Trade-offs in Learning Commonsense vs Factual Knowledge with Pretrained Language Models

Makesh Sreedhar
UW Madison
msreedhar@wisc.edu

Satya Sai Srinath
UW Madison
sgnamburi@wisc.edu

Abstract

Multiple probing techniques have been developed in order to understand how large transformer language models acquire semantic and syntactic knowledge during training. In this paper, we focus on the extent to which pretrained models capture factual and commonsense world knowledge. Differently from existing probing work that mostly analyze BERT-like models, we also analyze ELECTRA, which employs a different pretraining objective. Interestingly, our results reveal that ELECTRA acquires better commonsense and poorer factual knowledge compared with MLM models. By probing BERT at different pretraining steps, we discover that there seems to exist a tradeoff between storing explicit facts in memory and the ability of extracting more general, commonsense information from text.

1 Introduction

A plethora of recent works use *probing* to get insights on what type of linguistic knowledge large pretrained language models (PLMs) capture. These models can encode substantial amounts of syntax and semantics (Tenney et al., 2019a,c; Hewitt and Manning, 2019; Hu et al., 2020; Shwartz and Dagan, 2019). Moreover, recent evidence shows that these models can act as knowledge bases and store *factual* knowledge such as "*Dante was born in Florence*" or that "*The capital of France is Paris*" (Petroni et al., 2019; Heinzerling and Inui, 2021; Soares et al., 2019; Roberts et al., 2020). This evidence is obtained by measuring the ability of PLMs to *fill-in-the-blanks* verbalized versions of existing knowledge base triplets.

On the other hand, the ability of these models to acquire *commonsense* knowledge such as "*birds have feathers*" or "*rain makes the road slippery*" from text alone has been shown to happen only to some extent and appeared until now as a considerably more challenging task (Forbes et al., 2019; Hwang et al., 2020; Porada et al., 2019). The models can capture *IsA* relationships, which have a higher likelihood to be verbalized in text corpora, but struggle on more complex ones (Hwang et al., 2020). Recent investigations show that PLMs cannot judge the *plausibility* of different events, e.g. whether "*chef-bake-cookie*" is more plausible than "*fish-throw-elephant*", a particular type of commonsense reasoning (Porada et al., 2019).

The difficulty in capturing different aspects of commonsense knowledge derives from the fact that, contrary to factual knowledge, the former is implicit in the textual data, even in very large corpora, and therefore it must be inferred (Talmor et al., 2019; Zhang et al., 2020a; Forbes et al., 2019). Recent approaches have investigated the use of external knowledge bases (KBs) such as ATOMIC (Sap et al., 2019) or ConceptNet (Liu and Singh, 2004) to endow current PLMs with commonsense reasoning (Porada et al., 2019; Bosselut et al., 2019). This has the shortcoming of relying on existing KBs which have usually limited coverage and may require extensive manual annotation.

Investigating how to improve commonsense knowledge acquisition *from large amount of raw textual data* appears an important endeavor. Nevertheless, in the context of PLMs, multiple questions still remain open about how current pretraining objectives lead, to some extent, to commonsense knowledge acquisition from raw text, and whether it is possible to improve upon this mechanism. In this work, our focus is to provide additional empirical evidence towards answering these questions.

Our first contribution is to analyze how large PLMs acquire commonsense and factual knowledge *during training*. This is in contrast to recent probing studies which analyze only the final checkpoint of a model. Moreover, instead of analyzing BERT-like models, we gain additional insights by comparing the learning dynamics of models trained with starkly different pretraining objec-

tives, Masked Language Modeling (BERT; Devlin et al. 2019) and Replaced Token Detection (ELECTRA; Clark et al. 2020). By using the LAMA (Petroni et al., 2019) and CAT (Zhou et al., 2020) probes we find that: i) ELECTRA acquires *more* commonsense knowledge than BERT; ii) common-sense knowledge is generally learnt before factual knowledge; and iii) contrary to factual knowledge which improves over training, commonsense performance tends in general to *plateau or decrease* with additional training: it appears there exists a trade-off between factual and commonsense knowledge acquisition in current PLMs.

Our second contribution is to show that, by pruning small magnitude weights of the pre-trained checkpoints, the networks *forget factual knowledge* while *commonsense knowledge is impacted to a lesser extent*. These observations evoke recent observations in a vision setting (Hooker et al., 2020): *atypical* or *rare* examples are learnt later in training and are forgotten after a similar magnitude pruning technique. This speaks to the fact that such examples, akin to factual knowledge in our case, require larger effective model capacity (Feldman et al., 2019; Baratin et al., 2021). At this point, our working hypothesis is that the learning process gradually increases model capacity and leads to *memorizing* training examples containing factual knowledge thus hindering acquisition of more general concepts to explain observed data (Carlini et al., 2020; Sagawa et al., 2020).

2 Probing Knowledge During Training

Our first step is to analyze more closely the dynamic of knowledge learning during training. While most past work has focused on analyzing the representations learnt by PLMs at the end of pretraining, we adopt a complementary approach to answer the following questions:

1. Do different pretraining schemes cause models to acquire commonsense and factual knowledge differently?

2. At what stages of pretraining do the models acquire these different aspects of knowledge?

We next describe the models we consider and their pretraining objectives. We also describe our probing protocols and details of the experimentals conducted in the course of this first investigation.

2.1 Pretraining Objectives

Masked Language Modeling (MLM) is an instantiation of pseudo-likelihood maximization (Besag, 1974) and it involves masking a fraction of the input tokens in a sentence and then learning a conditional model to predict the tokens that have been masked out. It is essentially a *fill-in-the-blanks* task where the model is tasked with learning the conditional probability of a particular masked out token given the context. BERT, a bidirectional text encoder built by stacking several transformer layers, uses this objective to learn general purpose representations that achieve excellent results on downstream NLU tasks (Devlin et al., 2019). There have been multiple extensions to BERT training that still use MLM in different forms, either by predicting contiguous spans (Joshi et al., 2020) or by adversarial training (Liu et al., 2020b). In this paper, we analyze its original formulation as found in (Devlin et al., 2019) and report probing results for BERT-base and BERT-large.

Replaced Token Detection MLM, while highly effective in practice, is computationally inefficent as the models utilize in general 15% of the input tokens per example to learn the distribution. ELECTRA (Clark et al., 2020) proposes an efficient new pretraining task - Replaced Token Detection (RTD) - which utilizes all tokens instead of only a fraction of the example. Differently from MLM, RTD leverages a discriminator and a generator. The discriminator is trained to solve a binary classification loss, where tokens that have been *replaced or corrupted* in the input by a generator network are assigned a label of 0 and ground-truth tokens are assigned a label of 1. The generator is a BERT-like model trained with MLM. At the end of training, the generator is discarded and only the discriminator is used. ELECTRA converges faster and results in learning of better representations that lead to higher quantitative performance on NLU benchmark tasks. We will probe for knowledge in the discriminator which *doesn't* directly use MLM, but a binary cross-entropy loss with a non-stationary negative sampling distribution instead. This begs the question of whether the loss used during pretraining leads to different dynamics of knowledge learning. We analyze both ELECTRA-base and ELECTRA-large.

Probe	Type	Example
T-Rex	Factual	Francesco Bartolomeo Conti was born in **[MASK]**. (*Florence*)
Google-RE	Factual	Mareva Galanter is a **[MASK]** actress and former beauty queen. (*French Polynesia*)
SQuAD	Factual	Newton played as **[MASK]** during Super Bowl 50. (*quarterback*)
ConceptNet	Commonsense	Joke would make you want to **[MASK]**. (*laugh*)
CAT	Commonsense	Paul tried to call George on the phone, but Paul wasn't successful (✓).
		Paul tried to call George on the phone, but George wasn't successful (✗).

Table 1: Examples of the commonsense and factual probes used in the evaluation. The first four are from LAMA (Petroni et al., 2019) and the last is a collection of commonsense tasks (details in (Zhou et al., 2020)).

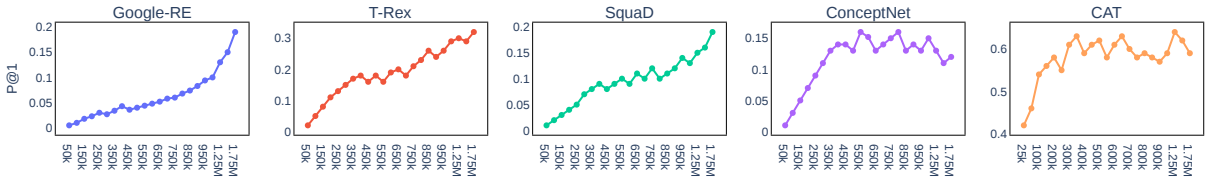

Figure 1: Probing of BERT and ELECTRA as a function of training steps. We plot the detail of each probe performance for BERT-Base. Commonsense probes generally converge faster but then plateaus, factual probes keep increasing.

2.2 Knowledge Probing Datasets

For testing commonsense and factual knowledge acquisition during training, we recur to LAMA (Petroni et al., 2019) and the Commonsense Acceptability Test (CAT; Zhou et al. 2020). We summarize these probes in Table 1.

LAMA The LAnguage Modeling Analysis (LAMA) probes are cloze-style sentence completion tasks where the model has to fill-in-the-blanks a missing word. Each component of the LAMA probes is designed to test the model for a specific type of knowledge that is learnt using pretraining. The examples are constrained to single token answers that examine different aspects of knowledge for both factual (T-Rex, Google-RE, SQuAD) and commonsense (ConceptNet). T-REx examples are constructed from Wikipedia triples and include 41 types of relations (facts about places, people, chemical compositions, etc). Google-RE consists of examples that consider three types of relations - place of birth, place of death and date of birth - extracted from Wikipedia and considerably harder than those in T-REx. SQuAD contains examples from the popular question answering dataset (Rajpurkar et al., 2016) rewritten as cloze-style statements with single token answers. Finally, examples in the ConceptNet split are verbalized triplets from the omonimous KB and contain commonsense relationships between words and phrases and consists of 16 types of relations.

CAT The Commonsense Acceptability Test (CAT) probes (Zhou et al., 2020) for commonsense knowledge and are adapted from exisiting commonsense datasets. The sentences in this corpus differ from each other by small phrases that alter the plausibility of such statements. The models are evaluated by measuring whether they score sentences that align with commonsense higher than those that don't. The dataset is also split in multiple sub-tasks: Conjunction Acceptability, Winograd, Sense making, Swag and Argument Reasoning.

2.3 Experimental Setup

In order to quantify the knowledge captured by the language models at different stages of pretraining, we store checkpoints at every 50k steps of pretraining and probe them with both LAMA and CAT probes. We use Wikipedia and Bookcorpus datasets[1] to pretrain BERT and ELECTRA (yet to complete).

For LAMA probes, we measure the amount of knowledge that is captured with the precision-at-one measure (P@1). We use the distribution of the MLM head to mimic a ranking of tokens with the token having the highest probability being the top ranked one. As ELECTRA is not trained using an MLM objective, we adapt the method for evaluating it as an MLM from (Clark et al., 2020). The particular setup we follow is explained in the Appendix. We compare the token ranked top by

[1] https://huggingface.co/datasets/

each model with the ground truth label for both BERT and ELECTRA type models. As P@1 is considered a strict method of evaluating the performance of models, we also measure P@10 and find the trends to be similar. For CAT probes, we only measure P@1 as the model is tasked with a binary classification task.

2.4 Results

Our results can be found in Figure 1. We are yet to run this experiment for ELECTRA and we will look to complete it after the course ends.

> Commonsense is acquired earlier in training, while factual knowledge improves with further training

As seen in Figure 1, commonsense performance appear to converge earlier than factual knowledge which for BERT requires a large number of pre-training steps. Among the three probing methods, the model performs best on the T-Rex dataset while the SQuAD and the Google-RE datasets prove to be more difficult in comparison but the performance still increases with training. Recent work highlighted how noise and atypical examples are learned later during training (Toneva et al., 2018; Hooker et al., 2020; Liu et al., 2020a; Baratin et al., 2021) in a vision setting. Similarly, our observation leads us to hypothesize that commonsense knowledge may be supported by a large variety of training examples and thus benefit from fast initial convergence (due to gradient alignment), while factual knowledge is shared by only few examples in the training set and therefore is learned later. We will expand on to this point in the next section.

> ELECTRA captures more commonsense and less factual knowledge than BERT

ELECTRA achieves better performance on the ConceptNet probe and on the CAT probes as compared to BERT across both variants (Base and Large). On the contrary, BERT outperforms ELECTRA considerably on all three of the factual knowledge probing tasks. This leads us to suggest that BERT is better at *memorizing* training data than ELECTRA, an effect already seen in Carlini et al. (2020) for a model trained with maximum likelihood, GPT-2. We give an intuition for this behavior in the final discussion.

> For BERT, commonsense performance plateaus after a given amount of training

There is an apparent trade-off between the ability of the model to learn commonsense and factual knowledge: it exists a point during training when the performance on the commonsense probes plateaus or even decreases while the performance on factual knowledge probes increases. The trade-off happens during the standard pre-training regime (400k-600k steps) in the case of BERT. This points to the interesting possibility that *memorizing* facts hinders further extraction of commonsense knowledge.

Summary Factual knowledge improves later in training. This points to the fact that the model starts using up more capacity to fit facts in the data. This elastic capacity allocation has been pointed out in previous work (Baratin et al., 2021). The increase in *effective capacity* may enhance memorization (Carlini et al., 2020; Sagawa et al., 2020) which, in turn, hinders extraction of more general commonsense knowledge. ELECTRA outperforms BERT in commonsense extraction which may be linked to its different objective function providing a different inductive bias to the model.

3 Effect of Magnitude Pruning

In this section, our aim is to explore what kind of knowledge is retained by a pruned network, i.e. with reduced capacity. This will help us clarify some of the observations made earlier.

3.1 Experimental Setup

Magnitude pruning (Han et al., 2015) usually consists of three simple steps:

1. Select a target percentage of model weights to be pruned denoted by $k\%$.
2. Calculate a threshold such that $k\%$ of weight magnitudes are under that threshold.
3. Remove those weights.

Our experimental procedure is as follows: we prune the fully-connected layers and biases of the publicly available pretrained BERT checkpoint (obtained after 1M steps of pre-training) and then compare the performance of the pruned BERT model on the LAMA probe suite to investigate what kind

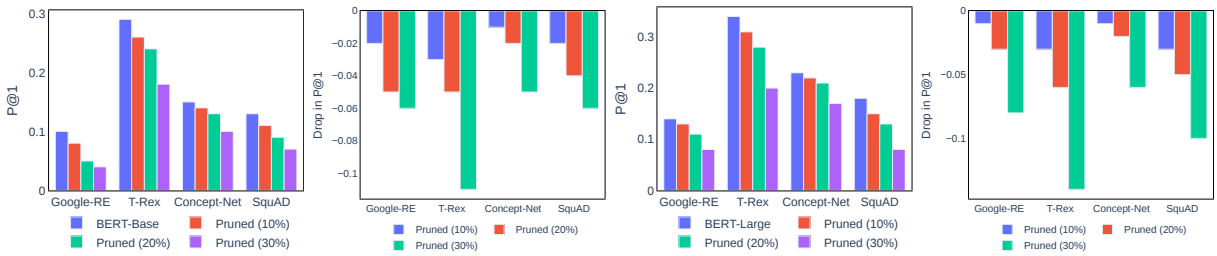

Figure 2: Pruning a pretrained BERT checkpoint impacts factual knowledge (Google-RE, T-Rex, SQuAD) more so than commonsense (ConceptNet) across all pruning threshold (10%, 20%, 30%).

of knowledge is lost or retained. We evaluate the decrease in performance for different values of pruning threshold $k \in \{10, 20, 30, 40\}$. We present the results for $k = 20\%$ in Figure 2 (and for other values, see Appendix).

3.2 Results

> Pruning BERT hurts factual knowledge more so than commonsense

This indicates that factual knowledge is stored to a greater extent in the smaller weights in the network. This pattern of results holds across model sizes (Base and Large). Because pruning diminishes the model capacity, this suggests that learning factual knowledge requires a larger capacity model.

This finding can be related to what has been empirically observed by Hooker et al. (2020) in the case of vision models. The authors investigate the effect of pruning and weight quantisation on the classification performance of ResNet models and determine that certain examples, which they label *Pruning Identified Examplars* (PIE), are affected by pruning to a significantly higher proportion as compared to other examples. It was found that PIEs are atypical or belonging to *long-tail* classes (under-represented in the dataset) and thus more challenging even to the non-pruned model.

Our findings suggest that, in the case of PLMs, sentences containing factual knowledge, especially when containing information pertaining to places of birth, death or date of birth, etc. may be characteristically similar to PIEs, i.e. examples containing idiosyncratic information not shared with many other examples in the dataset. Therefore, these must be fit by augmenting the model's *effective capacity*. Multiple works have highlighted the fact that neural networks tend to prefer simple, low-complexity solutions early in training and can increase capacity later during training to accomodate more challenging, atypical examples or noise (Baratin et al., 2021; Toneva et al., 2018; Liu et al., 2020a). This behavior is well-reflected by our previous observation, which is that factual knowledge seems to improve during the later stages of training.

3.3 Datasets

In addition to the LAMA and CAT probes, we report performance on commonsense downstream tasks such as Swag, HellaSwag and PEP3K plausibility classification. Finally, we also probe our models with the BLIMP (Warstadt et al., 2020) suite of syntactic evaluation tests, to evaluate how our modified training procedure impacts the acquisition of syntax. We present these datasets next.

Swag, HellaSwag Swag (Zellers et al., 2018) is a dataset for commonsense NLI which consists of multiple choice questions. For each question, the model is given a context from a caption and four choices for what might follow that particular statement. There exists only one correct answer which the model has to choose. HellaSwag (Zellers et al., 2019) is a more difficult version of the Swag dataset where the choices are created using an adversiarial filtering approach.

PEP3K The crowdsourced Physical Event Plausibility ratings datasets (PEP3K) of (Wang et al., 2018; Porada et al., 2021), measures the ability to identify plausible events as *chef-bake-cookie* from less plausible ones *fish-throw-elephant*. It is not a trivial problem for models to acquire this ability as plausibility often is something that is grounded in the real world which might not necessarily be captured in language. PEP3K consists of 3,062 events rated as physically plausible or implausible. We follow (Porada et al., 2021) and use AUC as our metric. We follow the valid/test split used by (Porada et al., 2021) and we create a train set using

Model	Swag	HellaSwag
BERT	81.6	40.5

Table 2: Performance of Base variants of BERT and ELECTRA on the dev set of Swag and test set of HellaSwag.

70% of their valid split and the rest is used as the validation set. The test set remains unchanged.

BLIMP BLIMP are *behavioral linguistic* probes built on top of linguistic minimal pairs. They test for grammatical acceptability and include syntax, semantics and morphology tests. The sentences in this corpus differ from each other by small phrases which alters the grammatical soundness of the statements. We probe the models by measuring whether they assign higher scores to grammatically acceptable sentences over incorrect ones. We compute the score of a sentence by sequentially masking words one at a time and compute the average of log probabilities of the masked words.

3.4 Discussion

ELECTRA vs BERT Our interpretation to the observation that ELECTRA and BERT capture factual and commonsense to different extents relies on the difference between the loss function they use during pre-training and its relationship to the amount of *memorization* that takes place during training. Carlini et al. (2020) shows that LMs trained with maximum likelihood can accurately store the training data and can be subject to extraction attacks. ELECTRA's discriminator is implicitly encouraged to do so only to the extent to which the distribution of noise samples is close to the true data distribution, e.g. the discriminator may not need to memorize that "The capital of Germany is [MASK]", where [MASK] = *Berlin*, if, during training, the generated samples for the masked position are cities of other countries, for example. The contrastive loss may be solved by relying on a feature indicating the masked token must be a city in Germany. In order for this hypothesis to be true, the generator must *make errors*. Curiously, (Clark et al., 2020) reports best performance when samples come from a generator smaller than the discriminator (thus more prone to errors).

Factual probes Although ELECTRA model improves commonsense extraction, one may wonder whether the drop in capturing factual knowledge may harm our model for downstream tasks that require it. Recent work investigates the use of external knowledge bases for knowledge-intensive NLP tasks to complement current PLMs models (Lewis et al., 2021). In that case, capturing factual knowledge during training may be less critical.

4 Related Works

Probing pretrained language models has been widely explored. Initial work started with exploring how the models captures linguistic capabilities (Peters et al., 2018; Goldberg, 2019; Ettinger et al., 2018; McCoy et al., 2019; Goldberg, 2019; Tenney et al., 2019b; Jawahar et al., 2019). There have been investigations into how these models acquire factual and commonsense knowledge (Petroni et al., 2019; Kassner and Schütze, 2020; Forbes et al., 2019; Rogers et al., 2020; Singh et al., 2020; Talmor et al., 2020; Kassner et al., 2020; Weir et al., 2020; Zhang et al., 2020b). In this work, we focus on gaining insights about at what stage during pre-training that knowledge acquired. This is also studied concurrently with our work in Liu et al. (2021), albeit using RoBERTa (Liu et al., 2019).

PLMs can be used as knowledge bases (Petroni et al., 2019). Some works used this to complete commonsense knowledge graphs (Bosselut et al., 2019; Roberts et al., 2020; Zhang et al., 2020a). Recent work exposed the risks for data privacy as private information can be easily extracted (Carlini et al., 2020), which raises the question whether such memorization is desired. Other works use external knowledge bases to induce commonsense in PLMs (Porada et al., 2021).

Pruning PLMs has been explored before with the focus of reducing the size of the models usually at the expense of downstream performance (Gordon et al., 2020; Fan et al., 2019; Prasanna et al., 2020; Chen et al., 2020; Tang et al., 2019). Our method is inspired by these works, although our motivation is different and we can improve performance on downstream tasks.

5 Conclusion

We demonstrated dynamics of learning of factual and commonsense knowledge. We used our insights to devise a technique to delay memorization of factual knoweldge, called Iterated Reset. This leads, both in Base and Large versions, to mod-

els capturing more commonsense and less factual knowledge. We believe that our results will benefit building data efficient models, which learn more robust features from input data by limiting memorization. Our results could also have ramifications for privacy considerations regarding large pretrained language models (Carlini et al., 2020).

References

Aristide Baratin, Thomas George, César Laurent, R Devon Hjelm, Guillaume Lajoie, Pascal Vincent, and Simon Lacoste-Julien. 2021. Implicit regularization via neural feature alignment. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2269–2277. PMLR.

Julian Besag. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models.

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.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

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

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout.

Joshua Feldman, Joe Davison, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained models. *CoRR*, abs/1909.00505.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities.

Mitchell A. Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing bert: Studying the effects of weight pruning on transfer learning.

Song Han, Jeff Pool, John Tran, and William J Dally. 2015. Learning both weights and connections for efficient neural networks. *arXiv preprint arXiv:1506.02626*.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

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

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2020. What do compressed deep neural networks forget?

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P. Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *CoRR*, abs/2005.03692.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *CoRR*, abs/2010.05953.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

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.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge?

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Leo Z. Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does roberta know and when?

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. 2020a. Early-learning regularization prevents memorization of noisy labels. *arXiv preprint arXiv:2007.00151*.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020b. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

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

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. 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.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *EMNLP*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Ian Porada, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. Can a gorilla ride a camel? learning semantic plausibility from text. *arXiv preprint arXiv:1911.05689*.

Ian Porada, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. Modeling event plausibility with consistent conceptual abstraction. *arXiv e-prints*, pages arXiv–2104.

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 (EMNLP)*, pages 3208–3229, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model?

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Jaspreet Singh, Jonas Wallat, and Avishek Anand. 2020. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 174–183, Online. Association for Computational Linguistics.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics–on what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. Bert rediscovers the classical nlp pipeline. In *Association for Computational Linguistics*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019b. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019c. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling semantic plausibility by injecting world knowledge. *arXiv preprint arXiv:1804.00619*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

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*.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020a. Transomcs: From linguistic graphs to commonsense knowledge. *arXiv preprint arXiv:2005.00206*.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R. Bowman. 2020b. When do you need billions of words of pretraining data?

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating Commonsense in Pre-Trained Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9733–9740.