

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2019. Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT. *arXiv preprint arXiv:1909.05840*. **DOI:** https://doi.org/10.1609/aaai.v34i05.6409

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does BERT Learn from Multiple-Choice Reading Comprehension Datasets? *arXiv:1910.12391 [cs]*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. *arXiv:2004.09297 [cs]*.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning. In *International Conference on Machine Learning*, pages 5986–5995.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *ACL 2019*.

Ta-Chun Su and Hsiang-Chih Cheng. 2019. SesameBERT: Attention for Anywhere. *arXiv: 1910.03176 [cs]*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. In *AAAI*. **DOI:** https://doi.org/10.1609/aaai.v34i05.6422

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019a. Patient Knowledge Distillation for BERT Model Compression. 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 4314–4323. **DOI:** https://doi.org/10.18653/v1/D19-1441

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019b. ERNIE: Enhanced Representation through Knowledge Integration. *arXiv:1904.09223 [cs]*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019c. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. *arXiv:1907.12412 [cs]*. **DOI:** https://doi.org/10.1609/aaai.v34i05.6428

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: Task-Agnostic Compression of BERT for Resource Limited Devices.

Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, and Lawrence Carin. 2019. Syntax-Infused Transformer and BERT models for Machine Translation and Natural Language Understanding. *arXiv:1911.06156 [cs, stat]*. **DOI:** https://doi.org/10.1109/IALP48816.2019.9037672, **PMID:** 31938450, **PMCID:** PMC6959198

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. oLMpics – On what Language Model Pre-Training Captures. *arXiv:1912.13283 [cs]*.

Hirotaka Tanaka, Hiroyuki Shinnou, Rui Cao, Jing Bai, and Wen Ma. 2020. Document Classification by Word Embeddings of BERT. In *Computational Linguistics*, Communications in Computer and Information Science, pages 145–154, Singapore, Springer.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. *arXiv preprint arXiv:1903.12136*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. **DOI:** https://doi.org/10.18653/v1/P19-1452

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. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

James Yi Tian, Alexander P. Kreuzer, Pai-Hung Chen, and Hans-Martin Will. 2019. WaLDORf: Wasteless Language-model Distillation On Reading-comprehension. *arXiv preprint arXiv:1912.06638*.

Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. 2020. A Cross-Task Analysis of Text Span Representations. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 166–176, Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.repl4nlp-1.20

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and Practical BERT Models for Sequence Labeling. *arXiv preprint arXiv:1909.00100*. **DOI:** https://doi.org/10.18653/v1/D19-1374

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-Read Students Learn Better: The Impact of Student Initialization on Knowledge Distillation. *arXiv preprint arXiv:1908.08962*.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. 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 5831–5837, Hong Kong, China. Association for Computational Linguistics. **DOI:**

https://doi.org/10.18653/v1/D19-1592

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Jesse Vig. 2019. Visualizing Attention in Transformer-Based Language Representation Models. *arXiv:1904.02679 [cs, stat]*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the Structure of Attention in a Transformer Language Model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W19-4808

David Vilares, Michalina Strzyz, Anders Søgaard, and Carlos Gómez-Rodríguez. 2020. Parsing as pretraining. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*. **DOI:** https://doi.org/10.1609/aaai.v34i05.6446

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives. 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 4387–4397. **DOI:** https://doi.org/10.18653/v1/D19-1448

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019b. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. *arXiv preprint arXiv:1905.09418*. **DOI:** https://doi.org/10.18653/v1/P19-1580

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. *arXiv:2003.12298 [cs]*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal Adversarial Triggers for Attacking and Analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1221

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019b. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. *arXiv preprint arXiv:1909.07940*. **DOI:** https://doi.org/10.18653/v1/D19-1534

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W18-5446

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2020a. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. *arXiv:2002.01808 [cs]*.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Liwei Peng, and Luo Si. 2019a. StructBERT: Incorporating Language Structures into Pre-Training for Deep Language Understanding. *arXiv:1908.04577 [cs]*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *arXiv preprint arXiv:2002.10957*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2020c. KEPLER: A Unified Model for Knowledge

Embedding and Pre-trained Language Representation. *arXiv:1911.06136 [cs]*.

Yile Wang, Leyang Cui, and Yue Zhang. 2020d. How Can BERT Help Lexical Semantics Tasks? *arXiv:1911.02929 [cs]*.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019b. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. *arXiv preprint arXiv:1912.07840*.

Alex Warstadt and Samuel R. Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? In *Proceedings of the 42nd Annual Virtual Meeting of the Cognitive Science Society*. Online.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. 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 2870–2880. **DOI:** https://doi.org/10.18653/v1/D19-1286

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. *arXiv preprint arXiv:1909.10430*.

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. **DOI:** https://doi.org/10.18653/v1/D19-1002

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. HuggingFace's Transformers: State-of-the-Art Natural Language Processing. *arXiv:1910.03771 [cs]*.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019a. Pay Less Attention with Lightweight and Dynamic Convolutions. In *International Conference on Learning Representations*.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019b. Conditional BERT Contextual Augmentation. In *ICCS 2019: Computational Science ICCS 2019*, pages 84–95. Springer. **DOI:** https://doi.org/10.1007/978-3-030-22747-0_7

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. 2020. BERT-of-Theseus: Compressing BERT by Progressive Module Replacing. *arXiv preprint arXiv:2002.02925*.

Junjie Yang and Hai Zhao. 2019. Deepening Hidden Representations from Pre-Trained Language Models for Natural Language Understanding. *arXiv:1911.01940 [cs]*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki

Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and Evaluating General Linguistic Intelligence. *arXiv: 1901.11373 [cs, stat]*.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, and Cho-Jui Hsieh. 2019. Large Batch Optimization for Deep Learning: Training BERT in 76 Minutes. *arXiv preprint arXiv:1904.00962*, 1(5).

Ali Hadi Zadeh and Andreas Moshovos. 2020. GOBO: Quantizing Attention-Based NLP Models for Low Latency and Energy Efficient Inference. *arXiv:2005.03842 [cs, stat]*. **DOI:** https://doi.org/10.1109/MICRO50266.2020.00071

Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8BERT: Quantized 8bit BERT. *arXiv preprint arXiv:1910.06188*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1139

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for Language Understanding. In *AAAI 2020*.

Sanqiang Zhao, Raghav Gupta, Yang Song, and Denny Zhou. 2019. Extreme Language Model Compression with Optimal Subwords and Shared Projections. *arXiv preprint arXiv:1909.11687*.

Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.429, **PMCID:** PMC7660194

Wenxuan Zhou, Junyi Du, and Xiang Ren. 2019. Improving BERT Fine-tuning with Embedding Normalization. *arXiv preprint arXiv:1911.03918*.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating Commonsense in Pre-Trained Language Models. In *AAAI 2020*. **DOI:** https://doi.org/10.1609/aaai.v34i05.6523

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. FreeLB: Enhanced Adversarial Training for Language Understanding. *arXiv:1909.11764 [cs]*.