Language Model Decomposition: Quantifying the Dependency and Correlation of Language Models

Hao Zhang
Supportiv Inc, Berkeley, CA
haozhang@alumni.princeton.edu

Abstract

Pre-trained language models (LMs), such as BERT (Devlin et al., 2018) and its variants, have led to significant improvements on various NLP tasks in past years. However, a theoretical framework for studying their relationships is still missing. In this paper, we fill this gap by investigating the *linear dependency* between pre-trained LMs. The linear dependency of LMs is defined analogously to the linear dependency of vectors. We propose Language Model Decomposition (LMD) to represent a LM using a linear combination of other LMs as basis, and derive the closed-form solution. A goodness-of-fit metric for LMD similar to the coefficient of determination is defined and used to measure the linear dependency of a set of LMs. In experiments, we find that BERT and eleven (11) BERT-like LMs are 91% linearly dependent. This observation suggests that current state-of-the-art (SOTA) LMs are highly "correlated". To further advance SOTA we need more diverse and novel LMs that are less dependent on existing LMs.

1 Introduction

Large-scale pre-trained language models (LMs) have pushed state-of-the-art (SOTA) of NLP recently (Han et al., 2021; Qiu et al., 2020). Following BERT (Devlin et al., 2018), many variants and improvements have been proposed since then. They differ from BERT in various aspects, in terms of training data (Liu et al., 2019), multilingual support (Conneau et al., 2019), model size (Sanh et al., 2019; Lan et al., 2019), pre-training objective (Yang et al., 2019), model architecture (Clark et al., 2020), attention structure (He et al., 2020), and sequence length (Beltagy et al., 2020). However, to the best of our knowledge, their relationships have not been studied from a mathematical perspective at the time of writing this paper.

In this work, we present a *quantitative* framework for studying LM *dependency* and *correlation*. Conceptually, we view LM (e.g., encoders like BERT (Devlin et al., 2018)) as a vector-valued function (or random vector) $\mathbf{u}(x) : \Omega \to \mathbb{R}^d$, where x is text sequence from the sequence space Ω, and $\mathbb{R}^d$ is the sequence embedding space. Sequence embedding is defined as the mean pooling of the last layer's token embeddings. We define linear combination and dependency of LMs analogously to vectors/functions, and based upon these definitions we propose the Language Model Decomposition (LMD) algorithm to analyze the linear dependency of LMs. A goodness-of-fit metric is defined for LMD to quantify the linear dependency and correlation of LMs. In experiments, we find BERT and its successors are highly linearly dependent.

Theoretically, the linear dependency between LMs implies redundancy, since they can be represented by each other. Practically, it simply means building more redundant LMs do not bring in new knowledge. If LMs are more linearly independent, we can potentially distill more knowledge (Alkhulaifi et al., 2021) from multiple diverse models, and combine them to create more powerful models.

The contributions of this paper are: (1) We formalize the notion of *linear dependency* for LMs, and propose Language Model Decomposition (LMD) to study linear dependency; (2) We define an universal metric based on LMD to quantify the dependency and correlation of LMs. (3) Our experiments reveal that BERT and its variants are 91% "correlated", suggesting current SOTA LMs are highly redundant. The code is available at https://github.com/haozhg/lmd.

2 Language Model Decomposition

2.1 Notations and Definitions

In this section, we formalize some necessary mathematical notations and definitions.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2508 - 2517

Definition 2.1.1 (Linear Combination of LMs) *Given n LMs $\{\mathbf{u}_i(x)\}_{i=1}^n$, a linear combination of these LMs is $\sum_{i=1}^k \mathbf{W_i}\mathbf{u_i}(x)$, where $\mathbf{W_i} \in \mathbb{R}^{d\times d}$ are matrices.*

Notice that here the coefficients are matrices, while for vectors/functions they are scalars.

Definition 2.1.2 (Linear Dependent LMs) *A set of LMs $\{\mathbf{u}_i(x)\}_{i=1}^n$ is linearly dependent, if there exists matrices $\{\mathbf{W}_i\}_{i=1}^n$ not all singular, such that $\sum_{i=1}^n \mathbf{W_i}\mathbf{u_i}(x) = \mathbf{0}$, $\forall x \in \Omega$, where $\mathbf{0}$ denotes the zero vector.*

Recall that a matrix $\mathbf{A}$ is singular $\iff$ $\mathbf{A}$ is not invertible $\iff$ $\det(\mathbf{A}) = 0$. If the LMs are not linearly dependent, they are said to be linearly independent, that is, the above equation can only be satisified by singular $\mathbf{W}_i, \forall i$.

Corollary 2.1.1 (Linear Dependency Condition) *A set of LMs is linearly dependent if and only if one of them is zero or a linear combination (as in Definition 2.1.2) of the others.*

2.2 Language Model Decomposition

To quantify the degree of "linear dependency" of a set of LMs, we propose Language Model Decomposition (LMD), where a LM is approximated by a linear combination of other LMs. LMD is motivated by the Galerkin projection method (Reddy, 2010) that is widely used for model reduction and reduced-order modeling. In particular, given a *target* LM $\mathbf{u}(x)$, and k *basis* LMs $\{\mathbf{v_i}(x)\}_{i=1}^k$, we fit a model in the following form

$$\mathbf{u}(x) = \sum_{i=1}^{k} \mathbf{W_i}\mathbf{v_i}(x) + \mathbf{e}(x), \tag{1}$$

where $\mathbf{e}(x)$ is the residual term. To simplify the derivation, we treat LMs as random vectors from now on. To minimize the residual, we solve the following optimization problem

$$\min_{\{\mathbf{W_i}\}_{i=1}^k} L(\mathbf{W_i}) = \mathrm{E}\,\|\mathbf{e}(x)\|^2, \tag{2}$$

where $\mathrm{E}[\cdot]$ is the expectation over $x \in \Omega$, and $\|\cdot\|$ is the L_2 norm. $L(\mathbf{W_i})$ is convex (not necessarily strictly convex), so global optimum exists (not necessarily unique). In the special case of $d = 1$, it reduces to multivariate linear regression.

Closed-form Solution For simplicity, let

$$\mathbf{W} = [\mathbf{W_1}, \mathbf{W_2}, \ldots, \mathbf{W_k}] \in \mathbb{R}^{d\times kd},$$
$$\mathbf{z} = [\mathbf{v_1}^\intercal, \mathbf{v_2}^\intercal, \ldots, \mathbf{v_k}^\intercal]^\intercal \in \mathbb{R}^{kd},$$

we can rewrite Eq (2) as

$$\begin{aligned} L &= \mathrm{E}[\|\mathbf{u} - \mathbf{W}\mathbf{z}\|^2] \\ &= \mathrm{E}[(\mathbf{u} - \mathbf{W}\mathbf{z})^\intercal(\mathbf{u} - \mathbf{W}\mathbf{z})] \\ &= \mathrm{E}[\mathbf{z}^\intercal\mathbf{W}^\intercal\mathbf{W}\mathbf{z} - 2\mathbf{u}^\intercal\mathbf{W}\mathbf{z} + \mathbf{u}^\intercal\mathbf{u}] \\ &= \mathrm{E}[\mathrm{tr}(\mathbf{z}\mathbf{z}^\intercal\mathbf{W}^\intercal\mathbf{W})] - 2\,\mathrm{E}[\mathrm{tr}(\mathbf{z}\mathbf{u}^\intercal\mathbf{W})] + c \\ &= \mathrm{tr}(\mathrm{E}[\mathbf{z}\mathbf{z}^\intercal]\mathbf{W}^\intercal\mathbf{W}) - 2\,\mathrm{tr}(\mathrm{E}[\mathbf{z}\mathbf{u}^\intercal]\mathbf{W}) + c, \end{aligned}$$

where $c = \mathrm{E}[\mathbf{u}^\intercal\mathbf{u}]$ is a constant, and we have used cyclic property of trace, linearity of trace, and linearity of expectation. Using the following matrix calculus identity (Petersen et al., 2008),

$$\frac{\partial\,\mathrm{tr}(\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{X}^\intercal\mathbf{C})}{\partial\mathbf{X}} = \mathbf{C}\mathbf{A}\mathbf{X}\mathbf{B} + \mathbf{A}^\intercal\mathbf{C}^\intercal\mathbf{X}\mathbf{B}^\intercal,$$

the gradient is

$$\frac{\partial L}{\partial \mathbf{W}} = 2(\mathbf{W}\,\mathrm{E}[\mathbf{z}\mathbf{z}^\intercal] - \mathrm{E}[\mathbf{u}\mathbf{z}^\intercal]). \tag{3}$$

Setting gradient to zero, the solution is

$$\mathbf{W} = \mathrm{E}[\mathbf{u}\mathbf{z}^\intercal](\mathrm{E}[\mathbf{z}\mathbf{z}^\intercal])^{-1}, \tag{4}$$

assuming $\mathrm{E}[\mathbf{z}\mathbf{z}^\intercal]$ is full rank (in this case there is an unique global optimal solution). $\mathrm{E}[\mathbf{z}\mathbf{z}^\intercal]$ is the covariance matrix (for simplicity we assume all LMs are mean-subtracted), which is (symmetric) positive semi-definite. Its eigenvalues are real and non-negative. In practice, expectation is approximated with finite samples.

If $\mathrm{E}[\mathbf{z}\mathbf{z}^\intercal]$ is not full rank, Eq (2) is convex but not strictly convex, and there are *infinitely many optimal solutions*. The *minimum-norm* optimal solution is

$$\mathbf{W} = \mathrm{E}[\mathbf{u}\mathbf{z}^\intercal](\mathrm{E}[\mathbf{z}\mathbf{z}^\intercal])^{+}, \tag{5}$$

where $(\mathrm{E}[\mathbf{z}\mathbf{z}^\intercal])^{+}$ is the Moore–Penrose inverse (Moore, 1920) of $\mathrm{E}[\mathbf{z}\mathbf{z}^\intercal]$. Moore–Penrose inverse exists even for non-invertible matrix or non-square matrix. In the special case of a full rank square matrix, Moore–Penrose inverse reduces the "standard" matrix inverse.

Regularization A small regularization term $\lambda\|\mathbf{W}\|^2$ can added to the loss function in Eq (2). Mathematically, it ensures that Eq (2) is a strictly convex hence the global optimum is unique regardless of the rank of $\mathrm{E}[\mathbf{z}\mathbf{z}^\intercal]$. Empirically, this increases the numerical stability and accuracy when computing matrix inverse. With regularization, the unique global optimal solution is

$$\mathbf{W} = \mathrm{E}[\mathbf{u}\mathbf{z}^\intercal](\lambda\mathbf{I} + \mathrm{E}[\mathbf{z}\mathbf{z}^\intercal])^{-1}, \tag{6}$$

where $\lambda\mathbf{I} + \mathrm{E}[\mathbf{z}\mathbf{z}^\intercal]$ is positive definite and always invertible.

Bias Term In the above derivation, for simplicity we assume all LMs are mean-subtracted. If we include a bias term in LMD equation (1), it becomes

$$\mathbf{u}(x) = \sum_{i=1}^{k} \mathbf{W_i v_i}(x) + \mathbf{b} + \mathbf{e}(x), \tag{7}$$

where $\mathbf{b} \in \mathbb{R}^d$ is the bias term. In this case, the solution is

$$\begin{aligned} \mathbf{W} &= \mathrm{cov}(\mathbf{u}, \mathbf{z})(\mathrm{cov}(\mathbf{z}, \mathbf{z}))^{+}, \\ \mathbf{b} &= \mathrm{E}[\mathbf{u}] - \mathbf{W}\,\mathrm{E}[\mathbf{z}], \end{aligned} \tag{8}$$

where $\mathrm{cov}(\mathbf{z}, \mathbf{z}) = \mathrm{E}[\mathbf{z}\mathbf{z}^\intercal] - \mathrm{E}[\mathbf{z}]\,\mathrm{E}[\mathbf{z}^\intercal]$ is the covariance matrix of $\mathbf{z}$, and $\mathrm{cov}(\mathbf{u}, \mathbf{z}) = \mathrm{E}[\mathbf{u}\mathbf{z}^\intercal] - \mathrm{E}[\mathbf{u}]\,\mathrm{E}[\mathbf{z}^\intercal]$ is the cross-covariance matrix of $\mathbf{u}, \mathbf{z}$.

2.3 Quantitative Measure of Dependency and Correlation

Dependency Between Multiple Language Models To measure the goodness-of-fit for LMD, we define R^2, analogous to the coefficient of determination used in linear regression (Draper and Smith, 1998). In particular,

$$\mathrm{R}^2(\mathbf{u}, \{\mathbf{v_i}\}_{i=1}^k) = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}}, \tag{9}$$

where $\mathrm{SSR} = \mathrm{E}[\|\mathbf{e}(x)\|^2]$ is the residual sum of squares, and $\mathrm{SST} = \mathrm{E}[\|\mathbf{u}(x) - \mathrm{E}[\mathbf{u}(x)]\|^2]$ is the total sum of squares. Here we view R^2 as a function of the target LM $\mathbf{u}(x)$ and basis LMs $\{\mathbf{v_i}(x)\}_{i=1}^k$. Note that $\mathrm{R}^2 \in [0, 1]$. $\mathrm{R}^2 = 1$ implies that the target model and the basis models are perfectly linearly dependent (by Corollary 2.1.1). A larger R^2 indicates that target model is well approximated by basis models, and LMs are more linearly dependent. Therefore, R^2 is a quantitative measure of the target LM's dependency on the basis LMs.

Correlation Between Multiple Language Models The degree of correlation among a group of n LMs $\{\mathbf{u}_i(x)\}_{i=1}^n$ is defined as

$$\rho(\{\mathbf{u}_i\}_{i=1}^n) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{R}^2(\mathbf{u}_i, \{\mathbf{u_{-i}}\}), \tag{10}$$

where $\{\mathbf{u_{-i}}\}$ are the remaining $n-1$ LMs after removing the *target* LM $\mathbf{u}_i$, which are used as the *basis* LMs. Note that $\rho(\{\mathbf{u}_i\}_{i=1}^n) \in [0, 1]$. A value of 1 implies that LMs are linear dependent (by Corollary 2.1.1)) and furthermore each of them can be represented by the rest.

Correlation Between Two Language Models The measure of correlation between two LMs is of particular interest, in this case it simplifies to

$$\rho(\mathbf{u}, \mathbf{v}) = \frac{1}{2}(\mathrm{R}^2(\mathbf{u}, \mathbf{v}) + \mathrm{R}^2(\mathbf{v}, \mathbf{u})), \tag{11}$$

where $\mathrm{R}^2(\mathbf{u}, \mathbf{v})$ is the shorthand for $\mathrm{R}^2(\mathbf{u}, \{\mathbf{v}\})$ when there is only one basis LM. Notice that $\rho(\mathbf{u}, \mathbf{v}) \in [0, 1]$. By definition, $\rho(\mathbf{u}, \mathbf{v})$ is symmetric, i.e, $\rho(\mathbf{u}, \mathbf{v}) = \rho(\mathbf{v}, \mathbf{u})$. However, $\mathrm{R}^2(\mathbf{u}, \mathbf{v})$ is asymmetric in general. $\rho(\mathbf{u}, \mathbf{u}) = 1$, meaning that a LM is perfectly "correlated" with itself.

In the special case of $d = 1, n = 2$, LMD reduces to simple linear regression with a single feature. Furthermore, both Eq (9) and Eq (11) reduce to the "standard" coefficient of determination, which is the square of the correlation coefficient.

model name	huggingface checkpoint name
XLM-R	xlm-roberta-base
M-BERT	bert-base-multilingual-cased
Longformer	allenai/longformer-base-4096
DeBERTa	microsoft/deberta-base
distil-M-BERT	distilbert-base-multilingual-cased
RoBERTa	roberta-base
XLNet	xlnet-base-cased
BERT	bert-base-uncased
ELECTRA	google/electra-base-discriminator
distil-RoBERTa	distilroberta-base
distil-BERT	distilbert-base-uncased
ALBERT	albert-base-v2

Table 1: Model name and huggingface checkpoint name

3 Experiments and Results

3.1 Experiments

Language Models BERT (Devlin et al., 2018) and many of its successors are pre-trained encoder LMs, which take in text sequence and output sequence level embedding (defined as the mean pooling of the last layer's token embeddings). In this work, we consider twelve (12) LMs, including BERT (Devlin et al., 2018), distil-BERT (Sanh et al., 2019), M-BERT (Devlin et al., 2018), distil-M-BERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), distil-RoBERTa (Sanh et al., 2019), XLM-R (Conneau et al., 2019), XLNet (Yang et al., 2019), ALBERT (Lan et al., 2019), ELECTRA (Clark et al., 2020), DeBERTa (He et al., 2020), and Longformer (Beltagy et al., 2020). The We apply LMD to examine the linear dependency between these LMs.

Data We utilize the English Wikipedia (Devlin et al., 2018)[1] and BooksCorpus (Zhu et al., 2015)[2], which are used in pre-training BERT (Devlin et al., 2018). The sequence length T ranges from 16 to 512 tokens, as determined by the BERT tokenizer. When fed into other models, all sequences are truncated at T tokens as determined by their respective tokenizers. We randomly sample 512,000 and 51,200 sequences as the train data and test data respectively. In our experiments, we find that further increasing the size of the data does not affect the results much. By central limit theorem, with enough samples the estimation of expectations (see Eq (4)) will become very accurate.

Implementation We download the pre-trained checkpoints for the considered LMs from the public huggingface model hub (Wolf et al., 2020), see Table 1 for the full list of checkpoint names. We ran experiments using PyTorch with a NVIDIA V100 GPU. To improve numerical stability, a small regularization term is added to ensure the minimal eigenvalue of $\lambda\mathbf{I} + \mathrm{E}[\mathbf{z}\mathbf{z}^\intercal]$ is 10^{-6}. In our experiments, the sequence embedding is the mean pooling of the last layer's token embeddings.

3.2 Results

The closed-form optimal solution (Eq (4)) depends on expectations, which are approximated using train data. The evaluation metrics R^2, ρ are reported on the test data. LMD Results on another dataset are in Appendix B.

Pairwise Correlation of Language Models The symmetric pairwise correlation $\rho(\mathbf{u}_i, \mathbf{u}_j)$ between LMs is visualized in Figure 1 (for $T = 512$), and the asymmetric dependency measure $\mathrm{R}^2(\mathbf{u}_i, \mathbf{u}_j)$ is shown in Figure 3 of Appendix A. First, notice that distilled models (distil-M-BERT, distil-RoBERTa, and distil-BERT) are highly "correlated" with the orignal full models, because they distil knowledge (Alkhulaifi et al., 2021) from the full model. Second, multilingual LMs (XLM-R and M-BERT) have lower correlation with other LMs. Our hypthesis is that multilingual LMs are trained to generalize to multiple languages, therefore they contain knowledge beyond a single language.

Dependency and Correlation of Multiple Language Models The dependency and correlation

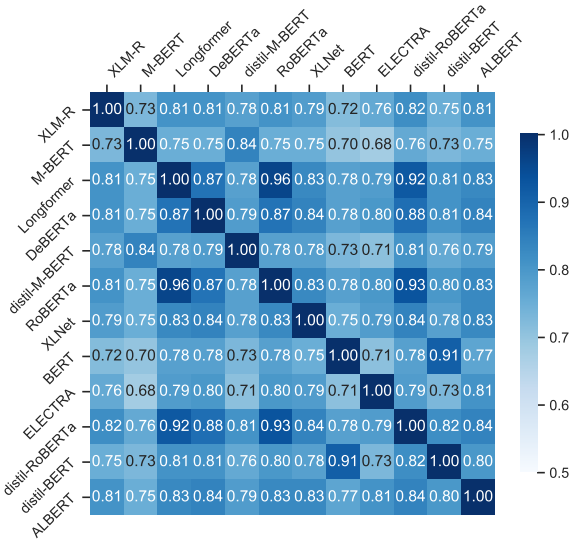


Figure 1: Results on `English Wikipedia` and `BooksCorpus`. Symmetric pairwise correlation $\rho(\mathbf{u}_i, \mathbf{u}_j)$ (as defined in Eq (10)) among 12 language models. The text sequence length is $T = 512$.

of multiple LMs $\mathrm{R}^2(\mathbf{u}_i, \{\mathbf{u}_{-\mathbf{i}}\})$ is shown in Figure 2. For exact numbers, see Table 2 in Appendix A. First, note that R^2 is around 90% for all models (for $T = 512$). This is consistent with the fact that all models are variants of BERT. The group correlation ρ is 91.25%, suggesting that these LMs are highly linearly dependent, and there is some redundancy in them. Second, R^2 is smaller for shorter sequence length T. We suspect the reason is that the sequence embedding is the mean pooling of the last layer's token embeddings, so by central limit theory the variations in shorter sequence embedding is larger. Therefore it is harder to approximate the target LM using others as basis LMs. Third, similar to pairwise correlation, note that multilingual LMs have lower dependency compared with others. Finally, for XLNet and ELECTRA, R^2 is lower compared to other LMs (especially for shorter sequences). Our hypothesis is that because they have very different training objectives: Permutation Language Modeling (Yang et al., 2019) for XLNet and Replaced Token Detection (Clark et al., 2020) for ELECTRA. Therefore they encode text differently from the rest of the models which mostly use Masked Language Modeling (Devlin et al., 2018).

4 Conclusion and Future Work

In this work, we present a theoretical framework to study the relationships between language models (LMs). We formalize the definitions of linear

[1]https://huggingface.co/datasets/wikipedia
[2]https://huggingface.co/datasets/bookcorpus

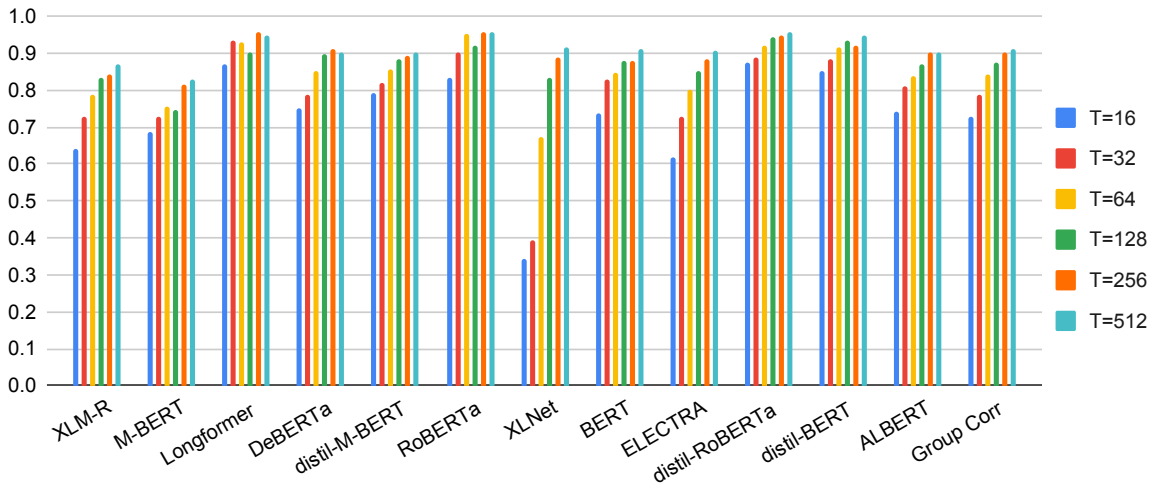


Figure 2: Results on `English Wikipedia` and `BooksCorpus`. Dependency $R^2(\mathbf{u}_i, \{\mathbf{u}_{-\mathbf{i}}\})$ (as defined in Eq (9)) of language models. The `Group Corr` refers to the group correlation measure $\rho(\{\mathbf{u}_i\}_{i=1}^n)$ (as defined in Eq (10)). T is the text sequence length. We take each model as *target* model, and apply LMD using the remaining eleven (11) models as *basis* (as defined in Eq (1)).

combination and linear dependency of LMs. We further propose the Language Model Decomposition (LMD) algorithm to represent one LM using other LMs as basis. A LMD metric is developed to quantify the linear dependency of LMs. Our experiments show that BERT and its variants are 91% "correlated". This suggests that there is redundancy in SOTA pre-trained language models. Preliminary results in this paper demonstrate the potential of LMD as a framework to quantitatively analyze the relationships among LMs. Finally, we leave some open questions to motivate future research.

1. Can we leverage the linear dependency or independency of LMs to improve pre-training and/or fine-tuning?
2. Is there any connection between R^2 and LM performance in the pre-training and fine-tuning stage? What does the "unexplained variance" (i.e., $1 - R^2$) of each LM represent?
3. To further improve SOTA, how can we learn complementary language representations that are less linearly dependent on existing LMs?
4. Are LMs still highly linearly dependent after fine-tuning on downstream tasks? How does their linear dependency change during fine-tuning?
5. Are multilingual LMs (e.g., M-BERT, XLM-R) linearly dependent on monolingual LMs?

5 Limitations

In our preliminary experiments, we study encoder LMs similar to BERT. It is worthwhile to investigate other types of pre-trained LMs, including encoder-decoder LMs (e.g, T5 (Raffel et al., 2020), BART (Lewis et al., 2020)), decoder LMs (e.g, GPT-2 (Radford et al., 2019)), pre-BERT models (e.g, ELMo (Peters et al., 2018), ULMFit (Howard and Ruder, 2018)), and even domain-specific LMs (e.g, FinBERT (Araci, 2019)). The proposed "dependency" and "correlation" measures quantify the "similarity" between LMs, but there are still open questions related to interpretation. Wikipedia and BooksCorpus are used in our experiments, while the results on a "neutral dataset" (such as a large corpus that is not used in the pre-training of any of the LMs) are also worth examination. We have chosen the mean pooling of the last layers embedding as the text sequence embedding because it is commonly used for many downstream tasks. This choice makes the LMD algorithm model agnostic, meaning that it treats LM as a black box. It only requires the final sequence embedding, but not the intermediate representations. However, for other task, the token level embedding is very important (e.g, determining the start and end location of the answer for question-answering). Therefore, layer-wise and token-level (including the [CLS] and [MASK] token) "correlation" is of interest as well, and further research is needed.

References

Abdolmaged Alkhulaifi, Fahad Alsahli, and Irfan Ahmad. 2021. Knowledge distillation in deep learning and its applications. *PeerJ Computer Science*, 7.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Norman R Draper and Harry Smith. 1998. *Applied regression analysis*, volume 326. John Wiley & Sons.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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. *arXiv preprint arXiv:1907.11692*.

Eliakim H Moore. 1920. On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.*, 26:394–395.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. 2008. The matrix cookbook. *Technical University of Denmark*, 7(15):510.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

JN Reddy. 2010. *An introduction to the finite element method*, volume 1221. McGraw-Hill New York.

Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, and German Rigau. 2010. Wikicorpus: A word-sense disambiguated multilingual wikipedia corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

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*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Acknowledgments

The author is grateful for the tremendous support from his wife and Corgi.

A Details for Experiment Results

In this section, we show more details for experiment results in Section 3.2.

Pairwise Dependency of Language Models Figure 3 shows the pairwise asymmetric dependency $R^2(\mathbf{u}_i, \mathbf{u}_j)$ between language models. Figure 1 shows the symmetric correlation $\rho(\mathbf{u}_i, \mathbf{u}_j)$.

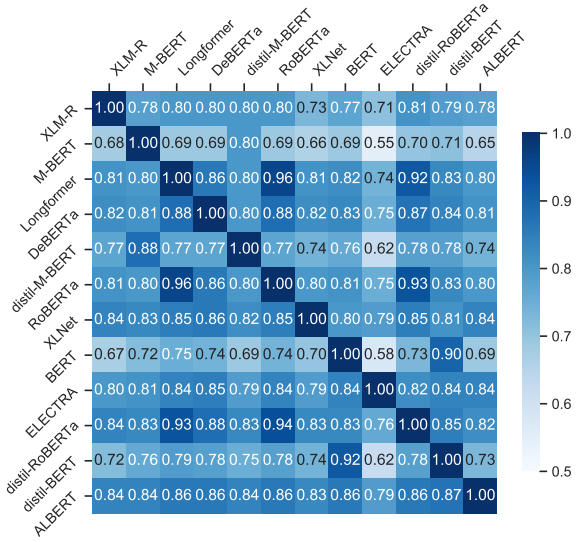


Figure 3: Results on `English Wikipedia` and `BooksCorpus`. Asymmetric pairwise dependency $R^2(\mathbf{u}_i, \mathbf{u}_j)$ (as defined in Eq (9)) of *target* LM (y-axis) on *basis* LM (x-axis). The text sequence length is $T = 512$.

Dependency and Correlation of Multiple Language Models Table 2 shows the exact numbers for dependency $R^2(\mathbf{u}_i, \{\mathbf{u}_{-i}\})$ (defined in Eq (9)) between language models. The visualization is shown in Figure 2.

B Additional Experiment Results

We also run LMD using the `raw English Wikicorpus` dataset (Reese et al., 2010)[3]. This corpus is not directly used in pre-training of aforementioend LMs, though it is in the same domain as the English Wikipedia.

The same group of language models are used (see Section 3.1, Table 1), and the train/validation/test sample size is 128,000/12,800/12,800. The regularization parameter is $\lambda = 10^{-6}$. We fit the LMD parameters using train set, validate on the validation set, and report dependency and correlation measures on the test set.

Pairwise Dependency of Language Models The pairwise dependency of language models are show in Figure 4. Overall, the dependency $R^2(\mathbf{u}_i, \mathbf{u}_j)$ (as defined in Eq (9)) increases with sequence length. This provides more evidence for our hypothesis in Section 3.2. Because the sequence embedding is the mean pooling of the last layer's token embeddings, by central limit theory the variations in shorter sequence embedding is larger. Therefore it is harder to approximate the target LM using others as basis LMs.

Dependency and Correlation of Multiple Language Models The dependency $R^2(\mathbf{u}_i, \{\mathbf{u}_{-i}\})$ (as defined in Eq (9)) and correlation $\rho(\{\mathbf{u}_i\}_{i=1}^n)$ (as defined in Eq (10)) of multiple language models are show in Figure 5, and the exact numbers are in Table 3. The results are similar to the results on `English Wikipedia` and `BooksCorpus` (see Figure 2, Table 2.).

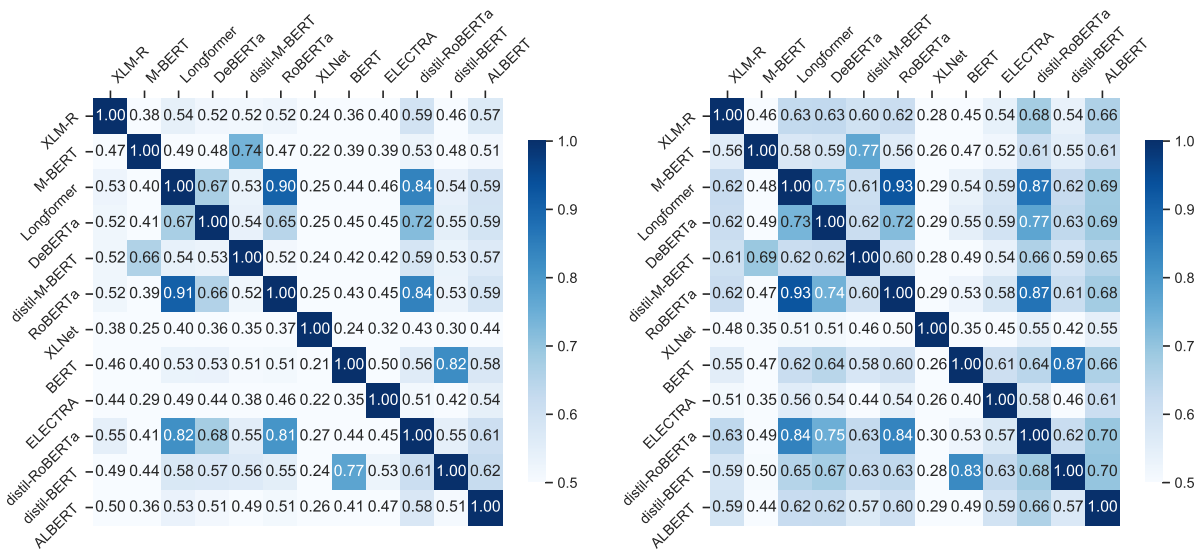
[3] https://huggingface.co/datasets/wikicorpus

	T=16	T=32	T=64	T=128	T=256	T=512
XLM-R	0.6415	0.7279	0.7866	0.8316	0.8430	0.8692
M-BERT	0.6861	0.7278	0.7541	0.7483	0.8139	0.8306
Longformer	0.8696	0.9329	0.9289	0.9015	0.9581	0.9481
DeBERTa	0.7492	0.7883	0.8503	0.8984	0.9105	0.9038
distil-M-BERT	0.7932	0.8190	0.8568	0.8830	0.8940	0.9035
RoBERTa	0.8322	0.9037	0.9512	0.9180	0.9566	0.9552
XLNet	0.3412	0.3925	0.6733	0.8341	0.8865	0.9154
BERT	0.7381	0.8293	0.8476	0.8803	0.8808	0.9124
ELECTRA	0.6190	0.7278	0.8029	0.8499	0.8852	0.9067
distil-RoBERTa	0.8722	0.8883	0.9221	0.9417	0.9469	0.9550
distil-BERT	0.8505	0.8827	0.9141	0.9323	0.9199	0.9472
ALBERT	0.7422	0.8093	0.8374	0.8696	0.9000	0.9027
Group Corr	0.7279	0.7858	0.8438	0.8741	0.8996	0.9125

Table 2: Results on `English Wikipedia` and `BooksCorpus`. Dependency $\mathrm{R}^2(\mathbf{u}_i, \{\mathbf{u}_{-\mathbf{i}}\})$ (as defined in Eq (9)) of language models. The `Group Corr` refers to the group correlation measure $\rho(\{\mathbf{u}_i\}_{i=1}^n)$ as defined in Eq (10). T is the text sequence length. We take each model as *target* model, and apply LMD using the remaining eleven (11) models as *basis* (as defined in Eq (1)).

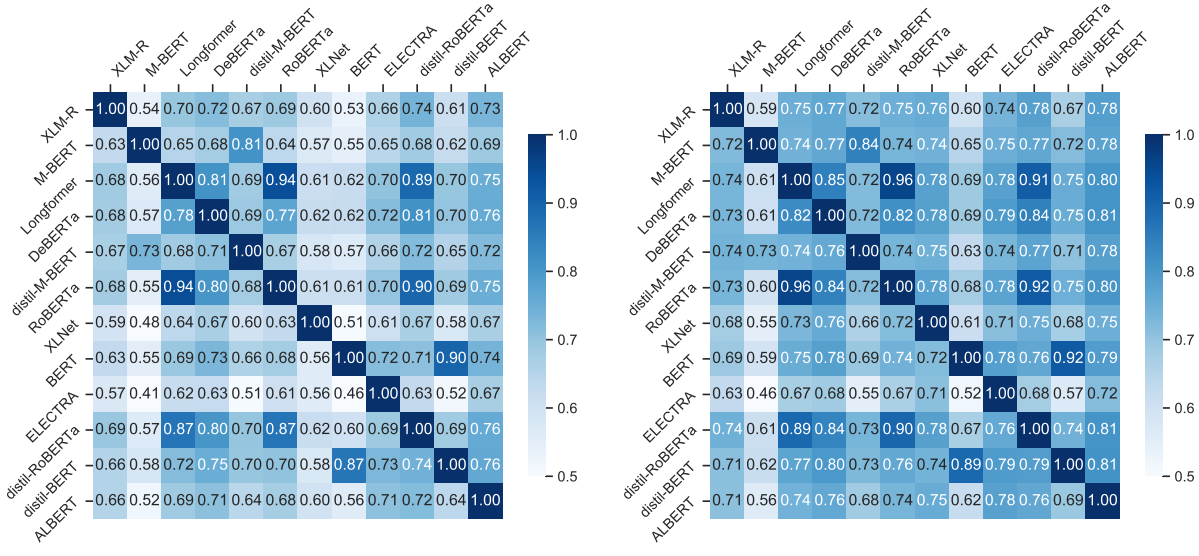
	T=16	T=32	T=64	T=128	T=256	T=512
XLM-R	0.6319	0.7179	0.7714	0.8215	0.8562	0.8808
M-BERT	0.6847	0.7234	0.7690	0.7753	0.8312	0.8741
Longformer	0.9240	0.9413	0.9542	0.9668	0.9753	0.9741
DeBERTa	0.7555	0.8260	0.8738	0.9022	0.9202	0.9311
distil-M-BERT	0.8048	0.8331	0.8659	0.8843	0.9035	0.9176
RoBERTa	0.9186	0.9401	0.9532	0.9667	0.9750	0.9761
XLNet	0.3255	0.3766	0.6868	0.8419	0.8959	0.9194
BERT	0.7829	0.8385	0.8784	0.9036	0.9189	0.9283
ELECTRA	0.6143	0.7173	0.8066	0.8618	0.8993	0.9228
distil-RoBERTa	0.8868	0.9077	0.9264	0.9418	0.9536	0.9612
distil-BERT	0.8593	0.8959	0.9230	0.9396	0.9477	0.9514
ALBERT	0.7240	0.7927	0.8435	0.8790	0.9044	0.9223
Group Corr	0.7427	0.7925	0.8543	0.8904	0.9151	0.9299

Table 3: Results on `raw English Wikicorpus`. Dependency $\mathrm{R}^2(\mathbf{u}_i, \{\mathbf{u}_{-\mathbf{i}}\})$ (as defined in Eq (9)) of language models. The `Group Corr` refers to the group correlation measure $\rho(\{\mathbf{u}_i\}_{i=1}^n)$ as defined in Eq (10). T is the text sequence length. We take each model as *target* model, and apply LMD using the remaining eleven (11) models as *basis* (as defined in Eq (1)).



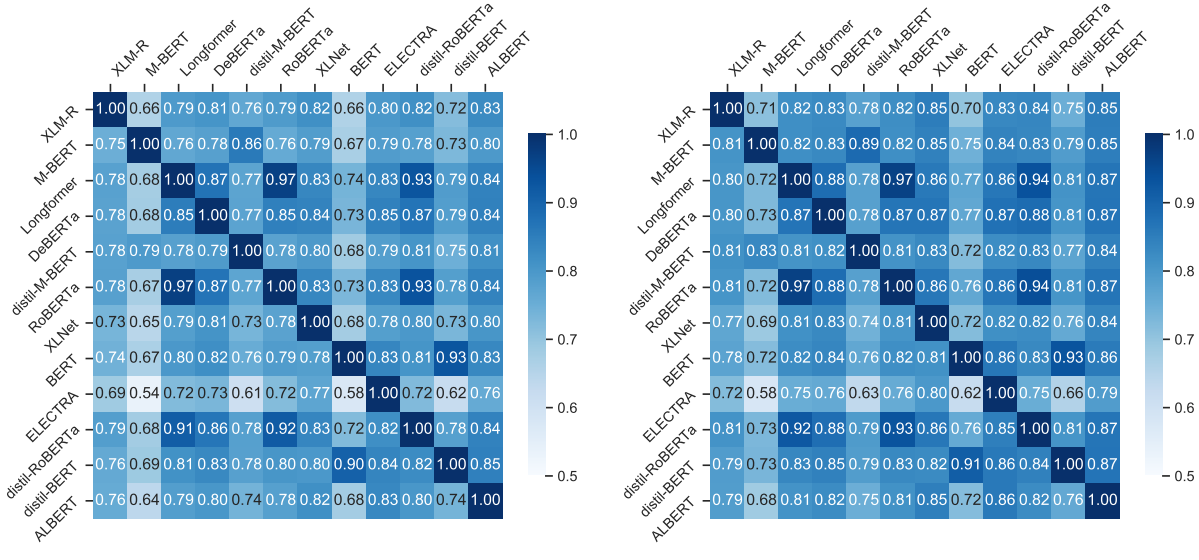
(a) $T = 16$

(b) $T = 32$



(c) $T = 64$

(d) $T = 128$



(e) $T = 256$

(f) $T = 512$

Figure 4: Results on `raw English Wikicorpus`. Asymmetric pairwise dependency $\mathrm{R}^2(\mathbf{u}_i, \mathbf{u}_j)$ (as defined in Eq (9)) of *target* LM (y-axis) on *basis* LM (x-axis).

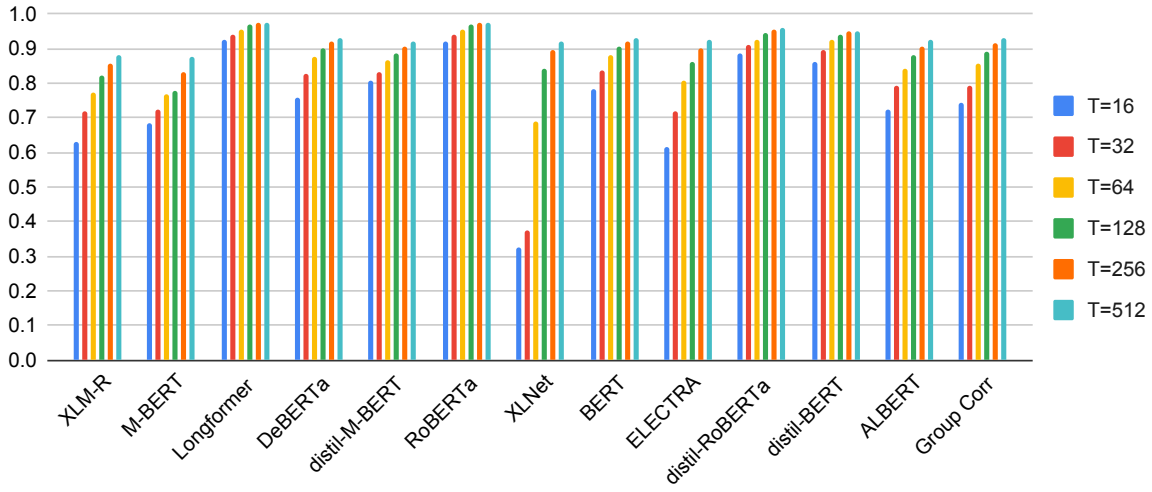


Figure 5: Results on `raw English Wikicorpus`. Dependency $\mathrm{R}^2(\mathbf{u}_i, \{\mathbf{u}_{-\mathbf{i}}\})$ (as defined in Eq (9)) of language models. The `Group Corr` refers to the group correlation measure $\rho(\{\mathbf{u}_i\}_{i=1}^n)$ (as defined in Eq (10)). T is the text sequence length. We take each model as *target* model, and apply LMD using the remaining eleven (11) models as *basis* (as defined in Eq (1)).