A Primer in BERTology: What we know about how BERT works

Anna Rogers, Olga Kovaleva, Anna Rumshisky
Department of Computer Science, University of Massachusetts Lowell
Lowell, MA 01854
{arogers, okovalev, arum}@cs.uml.edu

Abstract

Transformer-based models are now widely used in NLP, but we still do not understand a lot about their inner workings. This paper describes what is known to date about the famous BERT model (Devlin et al., 2019), synthesizing over 40 analysis studies. We also provide an overview of the proposed modifications to the model and its training regime. We then outline the directions for further research.

1 Introduction

Since their introduction in 2017, Transformers (Vaswani et al., 2017) took 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) which obtained state-of-the-art results in numerous benchmarks, and was integrated in Google search[1], improving an estimated 10% of queries.

While it is clear that BERT and other Transformer-based models work 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.

This paper provides an overview of what has been learned to date, highlighting the questions which are still unresolved. We focus on the studies investigating the types of knowledge learned by BERT, where this knowledge is represented, how it is learned, and the methods proposed to improve it.

2 Overview of BERT architecture

Fundamentally, BERT is a stack of Transformer encoder layers (Vaswani et al., 2017) which consist of multiple "heads", i.e., fully-connected neural networks augmented with a self-attention mechanism. For every input token in a sequence, each head computes key, value and query vectors, which are 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 layer normalization is applied after it.

The conventional workflow for BERT consists of two stages: pre-training and fine-tuning. Pre-training uses two semi-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 fol-

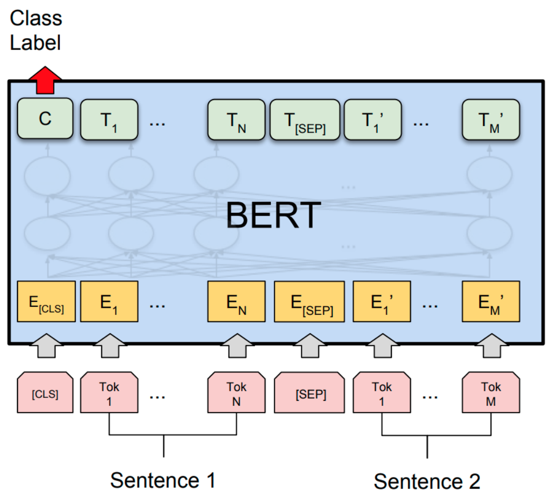

Figure 1: BERT fine-tuning (Devlin et al., 2019).

[1] https://blog.google/products/search/search-language-understanding-bert

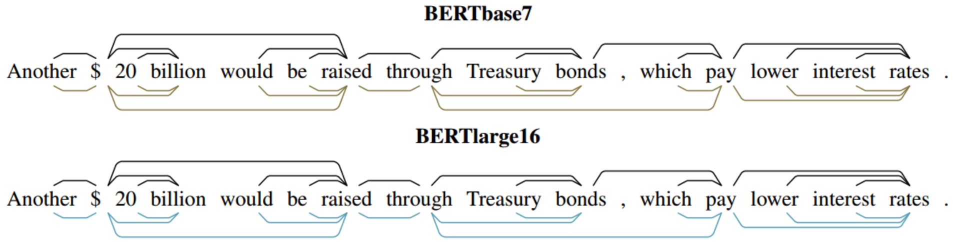

Figure 2: Parse trees recovered from BERT representations by Hewitt et al. (2019)

lows: BERT first tokenizes the given word into wordpieces (Wu et al., 2016b), and then combines three embedding layers (token, position, and segment) to obtain a fixed-length vector. Special token `[CLS]` is used for classification predictions, and `[SEP]` separates input segments.The original BERT comes in two versions: base and large, varying in the number of layers, their hidden size, and number of attention heads.

3 BERT embeddings

Unlike the conventional static embeddings (Mikolov et al., 2013a; Pennington et al., 2014), BERT's representations are contextualized, i.e., every input token is represented by a vector dependent on the particular context of occurrence. In the current studies of BERT's representation space, the term 'embedding' refers to the output vector of a given (typically final) Transformer layer.

Wiedemann et al. (2019) find that BERT's contextualized embeddings form distinct and clear clusters corresponding to word senses, which confirms that the basic distributional hypothesis holds for these representations. However, Mickus et al. (2019) note that representations of the same word varies depending on position of the sentence in which it occurs, likely due to NSP objective.

Ethayarajh (2019) measure how similar the embeddings for identical words are in every layer and find that later BERT layers produce more context-specific representations. They also find that BERT embeddings occupy a narrow cone in the vector space, and this effect increases from lower to higher layers. That is, two random words will on average have a much higher cosine similarity than expected if embeddings were directionally uniform (isotropic).

4 What knowledge does BERT have?

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

4.1 Syntactic knowledge

Lin et al. (2019) showed that **BERT representations are hierarchical rather than linear**, i.e. 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**. However, BERT's knowledge of syntax is partial, since probing classifiers could not recover the labels of distant parent nodes in the syntactic tree (Liu et al., 2019a).

As far as how syntactic information is represented, it seems that **syntactic structure is not directly encoded in self-attention weights, but they can be transformed to reflect it**. 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, Hewitt and Manning (2019) were able to learn transformation matrices that would successfully recover much of the Stanford Dependencies formalism for PennTreebank data (see Figure 2). Jawahar et al. (2019) try to approximate BERT representations with Tensor Product Decomposition Networks (McCoy et al., 2019a), concluding that the dependency trees are the best match among 5 decomposition schemes (although the reported MSE differences are very small).

Regarding syntactic competence of BERT's

MLM, Goldberg (2019) showed that **BERT takes subject-predicate agreement into account when performing the cloze task.** This was the case even for sentences with distractor clauses between the subject and the verb, and meaningless sentences. 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 evidence of syntactic knowledge is belied by the fact that **BERT does not "understand" negation and is insensitive to malformed input**. In particular, its predictions were not altered even with shuffled word order, truncated sentences, removed subjects and objects (Ettinger, 2019). This is in line with the recent findings on adversarial attacks, with models disturbed by nonsensical inputs (Wallace et al., 2019a), and suggests that **BERT's encoding of syntactic structure does not indicate that it actually relies on that knowledge**.

4.2 Semantic knowledge

To date, more studies were 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 for semantic roles** (Ettinger, 2019). BERT is even able to prefer the incorrect fillers for semantic roles that are semantically related to the correct ones, to those that are unrelated (e.g. "to tip a chef" should be 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.

4.3 World knowledge

MLM component of BERT is easy to adapt for knowledge induction by filling in the blanks (e.g. "Cats like to chase [___]"). There is at least one probing study of world knowledge in BERT (Ettinger, 2019), but the bulk of evidence comes from

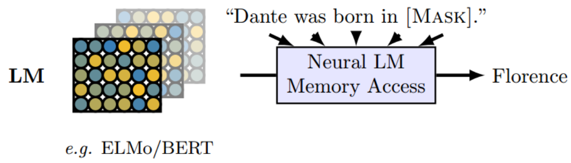

Figure 3: BERT's world knowledge (Petroni et al., 2019)

numerous practitioners using BERT to extract such knowledge.

Petroni et al. (2019) showed that, **for some relation types, vanilla BERT is competitive with methods relying on knowledge bases** (Figure 3). Davison et al. (2019) suggest that it generalizes better to unseen data. However, 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.)

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 does not have the information about their interactions (e.g. 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. At the same time, Poerner et al. (2019) show that some of BERT's success in factoid knowledge retrieval comes from learning stereotypical character combinations, e.g. it would predict that a person with an Italian-sounding name is Italian, even when it is factually incorrect.

5 Localizing linguistic knowledge

5.1 Self-attention heads

Attention is widely considered to be useful for understanding Transformer models, and several studies proposed classification of attention head types:

- attending to the word itself, to previous/next words and to the end of the sentence (Raganato and Tiedemann, 2018);
- attending to previous/next tokens, `[CLS]`, `[SEP]`, punctuation, and "attending broadly" over the sequence (Clark et al., 2019);
- the 5 attention types shown in Figure 4 (Kovaleva et al., 2019).

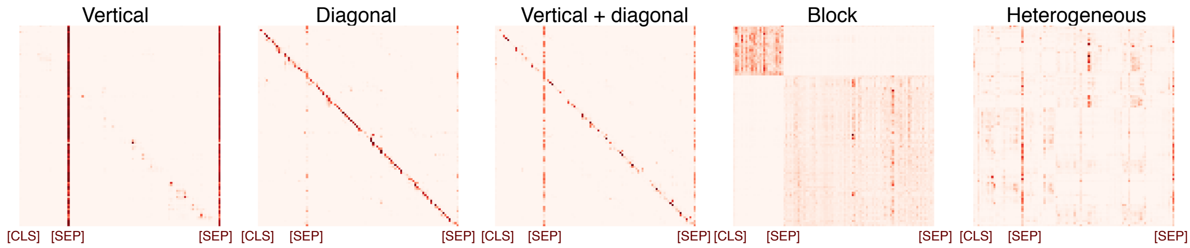

Figure 4: Attention patterns in BERT (Kovaleva et al., 2019)

According to Clark et al. (2019), "attention weight has a clear meaning: how much a particular word will be weighted when computing the next representation for the current word". However, Kovaleva et al. (2019) showed that **most self-attention heads do not directly encode any non-trivial linguistic information**, since less than half of them had the "heterogeneous" pattern[2]. Much of the model encoded the vertical pattern (attention to `[CLS]`, `[SEP]`, and punctuation tokens), consistent with the observations by Clark et al. (2019). This apparent redundancy must be related to the overparametrization issue (see section 7).

Attention to `[CLS]` is easy to interpret as attention to an aggregated sentence-level representation, but BERT also attends a lot to `[SEP]` and punctuation. Clark et al. (2019) hypothesize that periods and commas are simply almost as frequent as `[CLS]` and `[SEP]`, and the model learns to rely on them. They suggest also that the function of `[SEP]` might be one of "no-op", a signal to ignore the head if its pattern is not applicable to the current case. `[SEP]` gets increased attention starting in layer 5, but its importance for prediction drops. If this hypothesis is correct, attention probing studies that excluded the `[SEP]` and `[CLS]` tokens (as e.g. Lin et al. (2019) and Htut et al. (2019)) should perhaps be revisited.

Proceeding to the analysis of the "heterogeneous" self-attention pattern, a number of studies looked for specific BERT heads with linguistically interpretable functions.

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` has some variation between these two studies. The overall conclusion is also supported by Voita et al. (2019)'s data for 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.

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 (see subsection 4.1).

Lin et al. (2019) present evidence that **attention weights are weak indicators of subject-verb agreement and reflexive anafora.** 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.

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,.

Kovaleva et al. (2019) showed that **even when attention heads specialize in tracking semantic relations, they do not necessarily contribute to BERT's performance on relevant tasks**. Kovaleva et al. (2019) identified two heads of base BERT, in which self-attention maps were closely aligned with annotations of core frame semantic relations (Baker et al., 1998). Although such relations should have been instrumental to tasks such as inference, a head ablation study showed that these heads were not essential for BERT's success on GLUE tasks.

5.2 BERT layers

The first layer of BERT receives as input representations that are a combination of token, segment, and positional embeddings. It stands to reason that **the**

[2]The experiments were conducted with BERT fine-tuned on GLUE tasks (Wang et al., 2018).

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

There is a wide consensus among studies with different tasks, datasets and methodologies that **syntactic information is the most prominent in the middle BERT[3] layers.** 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) report the best subject-verb agreement around layers 8-9, and the performance on syntactic probing tasks used by Jawahar et al. (2019) also seemed to peak around the middle of the model.

The prominence of syntactic information in the middle BERT layers must be related to Liu et al. (2019a) observation that the middle layers of Transformers are overall the best-performing and the most transferable across tasks (see Figure 5).

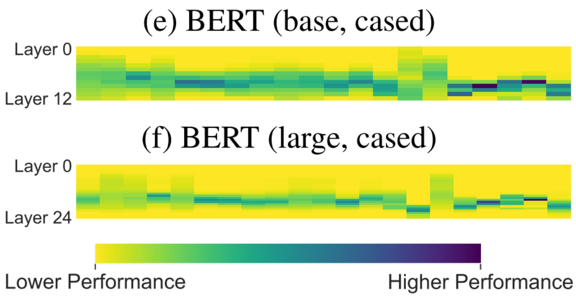

Figure 5: BERT layer transferability (columns correspond to probing tasks) (Liu et al., 2019a).

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 appears 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 also performed best at the middle layers, in both BERT-base and BERT-large.

The final layers of BERT are the most task-specific. In pre-training, this means specificity to the MLM task, which would explain 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). At the same time, Hao et al. (2019) report that if the weights of lower layers of the fine-tuned BERT are restored to their original values, it does not dramatically hurt the model performance.

Tenney et al. (2019a) suggest that while most of syntactic information can be localized in a few layers, **semantics is spread across the entire model**, which would explain why certain non-trivial examples get solved incorrectly at first but correctly at higher 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 actually achieves in BERT in terms of the spread of semantic knowledge, and whether that is beneficial. The authors' comparison between base and large BERTs shows that the overall pattern of cumulative score gains is the same, only more spread out in the large BERT.

The above view is disputed by Jawahar et al. (2019), who place "surface features in lower layers, syntactic features in middle layers and semantic features in higher layers". However, the conclusion with regards to the semantic features seems surprising, given that only one SentEval 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.

6 Training BERT

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

6.1 Pre-training BERT

The original BERT is a bidirectional Transformer pre-trained on two tasks: next sentence prediction (NSP) and masked language model (MLM). Multiple studies have come up with **alternative training objectives** to improve on BERT.

- *Removing NSP* does not hurt or slightly improves task performance (Liu et al., 2019b;

[3]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.

Joshi et al., 2020; Clinchant et al., 2019), especially in cross-lingual setting (Wang et al., 2019b). Wang et al. (2019a) replace NSP with the task of predicting both the next and the previous sentences. Lan et al. (2020) replace the negative NSP examples by the swapped sentences from positive examples, rather than sentences from different documents.

- *Dynamic masking* (Liu et al., 2019b) improves on BERT's MLM by using diverse masks for training examples within an epoch;
- *Beyond-sentence MLM*. Lample and Conneau (2019) replace sentence pairs with arbitrary text streams, and subsample frequent outputs similarly to Mikolov et al. (2013b).
- *Permutation language modeling*. Yang et al. (2019) replace MLM with training on different permutations of word order in the input sequence, maximizing the probability of the original word order. See also the n-gram word order reconstruction task (Wang et al., 2019a).
- *Span boundary objective* aims to predict a masked span (rather than single words) using only the representations of the tokens at the span's boundary (Joshi et al., 2020);
- *Phrase masking* and *named entity masking* (Zhang et al., 2019) aim to improve representation of structured knowledge by masking entities rather than individual words;
- *Continual learning* is sequential pre-training on a large number of tasks[4], each with their own loss which are then combined to continually update the model (Sun et al., 2019b).
- *Conditional MLM* by Wu et al. (2019b) replaces the segmentation embeddings with "label embeddings", which also include the label for a given sentence from an annotated task dataset (e.g. sentiment analysis).
- Clinchant et al. (2019) propose replacing the MASK token with `[UNK]` token, as this could help the model to learn certain representation for unknowns that could be exploited by a neural machine translation model.

Another obvious source of improvement is pre-training data. Liu et al. (2019c) explore the benefits of increasing the corpus volume and longer training. The data also does not have to be unstructured text: although BERT is actively used as a source of world knowledge (subsection 4.3), there are ongoing efforts to incorporate structured knowledge resources (Peters et al., 2019a).

Another way to integrate external knowledge is use entity embeddings as input, as in E-BERT (Poerner et al., 2019) and ERNIE (Zhang et al., 2019). Alternatively, SemBERT (Zhang et al., 2020) integrates semantic role information with BERT representations.

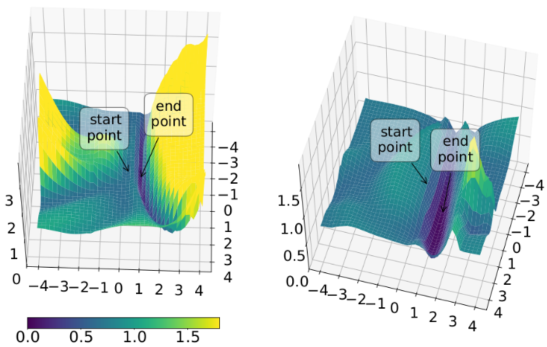

Figure 6: Pre-trained weights help BERT find wider optima in fine-tuning on MRPC (right) than training from scratch (left) (Hao et al., 2019)

Pre-training is the most expensive part of training BERT, and it would be informative to know how much benefit it provides. Hao et al. (2019) conclude 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 6). However, 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).

6.2 Model architecture choices

To date, the most systematic study of BERT architecture was performed by Wang et al. (2019b). They 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, which is consistent with the findings of Voita et al. (2019) and Michel et al. (2019), discussed in section 7, and also the observation by Liu et al. (2019a) that middle layers were the most transferable. Larger hidden representation size was consistently better, but the gains varied by setting.

[4] New token-level tasks in ERNIE include prediction whether a token is capitalized and whether it occurs in other segments of the same document. Segment-level tasks include sentence reordering, sentence distance prediction, and supervised discourse relation classification.

Liu et al. (2019c) show that large-batch training (8k examples) improves both the language model perplexity and downstream task performance. They also publish their recommendations for other model 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 embedding values of the trained [CLS] token are not centered around zero, their normalization stabilizes the training leading to a slight performance gain on text classification tasks.

Gong et al. (2019) note that, since self-attention patterns in higher layers resemble the ones in lower layers, 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 "warm-start" can lead to a 25% faster training speed while reaching similar accuracy to the original BERT on GLUE tasks.

6.3 Fine-tuning BERT

Pre-training + fine-tuning workflow is a crucial part of BERT. The former is supposed to provide task-independent linguistic knowledge, and the fine-tuning process would presumably teach the model to rely on the representations that are more 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 3 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*. Yang and Zhao (2019) learn a complementary representation of the information in the deeper layers that is combined with the output layer. Su and Cheng (2019) propose using a weighted representation of all layers instead of the final layer output.

- *Two-stage fine-tuning* introduces an intermediate supervised training stage between pre-training and fine-tuning (Phang et al., 2019; Garg et al., 2020).

- *Adversarial token perturbations* improve robustness of the model (Zhu et al., 2019).

With larger and larger models even fine-tuning becomes expensive, but Houlsby et al. (2019) show that it can be successfully approximated by inserting adapter modules. They adapt BERT to 26 classification tasks, achieving competitive performance at a fraction of the computational cost. Artetxe et al. (2019) also find adapters helpful in reusing monolingual BERT weights for cross-lingual transfer.

An alternative to fine-tuning is extracting features from frozen representations, but fine-tuning works better for BERT (Peters et al., 2019b).

Initialization can have a dramatic effect on the training process (Petrov, 2010). However, variation across initializations is not often reported, although the performance improvements claimed in many NLP modeling papers may be within the range of that variation (Crane, 2018). Dodge et al. (2020) report significant variation for BERT fine-tuned on GLUE tasks, where both weight initialization and training data order contribute to the variation. They also propose an early-stopping technique to avoid full fine-tuning for the less-promising seeds.

7 How big should BERT be?

7.1 Overparametrization

Transformer-based models keep increasing in size: e.g. T5 (Wu et al., 2016a) is over 30 times larger than the base BERT. This raises concerns about computational complexity of self-attention (Wu et al., 2019a), environmental issues (Strubell et al., 2019; Schwartz et al., 2019), as well as reproducibility and access to research resources in academia vs. industry.

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. (2019) showed that all but a few Transformer heads could be pruned without 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

[5]See also experiments with multilingual BERT by (Singh et al., 2019), where fine-tuning affected the top and the middle layers of the model.

		Compression	Performance	Speedup	Model	Evaluation
Distillation	DistilBERT (Sanh et al., 2019)	$\times$2.5	90%	$\times$1.6	BERT_6	All GLUE tasks
	BERT_6-PKD (Sun et al., 2019a)	$\times$1.6	97%	$\times$1.9	BERT_6	No WNLI, CoLA and STS-B
	BERT_3-PKD (Sun et al., 2019a)	$\times$2.4	92%	$\times$3.7	BERT_3	No WNLI, CoLA and STS-B
	(Aguilar et al., 2019)	$\times$2	94%	-	BERT_6	CoLA, MRPC, QQP, RTE
	BERT-48 (Zhao et al., 2019)	$\times$62	87%	$\times$77	$\mathrm{BERT}_{12}{}^{*\dagger}$	MNLI, MRPC, SST-2
	BERT-192 (Zhao et al., 2019)	$\times$5.7	94%	$\times$22	$\mathrm{BERT}_{12}{}^{*\dagger}$	MNLI, MRPC, SST-2
	TinyBERT (Jiao et al., 2019)	$\times$7.5	96%	$\times$9.4	$\mathrm{BERT}_4{}^{*\dagger}$	All GLUE tasks
	MobileBERT (Sun et al.)	$\times$4.3	100%	$\times$4	$\mathrm{BERT}_{24}{}^{\dagger}$	No WNLI
	PD (Turc et al., 2019)	$\times$1.6	98%	$\times$2.5^3	$\mathrm{BERT}_6{}^{\dagger}$	No WNLI, CoLA and STS-B
	MiniBERT(Tsai et al., 2019)	$\times$6§	98%	$\times$27§	$\mathrm{mBERT}_3{}^{\dagger}$	CoNLL-2018 POS and morphology
	BiLSTM soft (Tang et al., 2019)	$\times$110	91%	$\times$434‡	BiLSTM_1	MNLI, QQP, SST-2
Quant.	Q-BERT (Shen et al., 2019)	$\times$13	99%	-	BERT_{12}	MNLI, SST-2
	Q8BERT (Zafrir et al., 2019)	$\times$4	99%	-	BERT_{12}	All GLUE tasks
Other	ALBERT-base (Lan et al., 2019)	$\times$9	97%	$\times$5.6	$\mathrm{BERT}_{12}{}^{**}$	MNLI, SST-2
	ALBERT-xxlarge (Lan et al., 2019)	$\times$0.47	107%	$\times$0.3	$\mathrm{BERT}_{12}{}^{**}$	MNLI, SST-2
	BERT-of-Theseus (Xu et al., 2020)	$\times$1.6	98%	-	BERT_6	No WNLI

Table 1: Comparison of BERT compression studies. Compression, performance retention, and inference time speedup figures are given with respect to $\mathrm{BERT}_{\mathrm{base}}$, unless indicated otherwise. Performance retention is measured as a ratio of average scores achieved by a given model and by $\mathrm{BERT}_{\mathrm{base}}$. The subscript in the model description reflects the number of layers used. *Smaller vocabulary used. †The dimensionality of the hidden layers is reduced. **The dimensionality of the embedding layer is reduced. ‡Compared to $\mathrm{BERT}_{\mathrm{large}}$. §Compared to mBERT.

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 useless, but also harmful to the downstream task performance. Positive effects from disabling heads were reported for machine translation (Michel et al., 2019), and for GLUE tasks, both heads and layers could be disabled (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).

Many experiments comparing BERT-base and BERT-large saw the larger model perform better Liu et al. (2019a), but that is not always the case. In particular, the opposite was observed for 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 of the possible reasons is the use of attention dropouts, which causes some attention weights to be zeroed-out during training.

7.2 BERT compression

Given the above evidence of overparametrization, it does not come as a surprise that BERT can be efficiently compressed with minimal accuracy loss. Such efforts to date are summarized in Table 1. Two main approaches include *knowledge distillation* and *quantization*.

The studies in the knowledge distillation framework (Hinton et al., 2015) use a smaller student-network that is trained to mimic the behavior of a larger teacher-network (BERT-large or BERT-base). This is 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 different stages at the pre-training (Turc et al., 2019; Jiao et al., 2019; Sun et al.) or at the fine-tuning stage (Jiao et al., 2019)).

The quantization approach aims to decrease 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.

Other techniques include decomposing BERT's embedding matrix into smaller matrices (Lan et al., 2019) and progressive model replacing (Xu et al., 2020).

8 Multilingual BERT

Multilingual BERT (mBERT[6]) is a version of BERT that was trained on Wikipedia in 104 languages (110K wordpiece vocabulary). Languages with a lot of data were subsampled, and some were super-sampled using exponential smoothing.

[6] https://github.com/google-research/bert/blob/master/multilingual.md

mBERT performs surprisingly well in zero-shot transfer on many tasks (Wu and Dredze, 2019; Pires et al., 2019), although not in language generation (Rönnqvist et al., 2019). The model seems to naturally learn high-quality cross-lingual word alignments (Libovický et al., 2019), with caveats for open-class parts of speech (Cao et al., 2019). Adding more languages does not seem to harm the quality of representations (Artetxe et al., 2019).

mBERT generalizes across some scripts (Pires et al., 2019), and can retrieve parallel sentences, although Libovický et al. (2019) note that this task could be solvable by simple lexical matches. Pires et al. (2019) conclude that mBERT representation space shows some systematicity in between-language mappings, which makes it possible in some cases to "translate" between languages by shifting the representations by the average parallel sentences offset for a given language pair.

mBERT is simply trained on a multilingual corpus, with no language IDs, but it encodes language identities (Wu and Dredze, 2019; Libovický et al., 2019), and adding the IDs in pre-training was not beneficial (Wang et al., 2019b). It is also aware of at least some typological language features (Libovický et al., 2019; Singh et al., 2019), and transfer between structurally similar languages works better (Wang et al., 2019b; Pires et al., 2019).

Singh et al. (2019) argue that if typological features structure its representation space, it could not be considered as interlingua. However, Artetxe et al. (2019) show that cross-lingual transfer can be achieved by only retraining the input embeddings while keeping monolingual BERT weights, which suggests that even monolingual models learn generalizable linguistic abstractions.

At least some of the syntactic properties of English BERT hold for mBERT: its MLM is aware of 4 types of agreement in 26 languages (Bacon and Regier, 2019), and main auxiliary of the sentence can be detected in German and Nordic languages Rönnqvist et al. (2019).

Pires et al. (2019) and Wu and Dredze (2019) hypothesize that shared word-pieces help mBERT, based on experiments where the task performance correlated with the amount of shared vocabulary between languages. However, Wang et al. (2019b) dispute this account, showing that bilingual BERT models are not hampered by the lack of shared vocabulary. Artetxe et al. (2019) also show cross-lingual transfer is possible by swapping the model vocabulary, without any shared word-pieces.

Figure 7: Language centroids of the mean-pooled mBERT representations (Libovický et al., 2019)

To date, the following proposals were made for improving mBERT:

- fine-tuning on multilingual datasets is improved by freezing the bottom layers (Wu and Dredze, 2019);
- improving word alignment in fine-tuning (Cao et al., 2019);
- translation language modeling (Lample and Conneau, 2019) is an alternative pre-training objective where words are masked in parallel sentence pairs (the model can attend to one or both sentences to solve the prediction task);
- Huang et al. (2019) combine 5 pre-training tasks (monolingual and cross-lingual MLM, translation language modeling, cross-lingual word recovery and paraphrase classification);

A fruitful research direction is using monolingual BERT directly in cross-lingual setting. Clinchant et al. (2019) experiment with initializing the encoder part of the neural MT model with monolingual BERT weights. Artetxe et al. (2019) and Tran (2019) independently showed that mBERT does not have to be pre-trained on multiple languages: it is possible to freeze the Transformer weights and retrain only the input embeddings.

9 Discussion

9.1 Limitations

As shown in section 4, multiple probing studies report that BERT possesses a surprising amount of syntactic, semantic, and world knowledge. However, as Tenney et al. (2019a) aptly stated, "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 tradeoff between the complexity of the probe and the tested hypothesis (Liu et al., 2019a). A more complex probe might be able to recover more information, but it becomes less clear whether we are still talking about the original model.

Furthermore, different probing methods may reveal complementary or even contradictory information, in which case a single test (as done in most studies) would not be sufficient (Warstadt et al., 2019). Certain methods might also favor a certain model, e.g., RoBERTa is trailing BERT with one tree extraction method, but leading with another (Htut et al., 2019).

Head and layer ablation studies (Michel et al., 2019; Kovaleva et al., 2019) inherently assume that certain knowledge is contained in heads/layers, but there is evidence of more diffuse representations spread across the full network: the gradual increase in accuracy on difficult semantic parsing tasks (Tenney et al., 2019a), the absence of heads that would perform parsing "in general" (Clark et al., 2019; Htut et al., 2019). Ablations are also problematic if the same information was duplicated elsewhere in the network. To mitigate that, Michel et al. (2019) prune heads in the order set by a proxy importance score, and Voita et al. (2019) fine-tune the pre-trained Transformer with a regularized objective that has the head-disabling effect.

Many papers are accompanied by attention visualizations, with a growing number of visualization tools (Vig, 2019; Hoover et al., 2019). However, there is ongoing debate on the merits of attention as a tool for interpreting deep learning models (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Brunner et al., 2020). Also, visualization is typically limited to qualitative analysis (Belinkov and Glass, 2019), and should not be interpreted as definitive evidence.

9.2 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, together with the starting points that we already have.

Benchmarks that require verbal reasoning. While BERT enabled breakthroughs on many NLP benchmarks, a growing list of analysis papers are showing that its verbal reasoning abilities are not as impressive as it seems. In particular, it was shown to rely on shallow heuristics in both natural language inference (McCoy et al., 2019b; Zellers et al., 2019) and reading comprehension (Si et al., 2019; Rogers et al., 2020; Sugawara et al., 2020). As with any optimization method, if there is a shortcut in the task, we have no reason to expect that BERT will not learn it. To overcome this, the NLP community needs to incentivize dataset development on par with modeling work, which at present is often perceived as more prestigious.

Developing methods to "teach" reasoning. While the community had success extracting knowledge from large pre-trained models, they often fail if any reasoning needs to be performed on top of the facts they possess (see subsection 4.3). For instance, Richardson et al. (2019) propose a method to "teach" BERT quantification, conditionals, comparatives, and boolean coordination.

Learning what happens at inference time. Most of the BERT analysis papers focused on different probes of the model, but we know much less about what knowledge actually gets used. At the moment, we know that the knowledge represented in BERT does not necessarily get used in downstream tasks (Kovaleva et al., 2019). As starting points for work in this direction, we also have other head ablation studies (Voita et al., 2019; Michel et al., 2019) and studies of how BERT behaves in reading comprehension task (van Aken et al., 2019; Arkhangelskaia and Dutta, 2019).

10 Conclusion

In a little over a year, BERT has become a ubiquitous baseline in NLP engineering 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 will help the community to focus on the biggest unresolved questions.

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*.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does BERT answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1823–1832.

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:1910.11856 [cs]*.

Geoff Bacon and Terry Regier. 2019. Does BERT agree? evaluating knowledge of structure dependence through agreement relations. *arXiv preprint arXiv:1908.09892*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley Framenet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90. Association for Computational Linguistics.

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

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2019. Inducing Relational Knowledge from BERT. *arXiv:1911.12753 [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*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual Alignment of Contextual Word Representations. In *International Conference on Learning Representations*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does BERT look at? An analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*.

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.

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.

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.

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]*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Allyson Ettinger. 2019. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *arXiv:1907.13528 [cs]*.

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.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection. In *AAAI*.

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.

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 4134–4143.

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. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2019. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformers Models. *arXiv:1910.05276 [cs]*.

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*.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

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*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How Can We Know What Language Models Know?

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*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans.

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.

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. 2019. Albert: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How Language-Neutral is Multilingual BERT? *arXiv:1911.03310 [cs]*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside bert's linguistic knowledge. *arXiv preprint arXiv:1906.01698*.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

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]*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019b. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *arXiv preprint arXiv:1905.10650*.

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*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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.

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.

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.

Slav Petrov. 2010. Products of Random Latent Variable Grammars. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Los Angeles, California. Association for Computational Linguistics.

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

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

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*.

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.

Kyle Richardson, Hai Hu, Lawrence S Moss, and Ashish Sabharwal. 2019. Probing natural language inference models through semantic fragments. *arXiv preprint arXiv:1909.07521*.

Kyle Richardson and Ashish Sabharwal. 2019. What Does My QA Model Know? Devising Controlled Probes using Expert Knowledge. *arXiv:1912.13337 [cs]*.

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.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is Multilingual BERT Fluent in Language Generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

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

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]*.

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*.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does BERT learn from multiple-choice reading comprehension datasets? *arXiv preprint arXiv:1910.12391*.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is Not an Interlingua and the Bias of Tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

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*.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019a. Patient knowledge distillation for BERT model compression. *arXiv preprint arXiv:1908.09355*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2019b. ERNIE 2.0: A Continual Pre-training Framework for Language Understanding. *arXiv:1907.12412 [cs]*.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: Task-agnostic compression of bert for resource limited devices.

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. *arXiv preprint arXiv:1905.05950*.

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*.

Ke Tran. 2019. From English to Foreign Languages: Transferring Pre-trained Language Models.

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*.

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*.

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.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

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.

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*.

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.

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]*.

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, 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. *arXiv preprint arXiv:1909.02597*.

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.

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*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016a. Google's neural machine translation system: Bridging the gap between human and machine translation. abs/1609.08144.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016b. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

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]*.

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).

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? *arXiv preprint arXiv:1905.07830*.

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.

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*.

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 Pretrained Language Models. In *AAAI 2020*.

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]*.