Bidirectional Attention as Mixture of Continuous Word Experts

Abstract

Bidirectional attention has emerged as a key component of modern large language models, which includes self-attention mechanism, position encodings, and the masked language model (MLM) objective. Despite its widespread use, few studies have examined the inductive bias underlying bidirectional attention: What sets bidirectional attention models apart from its non-attention predecessors like continuous bag of words (CBOW)? Are they completely different model classes? In this paper, we show that, upon reparameterization, bidirectional attention—namely training single-head single-layer attention with position encodings using the MLM objective—is equivalent to CBOW with mixture-of-experts weights, hence *bidirectional attention as mixture of continuous word experts*. This viewpoint enables us to characterize when embeddings from bidirectional attention exhibit a similar linear structure to its non-attention predecessors (e.g., word2vec, GloVe). It also suggests immediate extensions of bidirectional attention beyond text, including general non-textual sequence data and tabular data. Empirically, we demonstrate the attention-based approach to tabular data improves out-of-distribution generalization.

1 INTRODUCTION

Bidirectional attention has become a cornerstone in the construction of modern large language models; it involves the self-attention mechanism, position encodings, and is trained with the masked language model (MLM) objective. First introduced by Vaswani et al. [2017], the attention-based architecture represents a departure from traditional recurrent or convolutional neural networks in language modeling. This architecture has since become the backbone of many large language models; they include BERT [Devlin et al., 2018], RoBERTa [Liu et al., 2019], and GPT-2 [Radford et al., 2019], all of which have achieved exceptional performance in natural language processing benchmarks.

At the heart of bidirectional attention lies the self-attention mechanism; it creates a holistic representation of a sentence by capturing pairwise relationships between tokens in each sentence. Equally important to bidirectional attention are position encodings, providing word ordering information that allows bidirectional attention to move beyond bag-of-words. Additionally, bidirectional attention employs the MLM objective. It is a self-supervised learning objective for unlabelled text data, optimizing the model's predictive accuracy on randomly masked words within each sentence.

Despite the widespread empirical success of attention-based language models, few works have examined their underlying inductive biases. What sets bidirectional attention models apart and contributes to their success? Are they fundamentally distinct from classical language models like continuous bag of words (CBOW) [Mikolov et al., 2013]? We explore these questions in this work.

We show that bidirectional attention—i.e. training a single-layer attention-based architecture with position encodings using MLM—is equivalent to learning CBOW with mixture-of-experts (MoE) weights [Jacobs et al., 1991], upon parameterization. Following this attention-as-CBOW perspective, we investigate two key aspects of bidirectional attention: (1) whether it can exhibit linear word analogy structure like its non-attention predecessors, and (2) whether it extends to general non-textual sequence data and tabular data.

To examine the ability of bidirectional attention to elicit word analogy structure, we begin by characterizing when token embeddings from bidirectional attention can yield meaningful similarities between tokens. To do so, we draw on a classical finding in Levy and Goldberg [2014]: the similarity between two tokens from word2vec embeddings is equal to their pointwise mutual information [Church and Hanks, 1990], provided that the embeddings have sufficient dimensionality and the models were trained using the skip-gram

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

with negative sampling (SGNS) objective. This result provides a useful basis for bidirectional attention's embeddings, given their connections to CBOW. We adopt the paraphrasing argument of Allen et al. (2019) for SGNS and identify the conditions under which both CBOW and attention-based embeddings exhibit a linear structure, such as the ability to solve analogies like $king - man + woman \approx queen$.

In addition, the attention-as-CBOW perspective opens up opportunities to extend bidirectional attention beyond text, including general non-textual sequence data and tabular data. Specifically, we first develop a probabilistic formulation of attention-based architectures that extends bidirectional attention to general non-textual sequence data, building on the exponential family embedding [Rudolph et al., 2016], a probabilistic formulation of CBOW. Employing the same attention-as-CBOW perspective, we further extends attention to tabular data, where we utilize feature encoding in the place of positional encoding. This attention-based approach to tabular data improves out-of-distribution generalization, which we demonstrate in empirical studies. This approach facilitates the integration of heterogeneous datasets with partially overlapping features: the learned feature encodings bring all features into the same embedding space; the mixture-of-experts nature of bidirectional attention further licenses a direct integration of heteorogeneous data sets.

Contributions. We develop the theoretical underpinnings of bidirectional attention, revealing its equivalence to CBOW with mixture-of-experts weights. This perspective allows us to characterize the conditions under which bidirectional attention's embeddings exhibit comparable linear structure to its non-attention predecessors like CBOW. It also suggests immediate extensions of bidirectional attention to general sequence data and tabular data. Empirically, we demonstrate that the attention-based approach to tabular data can improve out-of-distribution generalization.

Related work. Our work draws on three themes around attenion-based models.

The first is a body of work on theoretical foundations of attention-based models. Many studies have looked into different aspects of this broad topic. Elhage et al. [2021] analyzed how the components of decoder-only attention-based architectures relate to each other. Edelman et al. [2022] provided a rigorous justification on the ability of attention-based architectures to represent sparse functions. Tsai et al. [2019] viewed attention through the perspective of kernels. Peng et al. [2020] built a connection between multi-head attentions and mixture of experts. In this work, we demonstrate that single-layer single-head attention-based models can be interpreted as CBOW with mixture-of-experts weights, which suggests natural extensions to non-textual data via their probabilistic formulations.

The second theme relates to linear structures in word embeddings. Neural word embeddings such as word2vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014] have been empirically shown to exhibit linear structures, often manifested through analogies. Concretely, given an analogy "a is to b as c is to d", it is usually the case that $w_b + w_c - w_a \approx w_d$, where w_i denotes the embedding for word i. Numerous work has been devoted into providing theoretical justifications for this phenomenon. Arora et al. [2016] offered a latent variable argument which assumes that texts are generated from random walks of discourse vectors and word vectors are spatially isotropic. Ethayarajh et al. [2018] introduced the *co-occurrece shifted PMI* concept which characterizes when linear analogy holds in SGNS and GloVe. Allen and Hospedales [2019] adopted the paraphrasing framework of Gittens et al. [2017] and used *word transformation* to connect linear analogy in SGNS with paraphrases. This work expands on Allen and Hospedales's [2019] ideas to explain linear structure in CBOW and attention-based embeddings.

The third theme is the application of attention-based models for tabular data. [Gorishniy et al., 2021] presented an overview of the literature on this theme. One prominent work is TabTransformer [Huang et al., 2020], which utilizes self-attention and *unique identifiers*—in lieu of position encodings—to learn contextual embeddings for categorical features. Our work is different from theirs in that we represent each feature via a "position encoding".

2 BIDIRECTIONAL ATTENTION IN THE LENS OF CONTINUOUS BAG OF WORDS

In this section, we establish the equivalence between bidirectional attention and CBOW with mixture-of-experts weights. Section 2.1 derives an explicit form of the MLM objective for a simple attention-based architecture. Section 2.2 demonstrates that the resulting objective can be viewed as CBOW with mixture-of-expert weights.

2.1 THE MASKED LANGUAGE MODEL OBJECTIVE FOR ATTENTION-BASED MODELS

Suppose we have a corpus consisting of sentences of length S. The vocabulary size of this corpus is $|V|$, and each word i in the vocabulary is represented by a token embedding $x_i \in \mathbb{R}^p$. The MLM framework involves randomly substituting some tokens in each sentence with the `[MASK]` token. This suggests a need for a separate embedding for that token, which we denote by $x_{|V|+1} \in \mathbb{R}^p$. In this paper, we consider a simple attention-based architecture trained with the MLM objective. In an attempt to build a connection with CBOW, we mask the word in each position of each sentence one at a time and aim to predict the masked word. The detail of the architecture is given in Appendix A.

Now, fix a sentence and mask the word in position $m \in [S]$, which is the b-th word in the vocabulary where $b \in [|V|]$.

Denote by $a_1, \cdots, a_S$ the words in the masked sentence, in order. Clearly, we have $a_m = |V| + 1$. Given the masked sentence, we are interested in predicting b. The following proposition gives the MLM objective for this particular instance. The proof is provided in Appendix B.

Proposition 1. *The MLM objective for the above-mentioned instance is given by*

$$-\sum_{j=1}^{S} \frac{\theta(a_j, j, m)}{\sum_{j=1}^{S} \theta(a_j, j, m)} \chi(b, a_j, j, m) + \log\left(\sum_{k=1}^{|V|} \exp\left(\sum_{j=1}^{S} \frac{\theta(a_j, j, m)}{\sum_{j=1}^{S} \theta(a_j, j, m)} \chi(k, a_j, j, m)\right)\right),$$

where $p_j \in \mathbb{R}^p$ denotes the encoding for the j-th position, $W^{KQ} \in \mathbb{R}^{p\times p}$, $W^{LOV} \in \mathbb{R}^{|V|\times p}$, $W^{\ell} \in \mathbb{R}^{|V|\times p}$, $c \in \mathbb{R}^{|V|}$, $\theta(d, e, f) := \exp\left(\frac{(x_d+p_e)^\top W^{KQ}(x_{|V|+1}+p_f)}{\sqrt{d_w}}\right)$, and $\chi(d, e, f, g) := (W_d^{LOV})^\top (x_e + p_f) + c_d + (W_d^{\ell})^\top p_g$.

In Section 2.2, we prove the equivalence between the derived objective and CBOW with mixture-of-expert weights.

2.2 BIDIRECTIONAL ATTENTION AS MIXTURE OF CONTINUOUS WORD EXPERTS

In this subsection, we establish the equivalence between the MLM objective with self-attention and the continuous bag-of-words model with mixture-of-experts weights.

We begin with the CBOW formulation of word2vec [Mikolov et al., 2013], which aims to predict the center word based on the surrounding (or context) words. CBOW has two parameter matrices $U, V \in \mathbb{R}^{|V|\times p}$, referred to as the center and context embeddings, respectively. Each row $u_i^\top$ of U represents the embedding of the word i when it is the center word, and each row $v_i^\top$ of V represents the embedding of the word i when it is the context word. Consider the same setting as before where we have a sentence with S words, and we are interested in predicting the word in position $m \in [S]$ that is the b-th word in the vocabulary. Writing $\omega_{j,m,w}$ as a shorthand for $\mathbb{1}(1 \le |j - m| \le w)$, the loss corresponding to this instance is

$$-\sum_{j=1}^{S} \frac{\omega_{j,m,w}}{\sum_{j=1}^{S} \omega_{j,m,w}} u_b^\top v_{a_j} + \log\left(\sum_{k=1}^{|V|} \exp\left(\sum_{j=1}^{S} \frac{\omega_{j,m,w}}{\sum_{j=1}^{S} \omega_{j,m,w}} u_k^\top v_{a_j}\right)\right),$$

where w represents the window size. Here, it does not matter whether $a_m = b$ or $a_m = |V| + 1$ (i.e., whether or not the m-th word is replaced with the [MASK] token). In this case, the *weight* of the word in position j is determined by the distance between j and m and the number of integers between $m - w$ and $m + w$ (inclusive) that are in the range $[1, S]$. Also, the *similarity* of word α in the center and word β in the context is $u_\alpha^\top v_\beta$, regardless of their positions.

From the same perspective, we can also interpret the weight and similarity in the MLM objective. Concretely, the weight of the word in position j is given by

$$\frac{\exp\left(\frac{(x_{a_j}+p_j)^\top W^{KQ}(x_{|V|+1}+p_m)}{\sqrt{d_w}}\right)}{\sum_{j=1}^{S} \exp\left(\frac{(x_{a_j}+p_j)^\top W^{KQ}(x_{|V|+1}+p_m)}{\sqrt{d_w}}\right)},$$

which now depends on all words in the masked sentence and their corresponding positions, but does not depend on the center (masked) word b. In fact, the term inside the $\exp(\cdot)$ can be decomposed into four components: (1) $(x_{a_j})^\top W^{KQ} x_{|V|+1}/\sqrt{d_w}$, which depends on the word in position j; (2) $(x_{a_j})^\top W^{KQ} p_m/\sqrt{d_w}$, which depends on the word in position j and the position m; (3) $(p_j)^\top W^{KQ} x_{|V|+1}/\sqrt{d_w}$, which depends on the position j; and (4) $(p_j)^\top W^{KQ} p_m/\sqrt{d_w}$, which depends on the positions j and m. Also, the similarity of word α in the center (in position m) and word β in the context (in position j) is given by $(W_\alpha^{LOV})^\top x_\beta + (W_\alpha^{LOV})^\top p_j + c_\alpha + (W_\alpha^{\ell})^\top p_m$, which also contains four components.

Clearly, other parametric forms for the weight and similarity can be used as well. For example, Sonkar et al. [2020] uses a different weight function that depends on the center word b in their attention word embedding (AWE) model.

Moreover, we can interpret MLM weights as mixture-of-experts weights in that the weight (or contribution) of each position depends on the input sentence. As before, let $X = (a_1, \cdots, a_S)$ be the masked sentence, where $a_m = |V| + 1$. Also, let $y \in \mathbb{R}^{|V|}$ be a zero vector with 1 on the b-th entry. The following proposition gives an explicit mixture-of-experts formulation of the MLM objective.

Proposition 2. *Consider a mixture of S experts, one for each position. Let m denote the position of the masked word. For every $s \in [S]$, write*

$$f_s = \big((W_1^{LOV})^\top (x_{a_s} + p_s) + c_1 + (W_1^{\ell})^\top p_m, \cdots, (W_{|V|}^{LOV})^\top (x_{a_s} + p_s) + c_{|V|} + (W_{|V|}^{\ell})^\top p_m\big)^\top.$$

We also write $h(X) = (h_1(X), \cdots, h_S(X))^\top$, where

$$h_s(X) = \frac{(x_{a_s} + p_s)^\top W^{KQ}(x_{|V|+1} + p_m)}{\sqrt{d_w}}.$$

Define $\pi(X) = softmax(h(X)) \in \mathbb{R}^S$. The MLM loss for this instance can be written as the cross-entropy loss between $F(X)$ and y, where $F(X) = \sum_s \pi_s(X) f_s$.

Here, $\pi(X)$ represents the expert weights for input X. In the case where there are no position encodings or residual connections, we have an alternative mixture-of-experts formulation which contains $|V|$ instead of S experts. More detail on this formulation is provided in Appendix C.

3 LINEAR RELATIONSHIPS FOR ANALOGIES ON ATTENTION-BASED EMBEDDINGS

In this section, we analyze between-token similarities for attention-based models and adopt the framework described in Allen and Hospedales [2019] to understand the conditions under which attention-based embeddings for analogies exhibit linear relationships. For ease of explanation, we follow their notations whenever possible.

Let n and d denote the vocabulary size and embedding dimension, respectively. Also, let $W, C \in \mathbb{R}^{d\times n}$ respectively be the center and context embedding matrix. For $1 \le i \le n$, w_i (c_i) represents the i-th column of W (C), which is the center (context) embedding of word i. Levy and Goldberg [2014] showed that for a sufficiently large embedding dimension, skip-gram with negative sampling (SGNS) solves

$$w_i^\top c_j = \log\left(\frac{p(w_i, c_j)}{p(w_i)p(c_j)}\right) - \log k = \text{PMI}(w_i, c_j) - \log k,$$

where k is the number of negative samples. Using this result, Allen and Hospedales [2019] considered embeddings which factorize the unshifted PMI matrix, namely $w_i^\top c_j = \text{PMI}(w_i, c_j)$, compactly written as $W^\top C = \text{PMI}$. Through the ideas of *paraphrases* and *word transformations*, they explained why linear relationships exist for analogies on word embeddings trained with SGNS.

CBOW-based word embeddings have been shown to also exhibit linear structures, albeit not as strong as compared SGNS-based ones [Pennington et al., 2014]. However, most explanations about the linear structures in word2vec (e.g., Ethayarajh et al. [2018], Arora et al. [2016], Allen and Hospedales [2019]) have focused on SGNS. Due to the established connection between MLM with self-attention and CBOW, we first develop some results for CBOW before moving on to MLM with self-attention.

What matrix does CBOW (approximately) factorize? The following proposition is the CBOW version of Levy and Goldberg's [2014] classical result on between-token similarities for SGNS. The proof can be found in Appendix D.

Proposition 3. *Consider CBOW without negative sampling. Using the same notation as before, we have*

$$w_i^\top c_j \approx \left(\frac{p(w_i, c_j)}{p(c_j)} \cdot n\right) - 1 \approx \log\left(\frac{p(w_i, c_j)}{p(c_j)}\right) + \log n.$$

From Proposition 3, we know that CBOW approximately factorizes M, an $n \times n$ matrix such that $M_{i,j} = \log\left(p(w_i, c_j)/p(c_j)\right) + \log n$. It is worth noting that this formula is similar to that for noise-contrastive estimation (NCE), as mentioned in Levy and Goldberg [2014], with $\log n$ replaced by $-\log k$.

We empirically verify Proposition 3 using a toy corpus with a vocabulary size of 12. This corpus consists of 10,000 sentences, each of which has length 5. The corpus generation process is detailed in Appendix E. We then train a CBOW model with the whole sentence except the center word as the context. We choose the embedding dimension to be one of $\{30, 100, 300, 900\}$. For each dimension, we compute (1) the Spearman correlation between $w_i^\top c_j$ and $p(w_i, c_j)/p(c_j)$ for each i, j; and (2) the Pearson correlation between $w_i^\top c_j$ and $\log\left(p(w_i, c_j)/p(c_j)\right) + \log n$ for each i, j such that the latter is well-defined. We obtain values of $(0.74, 0.77, 0.77, 0.77)$ for (1) and $(0.67, 0.71, 0.70, 0.71)$ for (2), which are reasonably high.

From Proposition 3, we have $w_i^\top c_j > w_k^\top c_j$ iff $p(w_i, c_j) > p(w_k, c_j)$. Without invoking the Cauchy-Schwarz inequality as in the proof of Proposition 3, we do not have a nice closed form for $w_i^\top c_j$. Instead, we can derive a system of equations concerning the $w_i^\top c_j$'s as summarized in Proposition 4. The proof can be found in Appendix F.

Proposition 4. *Consider CBOW without negative sampling. Define $\Delta(j_1, j_2)$ as the number of times j_1 and j_2 occur together in the context on the whole corpus. For example, in an instance where the center word is c and the context words are a, a, b, b and c, we have $\Delta(a, a) = 1$, $\Delta(a, b) = 4$ and $\Delta(a, c) = 2$. Let $\Delta \in \mathbb{R}^{n\times n}$ be a matrix such that $(\Delta)_{k,l} := \Delta(k, l)$, and $\square := \text{diag}(\#(c_1) + \Delta(1, 1), \cdots, \#(c_n) + \Delta(n, n))$. Assuming a sufficiently large embedding dimension, we have*

$$\begin{pmatrix} w_i^\top c_1 \\ \cdots \\ w_i^\top c_n \end{pmatrix} \approx 2m(\Delta + \square)^{-1} \begin{pmatrix} n\#(w_i, c_1) + \#(c_1) \\ \cdots \\ n\#(w_i, c_n) + \#(c_n) \end{pmatrix},$$

where $2m$ is the window size, $\#(w_k, c_l)$ represents the number of center-context pairs with k as the center word and l as the context word, and $\#(c_l) = \sum_k \#(w_k, c_l)$.

The paraphrasing argument for CBOW. We look at what it means for two word sets to paraphrase each other.

Definition 5 (Definition D2 of [Allen and Hospedales, 2019])**.** *Let $\mathcal{E}$ be the set of all words in the vocabulary. Two word sets $\mathcal{W}, \mathcal{W}_* \subseteq \mathcal{E}$ are said to paraphrase each other if the paraphrase error $\rho^{\mathcal{W},\mathcal{W}_*} \in \mathbb{R}^n$ is element-wise small, where*

$$\rho_j^{\mathcal{W},\mathcal{W}_*} = \log\left(\frac{p(c_j|\mathcal{W}_*)}{p(c_j|\mathcal{W})}\right)$$

for every $c_j \in \mathcal{E}$.

Intuitively, "word sets paraphrase one another if they induce equivalent distributions over context words". When $\mathcal{W}$ and $\mathcal{W}_*$ paraphrase each other, we write $\mathcal{W} \approx_P \mathcal{W}_*$. From Definition 5, it is easy to see that $\mathcal{W} \approx_P \mathcal{W}_*$ iff $\mathcal{W}_* \approx_P \mathcal{W}$. Also, implicit in this definition is that we need both $p(\mathcal{W}_*)$ and $p(\mathcal{W})$ to be positive. This is exactly Assumption A3 in the original paper. We now provide an equivalent version of

their Lemma 2 for the matrix M. Here, M_i denotes the i-th column of M.

Lemma 6. *For any word sets $\mathcal{W}, \mathcal{W}_* \subseteq \mathcal{E}$ with the same cardinality, we have*

$$\sum_{w_i \in \mathcal{W}_*} M_i = \sum_{w_i \in \mathcal{W}} M_i + \rho^{\mathcal{W},\mathcal{W}_*} + \sigma^{\mathcal{W}} - \sigma^{\mathcal{W}_*} + \delta^{\mathcal{W},\mathcal{W}_*}$$
$$= \sum_{w_i \in \mathcal{W}} M_i + \xi^{\mathcal{W},\mathcal{W}_*} + \sigma^{\mathcal{W}} - \sigma^{\mathcal{W}_*},$$

where $\sigma_j^{\mathcal{W}} = \log\left(\frac{p(\mathcal{W}|c_j)}{\prod_i p(w_i|c_j)}\right)$, $\sigma_j^{\mathcal{W}_*} = \log\left(\frac{p(\mathcal{W}_*|c_j)}{\prod_i p(w_i|c_j)}\right)$, $\delta^{\mathcal{W},\mathcal{W}_*} = \log\left(\frac{p(\mathcal{W}_*)}{p(\mathcal{W})}\right)$, and $\xi_j^{\mathcal{W},\mathcal{W}_*} = \log\left(\frac{p(\mathcal{W}_*|c_j)}{p(\mathcal{W}|c_j)}\right)$.

The proof is similar to that of Lemma 2 in the original paper, and therefore omitted.

The following proposition, which is equivalent to Corollary 2.3 of Allen and Hospedales [2019], follows from multiplying both sides of the equations in Lemma 6 by $C^\dagger = (CC^\top)^{-1}C$ (assuming C has full row rank) and taking $\mathcal{W}$ and $\mathcal{W}_*$ to be sets of cardinality 2.

Proposition 7. *Given any $w_a, w_{a^*}, w_b, w_{b^*} \in \mathcal{E}$, we have*

$$w_{b^*}$$
$$= w_{a^*} - w_a + w_b + C^\dagger(\rho^{\mathcal{W},\mathcal{W}_*} + \sigma^{\mathcal{W}} - \sigma^{\mathcal{W}_*} + \delta^{\mathcal{W},\mathcal{W}_*})$$
$$= w_{a^*} - w_a + w_b + C^\dagger(\xi^{\mathcal{W},\mathcal{W}_*} + \sigma^{\mathcal{W}} - \sigma^{\mathcal{W}_*}),$$

where $\mathcal{W} = \{w_b, w_{a^*}\}$ and $\mathcal{W}_* = \{w_{b^*}, w_a\}$.

From Proposition 7, we see that when $\mathcal{W} \approx_P \mathcal{W}_*$, and $\sigma^{\mathcal{W}}$, $\sigma^{\mathcal{W}_*}$ and $\delta^{\mathcal{W},\mathcal{W}_*}$ are small, we have $w_{b^*} \approx w_{a^*} - w_a + w_b$. By definition, $\sigma^{\mathcal{W}}$ ($\sigma^{\mathcal{W}_*}$) is small when all $w_i \in \mathcal{W}$ ($w_i \in \mathcal{W}_*$) are approximately conditionally independent given c_j, and $\delta^{\mathcal{W},\mathcal{W}_*}$ is small when $p(\mathcal{W}) \approx p(\mathcal{W}_*)$. Following the connection between analogies and word transformations described in Sections 6.3 and 6.4 of Allen and Hospedales [2019], we now have an approximately linear relationship for CBOW embeddings with some error terms as mentioned above. Alternatively, we can modify Definition 5 so that $\mathcal{W} \approx_P \mathcal{W}_*$ iff $\xi^{\mathcal{W},\mathcal{W}_*}$ (instead of $\rho^{\mathcal{W},\mathcal{W}_*}$) is element-wise small. In this case, our error terms only depend on the approximate conditional independence of w_i's given c_j.

Does this linear relationship also hold for context embeddings? In other words, if $w_r + w_s \approx w_t + w_u$, do we have $c_r + c_s \approx c_t + c_u$? The following proposition, whose proof is provided in Appendix G, answers the question.

Proposition 8. *Let $\mathcal{W} = \{r, s\}$ and $\mathcal{W}_* = \{t, u\}$. Assume $p(\mathcal{W}) \approx p(\mathcal{W}_*)$ and $w_i \in \mathcal{W}$ ($w_i \in \mathcal{W}_*$) are approximately marginally independent. Also, assume that W has full row rank. If $w_r + w_s \approx w_t + w_u$, then $c_r + c_s \approx c_t + c_u$.*

So far, we have argued that both the center and context embeddings of CBOW exhibit a linear structure under some assumptions. We now extend this argument to MLM with self-attention, and show that the same conclusion holds under stronger assumptions.

What matrix does MLM with self-attention (approximately) factorize? For simplicity, we exclude both residual connections and position encodings in our discussion. We focus on the following instance: the masked sentence is $(a_1, \cdots, a_S)$, where $a_i \in \{1, 2, \cdots, n\}$ for every $i \neq m$ and $a_m = n + 1$ (the masked token); before masking, we have $a_m = b \in \{1, 2, \cdots, n\}$. Using the same notations as before, the loss for this instance is given by

$$-\sum_{j=1}^{S} \frac{\tau_{a_j}}{\sum_{j=1}^{S} \tau_{a_j}} (w_b)^\top (c_{a_j})$$
$$+ \log\left(\sum_{k=1}^{n} \exp\left(\sum_{j=1}^{S} \frac{\tau_{a_j}}{\sum_{j=1}^{S} \tau_{a_j}} (w_k)^\top (c_{a_j})\right)\right),$$

where $\tau_j := \exp\left(\frac{c_j^\top W^{KQ} c_{n+1}}{\sqrt{d_w}}\right)$. The following proposition approximates the matrix factorized by the attention objective, given all τ_j values for each j. The proof is similar to that of Proposition 3, and therefore omitted.

Proposition 9. *Consider the attention objective as above. Using the same approximation technique as used in Proposition 3, we have*

$$w_i^\top c_j \approx \frac{n \sum_{(i,j)} \gamma_j^i - \left(\sum_{(1,j)} \gamma_j^1 + \cdots + \sum_{(n,j)} \gamma_j^n\right)}{S\left(\sum_{(1,j)} (\gamma_j^1)^2 + \cdots + \sum_{(n,j)} (\gamma_j^n)^2\right)}, \tag{1}$$

where for a center-context pair (d, j) in a masked sentence $(a_1, \cdots, a_S)$, we define $\gamma_j^d := \tau_j / \sum_{s=1}^{S} \tau_{a_s}$.

In other words, MLM with self-attention approximately factorizes an $n \times n$ matrix whose (i, j)-th entry is given by Equation (1). It is worth noting that unlike in CBOW, the embedding for each word i is c_i (the "context" embedding), and not w_i (the "center" embedding).

In the case where τ_j is approximately the same for every j, our problem approximately reduces to a vanilla CBOW. In particular, we have $\gamma_j^d \approx 1/S$ for every j, d and masked sentence $(a_1, \cdots, a_S)$, whence Proposition 9 yields $w_i^\top c_j \approx \left(\frac{p(w_i, c_j)}{p(c_j)} \cdot n\right) - 1$. From Proposition 7, we have that the resulting embeddings approximately form a linear relationship, up to some error terms.

The paraphrasing argument for MLM with self-attention. To analyze the more general case, we first define

$$\tilde{c}_j := \frac{S\left(\sum_{(1,j)} (\gamma_j^1)^2 + \cdots + \sum_{(n,j)} (\gamma_j^n)^2\right)}{\sum_{(1,j)} \gamma_j^1 + \cdots + \sum_{(n,j)} \gamma_j^n} c_j$$

for every j. This means

$$
\begin{aligned}
& w_i^\top \tilde{c}_j \\
& \approx \frac{n \sum_{(i,j)} \gamma_j^i}{\sum_{(1,j)} \gamma_j^1 + \cdots + \sum_{(n,j)} \gamma_j^n} - 1 \\
& \approx \log \left(\frac{\sum_{(i,j)} \gamma_j^i}{\sum_{(1,j)} \gamma_j^1 + \cdots + \sum_{(n,j)} \gamma_j^n} \right) + \log n,
\end{aligned}
$$

where we used the approximation $x \approx \log(1+x)$. Previously, $p(w_i, c_j)$ represents a population quantity which is estimated by $\#(w_i, c_j)/D$, where D is a normalizing constant, and $p(c_j) = \sum_i p(w_i, c_j)$. We now define $\bar{p}(w_i, c_j)$, a population quantity which is estimated by $\sum_{(i,j)} \gamma_j^i / E$ for some normalizing constant E. We have

$$
w_i^\top \tilde{c}_j \approx \log \left(\frac{\bar{p}(w_i, c_j)}{\bar{p}(c_j)} \right) + \log n, \tag{2}
$$

where $\bar{p}(c_j) = \sum_i \bar{p}(w_i, c_j)$. Unlike p, $\bar{p}$ is not symmetric, i.e., $\bar{p}(w_i, c_j) \neq \bar{p}(w_j, c_i)$.

Having defined $\bar{p}$, we are ready to state the following lemma, which is a version of Lemma 6 for the matrix N, where $N_{i,j} = \log \left(\frac{\bar{p}(w_i, c_j)}{\bar{p}(c_j)} \right) + \log n$.

Lemma 10. *For any word sets $\mathcal{W}, \mathcal{W}_* \subseteq \mathcal{E}$ with the same cardinality, we have*

$$
\begin{aligned}
\sum_{w_i \in \mathcal{W}_*} N_i &= \sum_{w_i \in \mathcal{W}} N_i + \bar{\rho}^{\mathcal{W}, \mathcal{W}_*} + \bar{\sigma}^{\mathcal{W}} - \bar{\sigma}^{\mathcal{W}_*} + \bar{\delta}^{\mathcal{W}, \mathcal{W}_*} \\
&= \sum_{w_i \in \mathcal{W}} N_i + \bar{\xi}^{\mathcal{W}, \mathcal{W}_*} + \bar{\sigma}^{\mathcal{W}} - \bar{\sigma}^{\mathcal{W}_*},
\end{aligned}
$$

where $\bar{\rho}_j^{\mathcal{W}, \mathcal{W}_*} = \log \left(\frac{\bar{p}(c_j | \mathcal{W}_*)}{\bar{p}(c_j | \mathcal{W})} \right)$, $\bar{\sigma}_j^{\mathcal{W}} = \log \left(\frac{\bar{p}(\mathcal{W} | c_j)}{\prod_i \bar{p}(w_i | c_j)} \right)$, $\bar{\sigma}_j^{\mathcal{W}_*} = \log \left(\frac{\bar{p}(\mathcal{W}_* | c_j)}{\prod_i \bar{p}(w_i | c_j)} \right)$, $\bar{\delta}^{\mathcal{W}, \mathcal{W}_*} = \log \left(\frac{\bar{p}(\mathcal{W}_*)}{\bar{p}(\mathcal{W})} \right)$, *and* $\bar{\xi}_j^{\mathcal{W}, \mathcal{W}_*} = \log \left(\frac{\bar{p}(\mathcal{W}_* | c_j)}{\bar{p}(\mathcal{W} | c_j)} \right)$.

The following propositions are the attention versions of Propositions 7 and 8.

Proposition 11. *Given any $w_a, w_{a^*}, w_b, w_{b^*} \in \mathcal{E}$, we have*

$$
\begin{aligned}
& w_{b^*} \\
& = w_{a^*} - w_a + w_b + \tilde{C}^\dagger (\bar{\rho}^{\mathcal{W}, \mathcal{W}_*} + \bar{\sigma}^{\mathcal{W}} - \sigma^{\mathcal{W}_*} + \bar{\delta}^{\mathcal{W}, \mathcal{W}_*}) \\
& = w_{a^*} - w_a + w_b + \tilde{C}^\dagger (\bar{\xi}^{\mathcal{W}, \mathcal{W}_*} + \bar{\sigma}^{\mathcal{W}} - \bar{\sigma}^{\mathcal{W}_*}),
\end{aligned}
$$

where $\mathcal{W} = \{w_b, w_{a^*}\}$, $\mathcal{W}_* = \{w_{b^*}, w_a\}$ and $\tilde{C}^\dagger = (\tilde{C}\tilde{C}^\top)^{-1} C$, assuming $\tilde{C}$ has a full row rank.

Proposition 12. *Let $\mathcal{W} = \{r, s\}$ and $\mathcal{W}_* = \{t, u\}$. Assume $\bar{p}(\mathcal{W}) \approx \bar{p}(\mathcal{W}_*)$ and $w_i \in \mathcal{W}$ ($w_i \in \mathcal{W}_*$) are approximately marginally independent. Also, assume that W has full row rank and $\bar{p}(w_i, c_j) \approx \bar{p}(w_j, c_i)$. If $w_r + w_s \approx w_t + w_u$, then $\tilde{c}_r + \tilde{c}_s \approx \tilde{c}_t + \tilde{c}_u$.*

The proof can be found in Appendix H.

What do we learn from these results? One important takeaway is that the sufficient conditions to obtain linear relationships are stronger in the case of MLM with self-attention as compared to CBOW. Concretely, we need $\bar{p}$ to be approximately symmetric. Even when the conditions are satisfied, the linear relationships hold for the transformed embeddings $\tilde{c}_i$'s instead of the raw (token) embeddings c_i's.

In the case where

$$
\zeta_j := \frac{\sum_{(1,j)} (\gamma_j^1)^2 + \cdots + \sum_{(n,j)} (\gamma_j^n)^2}{\sum_{(1,j)} \gamma_j^1 + \cdots + \sum_{(n,j)} \gamma_j^n}
$$

is approximately the same for each j (e.g., when τ_j is approximately the same for every j), we approximately have linear relationships for the raw embeddings c_i's. Despite these technicalities, we show in Section 3.1 that BERT embeddings [Devlin et al., 2018], which are trained via MLM with self-attention, empirically exhibit linear relationships and outperform both CBOW and GloVe [Pennington et al., 2014] on syntactic analogy tasks.

3.1 EMPIRICAL STUDY: LINEAR RELATIONSHIPS IN ATTENTION-BASED EMBEDDINGS

So far, we have established sufficient conditions under which attention embeddings display approximate linear relationships on analogies. We now perform an empirical study to evaluate whether such linear relationships are observed in BERT [Devlin et al., 2018], a large language model which utilizes self-attention as its main building block.

Data description. We use the analogy data set first introduced in Pennington et al. [2014]. This data set contains 19,544 questions of the form "*a* is to *b* as *c* is to ?", together with the correct answers. As an example, the first question in the data set is "Athens is to Greece as Baghdad is to ?" (correct answer: Iraq). Overall, these questions can be categorized into two groups: semantic (about people and places) and syntactic (about word forms such as comparative, superlative and plural). For each question, we look for the word $d \neq a, b, c$ in the vocabulary such that the cosine similarity between x_d and $x_b + x_c - x_a$ is maximized. Here, x_i represents the embedding of word i.

To include a question, all words involved must be present in the vocabularies of all models. Out of 19,544 questions in the data set, 10,171 (52%) of them satisfy the condition. Analyzing each category separately, we find that the condition is satisfied for 2,612 (29%) out of 8,869 semantic questions, and 7,559 (71%) out of 10,675 syntactic questions.

Competing methods. We consider three models: (1) BERT model uncased, which is used in the original BERT paper [Devlin et al., 2018]; (2) GloVe trained on Wikipedia; (3) word2vec trained with CBOW. The embedding dimensions

Table 1: BERT achieves the highest accuracy overall and on syntactic analogies, where accuracy is defined as the proportion of correct answers.

	BERT	GloVe	word2vec
Semantic	0.633	**0.754**	0.236
Syntactic	**0.752**	0.690	0.654
Overall	**0.722**	0.706	0.547

Table 2: BERT achieves the highest average cosine similarity on syntactic analogies.

	BERT	GloVe	word2vec
Semantic	0.500	**0.598**	0.505
Syntactic	**0.610**	0.608	0.579
Overall	0.581	**0.606**	0.560

of these models are 768, 300 and 300, respectively, while the vocabulary size are around 30K, 400K and 3M, respectively. Since all questions in the data set consist of single words (e.g., not *golden_retriever*). In order to perform a fair comparison among these models, we only consider single words as possible answers to each question. With the same logic, we exclude non-words (e.g., *[unused9]*, *## ?*) from the list of possible answers.

Evaluation metrics. For each model, we are interested in (1) the overall and per-category accuracies, where accuracy is defined as the proportion of correct answers; and (2) the overall and per-category average cosine similarity between $x_b + x_c - x_a$ and the correct answer.

Results. The accuracy and average cosine similarity results are displayed in Tables 1 and 2. We observe that BERT and GloVe generally outperform word2vec in solving linear analogy tasks. Furthermore, BERT achieves the best performance on syntactic questions, while GloVe achieves the best performance on semantic questions.

4 BIDIRECTIONAL ATTENTION BEYOND TEXT

The equivalence between MLM with self-attention and CBOW with mixture-of-experts weights yield some methodological insights: (1) it enables us to integrate probabilistic models into attention-based models, thanks to the probabilistic formmulation of CBOW—namely, exponential family embeddings [Rudolph et al., 2016]; and (2) it allows for immediate generalization of attention-based architectures to model tabular data. In this section, we explore both ideas in more detail.

4.1 A PROBABILISTIC VIEW OF ATTENTION-BASED MODELS

Rudolph et al. [2016] proposed exponential family embeddings, a probabilistic framework of which CBOW is a special case. A concise summary of this framework is given in Appendix I. We demonstrate that the MLM objective in Proposition 1 can also be framed in terms of exponential family embeddings. Let our data (sentence) be $x = x_{1:I}$, where $x_i \in \mathbb{R}^{D+1}$ due to the introduction of the `[MASK]` token. Consider an instance in which the i-th word in the sentence is masked, i.e., $x_i = (0, \cdots, 0, 1)^\top$. In MLM, the whole masked sentence is the context, i.e., $c_i = [I]$. Let $\bar{x}_i \in \mathbb{R}^D$ be the one-hot encoding of the masked word and K be the embedding size. Also,

$$\eta_i := \sum_{j=1}^{S} \frac{\exp\left(\frac{(\alpha x_j + p_j)^\top W^{KQ}(\alpha x_i + p_i)}{\sqrt{d_w}}\right)}{\sum_{j=1}^{S} \exp\left(\frac{(\alpha x_j + p_j)^\top W^{KQ}(\alpha x_i + p_i)}{\sqrt{d_w}}\right)} \left(W^{LOV}(\alpha x_j + p_j) + c + W^{\ell} p_i\right),$$

and $p(\tilde{x}_i | x_{c_i}) \sim \text{Cat}(\eta_i)$, where

1. $\alpha \in \mathbb{R}^{K \times (D+1)}$ is the (context) embedding matrix, where the t-th column represents the (context) embedding of the t-th word in the vocabulary.
2. $p \in \mathbb{R}^{K \times I}$ is the position encoding matrix, where the t-th column represents the embedding of the t-th position.
3. $W^{LOV} \in \mathbb{R}^{D \times K}$ is the (center) embedding matrix, where the t-th row represents the (center) embedding of the t-th word in the vocabulary.
4. $W^{\ell} \in \mathbb{R}^{D \times K}$, where the t-th row represents the embedding of the t-th word in the vocabulary which interacts with the position of the masked word.
5. $W^{KQ} \in \mathbb{R}^{K \times K}$ and $c \in \mathbb{R}^D$ (similar to a bias term).

A potential application of this probabilistic formulation is to model user-item rating data in the presence of sequential item consumption information per user, whose detail is deferred to Appendix J.

4.2 ATTENTION-BASED MODEL FOR TABULAR DATA

We consider a classification problem in which the response variable Y_i is ordinal with C classes. For simplicity, assume each of the K-dimensional features X_i is also ordinal with C classes. In this subsection, we propose an attention-based method for predicting Y_i given X_i.

We first introduce "word" embeddings $w_1, \cdots, w_C \in \mathbb{R}^d$ for each class and w_0 for the `[MASK]` token. We also introduce "position" embeddings $p_1, p_2, \cdots, p_{K+1} \in \mathbb{R}^d$, one for each feature. We treat each data point (X_i, Y_i) as a sentence and learn the embeddings using the MLM objective. At prediction time, given X_i, the model returns the most probable class for the input $(X_i, \texttt{[MASK]})$. The main power of this method lies in the modeling of $p(X, Y)$ instead of $p(Y|X)$, which is what classification methods usually model. We believe that this will result in better out-of-distribution generalizations. Some other benefits of this model are:

1. It can be readily extended for unsupervised and semi-supervised learning settings.
   - For unsupervised learning, we can proceed in the same manner as in the supervised case by ignoring the Y_i's.
   - For semi-supervised learning, we can use both the labeled and unlabeled data after setting the Y_i's for the unlabeled data to `[MASK]`.
2. The learned feature embeddings can help us understand the relationships between different features across different data sets.
3. The learned feature embeddings allow us to simultaneously learn from multiple data sets by sharing the same embedding for the same feature across all data sets.

In Section 4.3, we empirically verify that the proposed model is indeed able to achieve better out-of-distribution generalizations.

4.3 EMPIRICAL STUDY: BIDIRECTIONAL ATTENTION FOR OUT-OF-DISTRIBUTION GENERALIZATION

In this subsection, we apply the method described in Section 4.2 to simulated data sets which exhibit covariate shift, i.e., $p(X_{\text{train}}) \neq p(X_{\text{test}})$ and $p(Y_{\text{train}}|X_{\text{train}}) = p(Y_{\text{test}}|X_{\text{test}})$. Each of the data sets contains 2,000 training and 2,000 test data points, with number of features $K = 5$. The embeddings are trained using a single-layer BERT [Devlin et al., 2018] with 5 heads. The embedding and feed-forward dimensions are 20 and 5, respectively.

Data generation. We briefly describe the data generation process (for a detailed description, see Appendix K). Five data sets are generated for each set of hyperparameters: (1) $n_c \in \{1, 5\}$, the number of features which generate Y; (2) corr $\in \{0.1, 0.5, 0.9\}$, where a larger value indicates a larger between-feature correlation in the training set as compared to the test set; (3) noise $\in \{1.5, 3.5\}$, where a larger value indicates a larger noise in the observed features; and (4) $C \in \{10, 50\}$, the number of classes.

Baseline model and evaluation metrics. We fit the proposed attention-based model and a logistic regression model (as a baseline) to each training set and evaluate the prediction accuracy and mean absolute error (MAE) on the corresponding test set. For each set of hyperparameters, we take the average of both metrics across the five data sets.

Results. Table 3 summarizes the experiment results for each set of hyperparameters, where each value is an average over the five replications. Across all hyperparameter settings, we generally find that the proposed attention-based model outperforms logistic regression, as evidenced by smaller test MAEs and larger test accuracies.

Table 3: The proposed attention-based model (Attn) achieves superior performance as compared to logistic regression (LR) in virtually all hyperparameter settings. Numbers written in boldface correspond to the best-performing model (small for MAE and large for accuracy).

	Test MAE		Test accurary	
$(n_c, \text{corr}, \text{noise}, C)$	Attn	LR	Attn	LR
(1, 0.1, 1.5, 10)	1.257	**1.231**	**0.387**	0.379
(1, 0.1, 1.5, 50)	**9.681**	10.136	**0.064**	0.062
(1, 0.1, 3.5, 10)	**1.931**	1.939	**0.251**	0.227
(1, 0.1, 3.5, 50)	**10.423**	11.200	**0.054**	0.050
(1, 0.5, 1.5, 10)	**1.604**	1.613	**0.278**	0.276
(1, 0.5, 1.5, 50)	**6.147**	9.331	**0.080**	0.062
(1, 0.5, 3.5, 10)	**1.586**	1.622	**0.310**	0.289
(1, 0.5, 3.5, 50)	**9.718**	11.551	**0.068**	0.054
(1, 0.9, 1.5, 10)	**1.487**	1.858	**0.306**	0.238
(1, 0.9, 1.5, 50)	**8.521**	11.429	**0.078**	0.053
(1, 0.9, 3.5, 10)	**2.141**	2.285	**0.208**	0.185
(1, 0.9, 3.5, 50)	**14.379**	14.921	**0.033**	0.025
(5, 0.1, 1.5, 10)	**0.530**	0.832	**0.580**	0.371
(5, 0.1, 1.5, 50)	**3.357**	8.174	**0.161**	0.030
(5, 0.1, 3.5, 10)	**0.779**	1.096	**0.532**	0.313
(5, 0.1, 3.5, 50)	**4.175**	9.095	**0.139**	0.032
(5, 0.5, 1.5, 10)	**0.604**	1.078	**0.535**	0.242
(5, 0.5, 1.5, 50)	**3.000**	9.661	**0.173**	0.029
(5, 0.5, 3.5, 10)	**0.932**	1.380	**0.462**	0.226
(5, 0.5, 3.5, 50)	**5.686**	10.853	**0.121**	0.024
(5, 0.9, 1.5, 10)	**1.055**	2.030	**0.335**	0.132
(5, 0.9, 1.5, 50)	**6.362**	13.350	**0.054**	0.019
(5, 0.9, 3.5, 10)	**1.516**	2.168	**0.239**	0.126
(5, 0.9, 3.5, 50)	**9.248**	12.693	**0.042**	0.029

5 DISCUSSION

In this paper, we study we focus on bidirectional attention, namely, training a single-layer attention-based architecture with the MLM objective. We prove bidirectional attention is equivalent to fitting a continuous bag of words with mixture-of-experts weights, given appropriate reparameterization. This perspective, which we call *bidirectional attention as mixture of continuous word experts*, allows us to identify the conditions under which embeddings from bidirectional attention exhibit linear structure on word analogies. Additionally, it suggests immediate extensions of attention to general non-textual sequence data and tabular data, leading to improved out-of-distribution generalization compared to existing approaches. Moving forward, an interesting avenue for future research would be to explore the statistical properties of estimating attention-based models and their implications for in-context learning.

References

C. Allen and T. Hospedales. Analogies explained: Towards understanding word embeddings, 2019. URL https://arxiv.org/abs/1901.09813.

S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016. doi: 10.1162/tacl_a_00106. URL https://aclanthology.org/Q16-1028.

K. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

B. L. Edelman, S. Goel, S. M. Kakade, and C. Zhang. Inductive biases and variable creation in self-attention mechanisms, 2022. URL https://openreview.net/forum?id=UjynxfqnGWG.

N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

K. Ethayarajh, D. Duvenaud, and G. Hirst. Towards understanding linear word analogies. *CoRR*, abs/1810.04882, 2018. URL http://arxiv.org/abs/1810.04882.

A. Gittens, D. Achlioptas, and M. W. Mahoney. Skip-gram Zipf + uniform = vector additivity. In *Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko. Revisiting deep learning models for tabular data. *CoRR*, abs/2106.11959, 2021. URL https://arxiv.org/abs/2106.11959.

X. Huang, A. Khetan, M. Cvitkovic, and Z. S. Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *CoRR*, abs/2012.06678, 2020. URL https://arxiv.org/abs/2012.06678.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc., 2014.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

H. Peng, R. Schwartz, D. Li, and N. A. Smith. A mixture of h - 1 heads is better than h heads. In *Meeting of the Association for Computational Linguistics*, pages 6566–6577, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.587. URL https://aclanthology.org/2020.acl-main.587.

J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

M. Rudolph, F. Ruiz, S. Mandt, and D. Blei. Exponential family embeddings. *Advances in Neural Information Processing Systems*, 29, 2016.

S. Sonkar, A. E. Waters, and R. G. Baraniuk. Attention word embedding. *CoRR*, abs/2006.00988, 2020. URL https://arxiv.org/abs/2006.00988.

Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4344–4353, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1443. URL https://aclanthology.org/D19-1443.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.