Bidirectional Attention as a Mixture of Continuous Word Experts

Kevin Christian Wibisono[1] **Yixin Wang**[1]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI, USA *

Abstract

Bidirectional attention—composed of the neural network architecture of self-attention with positional encodings, together with the masked language model (MLM) objective—has emerged as a key component of modern large language models (LLMS). Despite its empirical success, few studies have examined its statistical underpinnings: What statistical model is bidirectional attention implicitly fitting? What sets it apart from its non-attention predecessors? In this paper, we study these questions by theoretically analyzing bidirectional attention. The key observation is that fitting a single-layer single-head bidirectional attention, upon reparameterization, is equivalent to fitting a continuous bag of words (CBOW) model with mixture-of-experts (MOE) weights. Further, bidirectional attention with multiple heads and multiple layers are equivalent to stacked MOES and mixture of MOES, respectively. This statistical viewpoint reveals the distinct use of MOE in bidirectional attention, which aligns with its practical effectiveness in handling heterogeneous data. It also suggests an immediate extension to categorical tabular data, if we view each word location in a sentence as a tabular feature. Across empirical studies, we find that this extension outperforms existing tabular extensions of transformers in out-of-distribution (OOD) generalization. Finally, this statistical perspective of bidirectional attention enables us to theoretically characterize when linear word analogies are present in its word embeddings. These analyses show that bidirectional attention can require much stronger assumptions to exhibit linear word analogies than its non-attention predecessors.

*Correspondence to: {kwib,yixinw}@umich.edu; Software that replicates the empirical studies is at `https://github.com/yixinw-lab/attention-uai`.

1 INTRODUCTION

Bidirectional attention has recently emerged as a cornerstone in the construction of large language models (LLMS). It is composed of the self-attention mechanism with positional 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 the traditional recurrent or convolutional neural networks in language modeling. This architecture has since become the backbone of many large language models, including BERT [Devlin et al., 2018], RoBERTa [Liu et al., 2019], and GPT-2 [Radford et al., 2019]; all of them 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 positional encodings, supplying word ordering information that allows bidirectional attention to move beyond bag-of-words. Finally, 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 empirical success of attention-based language models, few works have examined their statistical underpinnings: What statistical models are these attention-based models implicitly fitting? What sets these models apart from their non-attention predecessors like continuous bag of words (CBOW) [Mikolov et al., 2013]? How does the use of self-attention mechanism contribute to their empirical success? We explore these questions in this work.

Main idea. We theoretically study bidirectional attention, i.e., a self-attention module with positional encodings trained using MLM. The key observation is that fitting a single-head and single-layer bidirectional attention, upon reparametrization, is equivalent to fitting CBOW word with mixture-of-

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

experts (MOE) weights [Jacobs et al., 1991]. Moreover, bidirectional attention with multiple heads and multiple layers are equivalent to stacked MOEs and mixture of MOEs, respectively. These analyses reveal the distinct use of MOE in bidirectional attention as compared with its non-attention predecessor; they partially explain its practical effectiveness in capturing heterogeneous patterns in natural language [Devlin et al., 2018, Liu et al., 2019].

This statistical interpretation of bidirectional attention suggests an immediate extension of bidirectional attention to (categorical) tabular data: one can view each word location in a sentence as a tabular feature, and each word as the value that the feature takes. Across empirical studies, we find that this tabular extension of attention improves out-of-distribution (OOD) generalization, compared with existing tabular data algorithms or tabular extensions of attention. Moreover, this tabular extension of attention facilitates the integration of heterogeneous datasets with partially overlapping features; the learned feature encodings (akin to the positional encodings in the original attention module) bring all features into the same embedding space.

Finally, this connection between bidirectional attention and CBOW+MOE empowers us to theoretically characterize when linear word analogies (e.g. $\text{king} - \text{man} + \text{woman} \approx \text{queen}$) can be present in its word embeddings. 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 with negative sampling (SGNS) objective. This result enables us to analyze the embeddings of bidirectional attention, given their connections to CBOW. Adopting the paraphrasing argument of Allen and Hospedales [2019] for SGNS, we characterize the conditions under which both CBOW and attention-based embeddings exhibit linear word analogies. We show that bidirectional attention can require much stronger assumptions to exhibit linear word analogies than its non-attention predecessors. These results partially explain the empirical observations that bidirectional attention may not always achieve meaningful improvements over classical word embeddings in capturing abstract and complex relationships [Ushio et al., 2021].

Contributions. We prove that bidirectional attention, upon reparametrization, is equivalent to CBOW with MOE weights. Moreover, bidirectional attention with multiple heads and multiple layers are equivalent to stacked MoEs and mixture of MoEs, respectively. This statistical interpretation with MOE partially explains the power of bidirectional attention in handling heterogeneous data. Further, it suggests an immediate extension of bidirectional attention to categorical tabular data. Across empirical studies, it outperforms existing tabular algorithms or tabular extensions of attention in OOD generalization. Finally, we leverage this statistical perspective of attention to characterize the presence of linear word analogies in bidirectional attention's embeddings. We show that it can require much stronger assumptions to exhibit linear word analogies than its non-attention predecessors. These results align with the empirical observations that bidirectional attention can sometimes perform worse n complex analogy tasks than classical word embeddings.

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

The first is a body of work on the theoretical foundations of attention-based models. Elhage et al. [2021] analyzed how the different components of decoder-only attention-based architectures relate to each other. Edelman et al. [2022] provided a rigorous justification of 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] established a connection between the use of multiple heads in transformers and MOE. Li et al. [2023] showed that the embedding and self-attention layers in a transformer architecture are capable of capturing topic structures. Bai et al. [2023], Bietti et al. [2023], Xie et al. [2022], Han et al. [2023] provided theoretical analyses about the in-context learning ability of attention-based models. In contrast to these works, we provide a statistical interpretation of the bidirectional attention objective, showing that fitting a single-layer single-head attention-based architectures is equivalent to fitting a CBOW model with MOE weights; this statistical interpretation provides a theoretical basis for the empirical effectiveness of bidirectional attention in handling heterogeneous data [Devlin et al., 2018, Liu et al., 2019].

The second theme is the extension of attention-based models to tabular data. One prominent work along this line is TabTransformer [Huang et al., 2020], which utilizes a concatenation of token embeddings and unique feature identifiers—in lieu of positional encodings—to learn contextual embeddings for categorical features with self-attention. Different from TabTransformer, we view each word location in a sentence as a tabular feature; our extension thus represents each feature in tabular data via an encoding akin to the positional encodings. Other tabular extensions of self-attention include FTTransformer (tokenizing each feature, applying transformer layers, and using the `[CLS]` token for prediction) [Gorishniy et al., 2021], AutoInt (mapping all features into the same space and applying self-attention to model between-feature interactions) [Song et al., 2019] and TabNet (utilizing sequential attention for feature selection in different learning steps) [Arik and Pfister, 2020]. Compared with these existing approaches, our approach is more robust to covariate shifts across empirical studies; it also facilitates the integration of heterogeneous datasets with partially overlapping features.

The third theme relates to linear word analogy 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", we often find $w_b + w_c - w_a \approx w_d$, where w_i denotes the embedding of word $i \in \{a, b, c, d\}$. Many works provide theoretical justifications for this phenomenon. Arora et al. [2016] offered a latent variable argument, assuming that texts are generated from random walks of discourse vectors and word vectors are spatially isotropic. Ethayarajh et al. [2018] introduced the *co-occurrence 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. In contrast to these existing works, our work moves beyond SGNS and GloVe; we characterize when linear word analogies may be present in CBOW and attention-based embeddings.

2 BIDIRECTIONAL ATTENTION AS A MIXTURE OF CONTINUOUS WORD EXPERTS

In this section, we first review bidirectional attention, a language model composed of the self-attention architecture, positional encodings, and the use of MLM training objective. En route, we derive an explicit form of the MLM objective for a single-layer single-head attention-based architecture in Section 2.1. We then formally establish the equivalence between fitting bidirectional attention and fitting the CBOW model with MOE weights in Section 2.2, with extensions to multi-head and multi-layer attention-based architectures.

2.1 BIDIRECTIONAL ATTENTION: SELF-ATTENTION, POSITIONAL ENCODINGS, AND THE MLM OBJECTIVE

We begin with describing the structure of bidirectional attention—self-attention, positional encodings, and the MLM objective—in the context of language modeling. (Appendix A contains a summary of the notations used in this section.)

Building blocks of bidirectional attention. Consider a corpus that consists of sentences of length S, with a vocabulary size of $|V|$. The self-attention mechanism takes sentences and outputs their sentence embeddings, via transforming the token embeddings and positional encodings of each token in the sentence. We denote $C \in \mathbb{R}^{(|V|+1)\times p}$ as the matrix such that each row $c_i^\top$ corresponds to the token embedding of the i-th token in the vocabulary. The $(|V|+1)$-th token is the `[MASK]` token, representing a token in the training corpus that is masked. Further denote $P \in \mathbb{R}^{S\times p}$as the positional encoding matrix.

To learn these token embeddings and positional encodings, bidirectional direction employs an MLM objective: it randomly masks a random subset of the tokens in the training corpus; then it aims to predict these masked tokens from the sentence embeddings, which are produced from transforming the token embeddings and positional encodings through the attention mechanisms. To operationalize the MLM objective, we use $\overline{X} \in \{0,1\}^{S\times(|V|+1)}$ to denote the one-hot encoding matrix of the S tokens (including the masked tokens) in each sentence. For notational simplicity, we consider a simple masking strategy: each sentence produces S prediction tasks in the MLM objective, each of which involves masking exactly one of the S positions in the sentence and predicting the token in that position. Results in this section can be easily generalized to other masking strategies.

Predicting masked tokens with self-attention. We next describe how the self-attention mechanism (with positional encodings) produces predictions of masked tokens. For ease of exposition, we focus on a single-head single-layer attention module. It takes in $\overline{X}$, the one-hot encoding matrix of the S tokens (including the masked tokens) in a sentence; it outputs a probability vector $\hat{y} \in \Delta^{|V|}$ as a prediction of the masked token, indicating the probability of the masked token being each of the $|V|$ words in the vocabulary.

The self-attention architecture transforms $\overline{X}$ into the prediction $\hat{y}$ following steps:

1. **Token embeddings with positional encodings**: Produce a matrix consisting of the token embeddings of all the tokens in the masked sentence: $X = \overline{X}C \in \mathbb{R}^{S\times p}$. Then add positional encodings to the matrix: $X' = X + P \in \mathbb{R}^{S\times p}$.
2. **Sentence embeddings with attention weight matrices:** Employing value mapping $W^V \in \mathbb{R}^{d\times p}$, query mapping $W^Q \in \mathbb{R}^{d_w\times p}$, and key mapping $W^K \in \mathbb{R}^{d_w\times p}$, we obtain the attention weight matrix $X^{\text{attn}} \in \mathbb{R}^{S\times d}$ as
$$X^{\text{attn}} = \text{softmax}\left(\frac{X'(W^Q)^\top W^K (X')^\top}{\sqrt{d_w}}\right) X'(W^V)^\top,$$
where the softmax is taken row-wise.
3. **Intermediate representations with residual connections:** Obtain an intermediate representation with coefficient matrix $W^O \in \mathbb{R}^{d\times p}$ and a residual connection: $Z = X^{\text{attn}} W^O \in \mathbb{R}^{S\times p}$; then $Z' = X' + Z \in \mathbb{R}^{S\times p}$.
4. **Final predictions with linear layer and residual connections.** For each position $i \in [S]$ of the sentence, apply a linear layer $\text{LIN}_1(Z'_i) = W'Z'_i \in \mathbb{R}^p$ with a weight matrix $W' \in \mathbb{R}^{p\times p}$; then another residual connection $Z'' = Z' + \text{LIN}_1(Z') \in \mathbb{R}^{S\times p}$; finally another linear layer and softmax operation
$$\hat{y} = \text{softmax}(\text{LIN}_2(Z''_i)),$$
where $\text{LIN}_2(Z''_i) = W''Z''_i \in \mathbb{R}^{|V|}$ with weight matrix $W'' \in \mathbb{R}^{|V|\times p}$.

Given the self-attention transformations from input sentences $\overline{X}$ to masked token predictions $\hat{y}$, bidirectional attention learns the token embeddings, positional encodings, and weight matrices by optimizing the cross entropy loss of $\hat{y}$ in predicting the masked tokens. This loss objective is also known as the MLM objective.

The loss objective of bidirectional attention. We next derive an explicit form for the loss objective of bidirectional attention. This derivation will pave the road for statistical interpretations of bidirectional attention.

In more detail, we consider an input-output pair $(\overline{X}, \overline{y})$ for the masked token prediction task, where $\overline{X}$ is the one-hot encoding matrix of all the tokens (including the masked tokens) in the sentence, and $\overline{y} \in \{0,1\}^{|V|}$ is the one-hot encoding of the token being masked. We denote $m \in [S]$ and $b \in [|V|]$ as the masked position and masked token, respectively. Lemma 1 below derives an explicit form of the MLM objective $L_{\text{MLM}}(m, b)$.

Lemma 1 (The loss objective of bidirectional attention)**.** *Upon reparametrization, the MLM objective for predicting token b in the mth position is given by*

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

where

$$\theta(j,m) \triangleq \exp\left(\frac{e_j^\top(\overline{X}C+P)W^{KQ}(c_{|V|+1}+P^\top e_m)}{\sqrt{d_w}}\right),$$

$$\chi(j,m,k) \triangleq \left(W^{LOV}(\overline{X}C+P)^\top e_j + g + De_m\right)_k,$$

and $g \in \mathbb{R}^{|V|}$, $D \in \mathbb{R}^{|V|\times S}$, $W^{LOV} \in \mathbb{R}^{|V|\times p}$, $W^{KQ} \in \mathbb{R}^{p\times p}$; $e_j \in \{0,1\}^S$ denotes a zero vector with 1 on the j-th entry. (The proof is in Appendix B.)

Lemma 1 performs a reparametrization over the weight matrices W^V, W^Q, W^K, W^O, arriving at an explicit form of the MLM objective with only two weight matrices W^{KQ}, W^{LOV}. Lemma 1 also reveals two key components of the MLM objective: $\theta(j,m)$, the attention weight of token m on token j, and $\chi(j,m,\cdot)$, the similarity between token m and token j. These quantities will play a key role in facilitating the statistical interpretation of bidirectional attention.

2.2 BIDIRECTIONAL ATTENTION AS A MIXTURE OF CONTINUOUS WORD EXPERTS

Building on the derivations in Lemma 1, we next establish the equivalence between the loss objective of bidirectional attention and that of the continuous bag of words (CBOW) model with mixture-of-experts (MOE) weights. This equivalence will enable us to interpret bidirectional attention as fitting a statistical model of CBOW+MOE.

The continuous bag of words model (CBOW). We begin with reviewing the CBOW formulation of word2vec [Mikolov et al., 2013]. CBOW aims to predict the center token based on the surrounding tokens (a.k.a. context tokens). It has two parameter matrices, representing the center and context embeddings respectively.

In more detail, we consider an input-output pair $(\overline{X}, \overline{y})$ as in Section 2.1, where $m \in [S]$ and $b \in [|V|]$ represent the masked position and masked token. We note that, while masking is never employed in CBOW, introducing masking into CBOW does not change its objective. The reason is that the context of a token in CBOW does not include the token itself. Thus the loss objective for predicting the token in mth position of CBOW (a.k.a. the negative log-likelihood) with window size w is

$$L_{\text{CBOW}}(m,b) = \log\left(\sum_{k=1}^{|V|}\exp\left(\sum_{j=1}^{S}\frac{\omega_{j,m,w}\xi(j,k)}{\sum_{j=1}^{S}\omega_{j,m,w}}\right)\right) - \sum_{j=1}^{S}\frac{\omega_{j,m,w}\xi(j,b)}{\sum_{j=1}^{S}\omega_{j,m,w}},$$

$$\text{where} \quad \omega_{j,m,w} = \mathbb{1}(1 \le |j-m| \le w), \quad \xi(j,k) = \left(W^{LOV}(\overline{X}C)^\top e_j\right)_k,$$

if we denote the center and context matrices by W^{LOV} and C to match the notations of bidirectional attention.

Weight and similarity matrices in CBOW and bidirectional attention. The CBOW model appears related to bidirectional attention: it admits natural notions of (attention) *weight* and (token) *similarity* as in bidirectional attention. Specifically, the *weight* of the token in position $j \in [S]$ is determined by the distance between j and m and the number of integers between $m-w$ and $m+w$ (inclusive) that are within the range $[1, S]$. The *similarity* of token $\alpha \in [|V|]$ in the center and token $\beta \in [|V|]$ in the context is $(W_\alpha^{LOV})^\top c_\beta$, regardless of their positions in the sentence.

To compare CBOW and bidirectional attention, we next inspect the weight matrices in the MLM objective of bidirectional attention. Specifically, the weight of the token in position j in L_{MLM} is given by[1]

$$\frac{\exp\left(e_j^\top(\overline{X}C+P)W^{KQ}(c_{|V|+1}+P^\top e_m)/\sqrt{d_w}\right)}{\sum_{j=1}^{S}\exp\left(e_j^\top(\overline{X}C+P)W^{KQ}(c_{|V|+1}+P^\top e_m)/\sqrt{d_w}\right)}.$$

[1]The weight and similarity matrices can take other parametric forms; e.g. Sonkar et al. [2020] uses a different weight function that depends on the center token b in their attention word embedding (AWE) model.

Unlike that of CBOW, this weight matrix of bidirectional attention depends on all tokens in the masked sentence and their corresponding positions. Yet, it does not depend on the center (masked) token b. Further, the term inside the $\exp(\cdot)$ can be decomposed into four components: (1) $e_j^\top \overline{X} C W^{KQ} c_{|V|+1}/\sqrt{d_w}$, which depends only on the token in position j; (2) $e_j^\top \overline{X} C W^{KQ} P^\top e_m/\sqrt{d_w}$, which depends on both position j and position m; (3) $e_j^\top P W^{KQ} c_{|V|+1}/\sqrt{d_w}$, which depends only on position j; and (4) $e_j^\top P W^{KQ} P^\top e_m/\sqrt{d_w}$, which depends on both position j and position m.

The similarity matrix of bidirectional attention also appears related to that of CBOW. In bidirectional attention, the similarity of token α in the center (in position m) and token β in the context (in position j) is given by $(W_\alpha^{LOV})^\top c_\beta + (W_\alpha^{LOV})^\top P^\top e_j + g_\alpha + (D_\alpha)^\top e_m$, which also contains four components as above. Moreover, the first component coincides with the similarity matrix of CBOW.

Bidirectional attention as a mixture of continuous word experts. Following these observations that bidirectional attention appears closely related to CBOW, we conclude this section with Theorem 2: it proves that the MLM objective of bidirectional attention in Lemma 1 is equivalent to the CBOW objective with MOE weights, where the token in each position serves as an expert.

Theorem 2 (Bidirectional attention as a mixture of continuous word experts)**.** *The MLM objective of bidirectional attention is equivalent to the cross-entropy loss between the token being masked $\overline{y}$ and the prediction probabilities softmax$(F(\overline{X}))$ from a mixture-of-experts (MOE) predictor:*

$$F(\overline{X}) = \sum_{j\in[S]} \pi_j(\overline{X}) f_j(\overline{X}),$$

where the jth expert $f_j(\overline{X})$ relies on the embedding of the token in position j,

$$f_j(\overline{X}) = W^{LOV}(\overline{X}C + P)^\top e_j + g + De_m,$$

and its weight (namely the contribution of expert j to prediction) is $\pi_j(\overline{X}) = \big(\mathrm{softmax}(h(\overline{X}))\big)_j$ with

$$h_j(\overline{X}) = e_j^\top(\overline{X}C + P)W^{KQ}(c_{|V|+1} + P^\top e_m)/\sqrt{d_w}.$$

Theorem 2 is an immediate consequence of Lemma 1. It formally establishes the equivalence between bidirectional attention and CBOW+MOE, enabling a statistical interpretation of bidirectional attention. In particular, Theorem 2 reveals the distinct use of MOE in bidirectional attention, which is a machine learning technique that excels at handling heterogeneous data. It thus can partially explain the empirical effectiveness of attention-based models in capturing heterogeneous patterns in complex natural language data [Devlin et al., 2018, Liu et al., 2019].

Extensions to multi-head and multi-layer bidirectional attention. We finally extend Theorem 2 to multi-head and multi-layer bidirectional attention. For bidirectional attention with multiple attention heads, its MLM objective can be shown to be equivalent to a stacked MOE of CBOW. For example, for bidirectional attention with two attention heads, its MLM objective is equivalent to cross entropy loss with the following stacked MOE predictor:

$$F(\overline{X}) = \sum_{j\in[S]} \pi_j^1(\overline{X}) f_j^1(\overline{X}) + \sum_{j\in[S]} \pi_j^2(\overline{X}) f_j^2(\overline{X}),$$

where the jth expert of the ith head is

$$f_j^i(\overline{X}) = W^{LOV_i}(\overline{X}C + P)^\top e_j + \frac{g}{2} + \frac{De_m}{2},$$

whose MOE weight is $\pi_j^i(\overline{X}) = \big(\mathrm{softmax}(h^i(\overline{X}))\big)_j$ with

$$h_j^i(\overline{X}) = e_j^\top(\overline{X}C + P)W^{KQ_i}(c_{|V|+1} + P^\top e_m)/\sqrt{d_w}.$$

Following similar derivations, one can show that bidirectional attention with multiple attention layers is equivalent to a mixture of MOEs.

3 BIDIRECTIONAL ATTENTION FOR TABULAR DATA

The equivalence between MLM with self-attention and CBOW with MOE weights (Theorem 2) suggests an immediate extension to categorical tabular data. We develop this tabular extension in this section. Across empirical studies, we find that this tabular extension of attention achieves significant improvement in OOD generalization over existing methods, including existing algorithms for tabular data (e.g. random forest, gradient boosting) and existing tabular generalizations of attention modules (e.g. TabTransformer, FTTransformer).

3.1 TABULAR EXTENSION OF BIDIRECTIONAL ATTENTION

To extend bidirectional attention to tabular data, we consider a classification problem with categorical features. For simplicity, we assume the response variable Y_i is ordinal with C classes. Further assume each of the K-dimensional features X_i is also ordinal with C classes. The training data contains pairs of features and responses (X_i, Y_i). The goal is to predict the response for some test X.

Extending bidirectional attention to this tabular setting requires that we handle tabular features with bidirectional attention. To this end, we leverage the observations in Theorem 2 that bidirectional attention can be viewed as prediction with MOE, where the token in each position of the sentence (endowed with positional encodings) serves as an expert. This MOE perspective of bidirectional attention immediately

suggests that we consider each tabular feature as an expert in tabular, since each position in a sentence can be viewed as a tabular feature for predicting masked tokens. One can thus consider using tabular feature encodings in the place of positional encodings for analyzing tabular data with bidirectional attention.

In more detail, for tabular extensions of bidirectional attention, 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 then introduce "position" encodings $p_1, \cdots, p_{K+1} \in \mathbb{R}^d$, one for each feature. Finally, we consider the concatenation of features and covariates (X_i, Y_i) of each data point as a sentence in bidirectional attention. These mapping enables us to learning the embeddings and encodings using the MLM objective. At test time, given a test X, one can use the bidirectional attention model to predict the most probable class for the input $(X_i, \texttt{[MASK]})$.

We note that this use of MLM objective for tabular data implicitly models the joint distribution $p(X, Y)$, as opposed to the conditional distribution $p(Y|X)$ that standard supervised algorithms commonly model. As a consequence, tabular extensions of bidirectional attention can potentially achieve better OOD, as we demonstrate empirically next.

Finally, this tabular extension of bidirectional attention can be applied beyond supervised classification. It immediately extends to unsupervised settings (if we ignore the Y_i's) and semi-supervised settings (if we consider both the labeled and unlabeled data and set the Y_i's for the unlabeled data to `[MASK]`). This approach is also applicable to handling multiple datasets with only partially overlapping features; the reason is that the learned feature encodings allow us to bring all features into the same embedding space. In the mean time, these learned encodings can also reveal the relationships between different tabular features across different data sets.

3.2 EMPIRICAL STUDIES OF TABULAR BIDIRECTIONAL ATTENTION

In this section, we empirically study the tabular extension of bidirectional attention using simulated and real datasets. Across empirical studies, we find that our approach outperforms in OOD generalization for tabular data, as compared with both existing tabular data algorithms and existing tabular extensions of attention modules.

3.2.1 Simulated data

Begin with evaluating tabular bidirectional attention on simulated. We focus on the common OOD generalization setting of covariate shift; it refers to prediction tasks where $p(X_{\text{train}}) \neq p(X_{\text{test}})$ and $p(Y_{\text{train}}|X_{\text{train}}) = p(Y_{\text{test}}|X_{\text{test}})$.

Data generation. We describe the key components of data generation process; we refer the readers to Appendix H for full details. We set the number of features K to be 5, the number of classes C to be 10, and the training and test set size to be 2,000 each. Twenty data sets are generated for each combination of hyperparameters.

Competing methods and evaluation metrics. We fit the proposed tabular extension of bidirectional attention model to each training set, together with a few competing methods, namely logistic regression (LR), random forests (RF), gradient boosting (GB) and multilayer perceptron (MLP). See Appendix I for implementation details.

Results. Table 1 summarizes the test accuracy and mean squared error of all methods. We find that the proposed tabular extension of bidirectional attention outperforms or competitively compares to all competing methods. Moreover, its performance gain is more apparent when corr $= 0.9$ (very correlated training features) as compared to when corr $= 0.1$; the former corresponds to a more challenging case of covariate shift.

3.2.2 UCI's auto-mpg data

We next study the tabular extension of bidirectional attention on a real dataset, namely the `auto-mpg` data from the UCI data set. This data set contains the following information from 398 different car models: *mpg*, *cylinders*, *displacement*, *horsepower*, *weight*, *acceleration*, *model year*, *origin*, and *car name*.

Data processing. To simulate covariate shift, we follow the approach of Sugiyama and Storkey [2006]: we assigns cars from origin 1 to the training set, and origins 2 and 3 to the test set. In addition, we only consider cars with 4, 6 or 8 cylinders and remove data points with missing values. Lastly, similar to the synthetic data experiments, we convert each column into three quantile-based categories. The final data set has 385 data points, where 245 belong to the training set and 140 belong to the test set.

Competing methods and evaluation metrics. We use the same competing methods and evaluation metrics as in Section 3.2.1. Additionally, we compare with other existing tabular extensions of attention modules, including CategoryEmbedding (CE) [Joseph, 2021], FTTransformer (FT) [Gorishniy et al., 2021], TabTransformer (TT) [Huang et al., 2020], AutoInt (AI) [Song et al., 2019], and TabNet (TN) [Arik and Pfister, 2020].[2]

Results. Table 2 summarizes the test accuracy and mean squared error of all methods. We find that the proposed tabular extension of bidirectional attention outperforms all competing methods. This performance gain is likely due to its focus on modeling the joint distribution of the covariates

[2] We use `pytorch_tabular`'s [Joseph, 2021] implementation with the default parameters. The batch and epoch sizes are set to be 128 and 200, respectively.

Table 1: The proposed tabular extension of bidirectional attention (ATN) achieves better or competitive accuracy and MSE than competing methods, across all parameter settings. The parameter tuples indicate different choices of $(n_c, \text{noise}, \text{corr})$.

Param. \ **Acc.**	LR	RF	GB	MLP	ATN
$(1, 0, 0.1)$	0.388	0.409	**0.413**	0.323	0.404
$(1, 0, 0.9)$	0.313	0.298	0.350	0.237	**0.389**
$(1, 0.5, 0.1)$	0.345	0.361	**0.366**	0.292	0.359
$(1, 0.5, 0.9)$	0.270	0.253	0.299	0.202	**0.306**
$(1, 1.5, 0.1)$	0.250	0.243	**0.253**	0.204	0.252
$(1, 1.5, 0.9)$	0.169	0.158	**0.172**	0.142	0.170
$(5, 0, 0.1)$	0.250	0.207	0.244	0.306	**0.419**
$(5, 0, 0.9)$	0.162	0.150	0.156	0.169	**0.392**
$(5, 0.5, 0.1)$	0.227	0.173	0.214	0.252	**0.318**
$(5, 0.5, 0.9)$	0.154	0.133	0.153	0.151	**0.269**
$(5, 1.5, 0.1)$	0.167	0.099	0.157	0.165	**0.171**
$(5, 1.5, 0.9)$	0.125	0.108	0.114	0.118	**0.133**

Param. \ **MSE**	LR	RF	GB	MLP	ATN
$(1, 0, 0.1)$	3.015	**2.694**	2.730	4.059	2.941
$(1, 0, 0.9)$	5.163	9.331	4.855	7.911	**3.078**
$(1, 0.5, 0.1)$	3.416	3.201	**3.123**	4.704	3.281
$(1, 0.5, 0.9)$	5.955	10.106	6.070	8.123	**4.465**
$(1, 1.5, 0.1)$	5.725	5.685	**5.415**	7.199	5.594
$(1, 1.5, 0.9)$	8.942	12.340	9.837	9.874	**7.339**
$(5, 0, 0.1)$	5.333	8.498	5.967	2.814	**1.521**
$(5, 0, 0.9)$	5.674	10.101	8.858	7.842	**1.633**
$(5, 0.5, 0.1)$	6.021	10.236	6.844	4.056	**2.605**
$(5, 0.5, 0.9)$	6.118	10.427	8.283	7.884	**2.355**
$(5, 1.5, 0.1)$	9.159	16.154	9.538	**8.313**	8.316
$(5, 1.5, 0.9)$	8.410	10.409	10.110	9.966	**6.501**

and response variable, as opposed to only the conditional distribution of the response variable given the covariates.

4 LINEAR WORD ANALOGIES IN ATTENTION-BASED EMBEDDINGS

In this section, we explore the presence (or the lack of presence) of linear word analogies in the embeddings of bidirectional attention and its non-attention predecessors, leveraging the close connections between CBOW and bidirectional attention in Theorem 2. This exploration is motivated by a a curious empirical observation: While bidirectional attention (e.g. BERT) often significantly outperforms its non-attention predecessors in natural language processing benchmarks, it does not seem to outperform its predecessors in word analogy tasks. In particular, it can sometimes perform worse in word analogy tasks than classical word embedding algorithms like word2vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014].

Thanks to these empirical observations, we characterize under which conditions bidirectional attention and CBOW can exhibit linear word analogies in their embeddings. We find that bidirectional attention requires much stronger conditions to exhibit linear word analogies than its non-attention predecessors. These results partially explain the curious empirical performance of bidirectional attention in word analogy tasks.

4.1 A CURIOUS EMPIRICAL STUDY: DO ATTENTION-BASED TOKEN EMBEDDINGS EXHIBIT LINEAR WORD ANALOGIES?

We begin with a curious empirical study about the presence of linear word analogies in attention-based and non-attention-based token embeddings. Linear structure in neural word embeddings such as word2vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014] is a well-known empirical phenomenon. However, most studies focused on embeddings trained via SGNS [Ethayarajh et al., 2018, Allen and Hospedales, 2019]. This phenomenon is less studied in more recent language modeling approaches, e.g. CBOW and bidirectional attention, with few exceptions [Ushio et al., 2021].

To this end, we perform an empirical study about whether linear relationships are observed in embeddings from word2vec trained with the CBOW objective and BERT [Devlin et al., 2018], a large language model based on bidirectional attention. Following other existing studies, we use the analogy identification task as a proxy for identifying the presence of linear relationships, using the analogy data set first introduced in Pennington et al. [2014]. We refer the readers to Appendix J for dataset and implementation details.

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. We note that (2) is a better metric than (1) due to the difference in vocabulary sizes across models.

Results. The accuracy and average cosine similarity for each model is displayed in Table 3. We observe that all three models generally result in word embeddings that exhibit certain linear word analogies. However, the bidirectional attention model BERT can often perform worse than its non-attention predecessor GloVe in this task, despite it being a much more powerful language model in common natural language benchmarks.

What factors have limited BERT's (and bidirectional attention's) ability to exhibit linear word analogies? What about CBOW and GloVe? Below we study these questions theoretically, leveraging the close connection between CBOW and bidirectional attention in Theorem 2. In particular, we characterize the conditions under which CBOW and bidirectional attention may exhibit linear word analogies respectively. We find that the conditions required by bidirectional attention is much stronger, which partially explains the curious empirical

Table 2: The proposed tabular extension of attention (ATN) achieves superior performance as compared to all baselines. (Lower MSE and higher accuracy is better.)

	LR	RF	GB	MLP	CE	FT	TT	AI	TN	ATN (ours)
Accuracy	0.657	0.721	0.657	0.700	0.764	0.707	0.707	0.364	0.600	**0.793**
MSE	0.343	0.279	0.343	0.300	0.236	0.293	0.293	0.636	0.486	**0.207**

observations above.

4.2 LINEAR WORD ANALOGIES IN CBOW AND BIDIRECTIONAL ATTENTION EMBEDDINGS

We begin with theoretically characterize under which conditions can CBOW embeddings exhibit linear word analogies. Adopting Allen and Hospedales's [2019] argument for SGNS, we extend the argument to both CBOW and attention-based token embeddings, thanks to the equivalence we established in Theorem 2.

To perform this theoretical analysis, we follow existing analyses about SGNS: Levy and Goldberg [2014] showed that for a sufficiently large embedding dimension, embeddings from SGNS satisfy $w_i^\top c_j = \log\left(\frac{p(w_i,c_j)}{p(w_i)p(c_j)}\right) - \log k = \text{PMI}(w_i, c_j) - \log k$, where k is the number of negative samples for each positive sample; $W^{LOV}, C \in \mathbb{R}^{|V|\times p}$ are the center and context embedding matrix, respectively. For each $i \in [|V|]$, $w_i^\top$ ($c_i^\top$) is the i-th row of W^{LOV} (C), which represents the center (context) embedding of word i.

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 SGNS word embeddings.

We next perform similar analyses for CBOW and bidirectional attention; the goal is to characterize the conditions under which CBOW and bidirectional attention can exhibit linear word analogies respectively. Below we state the main results we obtain, leaving full details to Appendix K.

Linear word analogies in CBOW embeddings. We first characterize the inner product of center and contextual embeddings of CBOW.

Proposition 3. *Embeddings from fitting CBOW without negative sampling must satisfy* $w_i^\top c_j \approx \log\left(\frac{p(w_i,c_j)}{p(c_j)}\right) + \log|V|$.

This result suggests that CBOW approximately factorizes M, a $|V| \times |V|$ matrix such that $M_{i,j} = \log\left(\frac{p(w_i,c_j)}{p(c_j)}\right) + \log|V|$.

Following this result, we next argue that the CBOW embeddings approximately form a linear relationship, up to some error terms.

Proposition 4. *Given any* $w_a, w_{a^*}, w_b, w_{b^*} \in \mathcal{E}$, *we have*

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

where $\mathcal{E}$ *is the set of all words in the vocabulary,* $\mathcal{W} = \{w_b, w_{a^*}\}$, $\Delta^{\mathcal{W},\mathcal{W}_*} = \sigma^{\mathcal{W}} - \sigma^{\mathcal{W}_*}$ *and* $\mathcal{W}_* = \{w_{b^*}, w_a\}$. *The quantities* $\rho^{\mathcal{W},\mathcal{W}_*}, \Delta^{\mathcal{W},\mathcal{W}_*}, \delta^{\mathcal{W},\mathcal{W}_*}, \xi^{\mathcal{W},\mathcal{W}_*}$ *are all statistics that characterize the relationships between the two word sets* $\mathcal{W}, \mathcal{W}_*$. *We refer the reader to Appendix K for their precise definitions and complete details of the results.*

Proposition 4 reveals that we have linear word analogies $w_{b^*} \approx w_{a^*} - w_a + w_b$ when $\mathcal{W}$ paraphrases $\mathcal{W}_*$ in the sense of Allen and Hospedales [2019], and $\sigma^{\mathcal{W}}$, $\sigma^{\mathcal{W}_*}$ and $\delta^{\mathcal{W},\mathcal{W}_*}$ are small. The latter holds true only when all $w_i \in \mathcal{W}$ ($w_i \in \mathcal{W}_*$) are approximately conditionally independent given c_j, and $p(\mathcal{W}) \approx p(\mathcal{W}_*)$. If we consider alternative definitions of paraphrase—which we detail in Appendix K, then the linear analogy error may only depend on the approximate conditional independence of w_i's given c_j.

Finally, we characterize conditions under which, if token embeddings of CBOW exhibit linear word analogies, then its contextual embedding will also exhibit this structure.

Proposition 5. *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. Further, assume that* W *has full row rank. If* $w_r + w_s \approx w_t + w_u$, *then* $c_r + c_s \approx c_t + c_u$.

Linear word analogies for bidirectional attention. We next extend these CBOW arguments to bidirectional attention, leveraging the close connection established in Theorem 2. We will show that the same linear word analogies may emerge in bidirectional attention, but under much stronger assumptions.

Proposition 6. *Token embeddings from bidirectional attention must satisfy*

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

where for a center-context pair (d, j) *in the masked sentence* $(a_1, \cdots, a_S)$, *we define* $\gamma_j^d = \tau_j / \sum_{s=1}^S \tau_{a_s}$, *and* $\tau_j = \exp\left(c_j^\top W^{KQ} c_{|V|+1} / \sqrt{d_w}\right)$.

Table 3: Classical word embedding methods can achieve similar or higher performance than attention-based model in word analogy tasks: GloVe achieve higher or the same average cosine similarity than BERT on both syntactic and semantic analogies; GloVe also outperforms BERT in accuracy for semantic analogies. (Higher is better.)

Accuracy	BERT	GloVe	CBOW
Semantic	0.641	0.759	0.234
Syntactic	0.754	0.692	0.667
Overall	0.727	0.708	0.563

Cosine similarity	BERT	GloVe	CBOW
Semantic	0.500	0.600	0.504
Syntactic	0.610	0.610	0.582
Overall	0.584	0.607	0.564

Proposition 6 shows that bidirectional attention approximately factorizes a $|V| \times |V|$ matrix whose (i, j)-th entry is given by the equation above. Unlike in CBOW the token 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 $j \in [|V| + 1]$, the problem approximately reduces to a vanilla CBOW. In particular, we always have $\gamma_j^d \approx 1/S$, whence Proposition 6 yields $w_i^\top c_j \approx \frac{p(w_i,c_j)}{p(c_j)} \cdot |V| - 1 \approx \log\left(\frac{p(w_i,c_j)}{p(c_j)}\right) + \log|V|$.

Following a similar argument as Proposition 4, we argue that the bidirectional attention embedding can also exhibit linear word analogies, up to some error.

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

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

where $\overline{\Delta}^{\mathcal{W},\mathcal{W}_*} = \bar{\sigma}^{\mathcal{W}} - \bar{\sigma}^{\mathcal{W}_*}$, $\mathcal{W} = \{w_b, w_{a^*}\}$ and $\mathcal{W}_* = \{w_{b^*}, w_a\}$.

Under additional conditions, similar linear word analogy relationships may also emerge for contextual embeddings of bidirection attention.

Proposition 8. *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 $\tilde{c}_r + \tilde{c}_s \approx \tilde{c}_t + \tilde{c}_u$.*

While we leave the full details of these results to Appendix K, Propositions 7 and 8 suggest that bidirectional attention requires much stronger conditions to exhibit linear relationships than CBOW. Specifically, it requires the quantity $\bar{p}(w_i, c_j) = \sum_{(i,j)} \gamma_j^i / E$ to be approximately symmetric. Even when this condition holds, linear word analogy would only hold for some transformed embeddings $\tilde{c}_i$'s, as opposed to the token embeddings c_i's. Only under an additional assumption that $\zeta_j := \frac{\sum_{(1,j)}(\gamma_j^1)^2 + \cdots + \sum_{(|V|,j)}(\gamma_j^{|V|})^2}{\sum_{(1,j)} \gamma_j^1 + \cdots + \sum_{(|V|,j)} \gamma_j^{|V|}}$ is approximately the same for each j (e.g., when τ_j is approximately the same for every j), we will approximately have linear word analogies for the token embeddings c_i's.

Finally, we note that all these results can be easily extended to incorporate positional encodings by considering each (word, position) pair as a unit. As a result, analogies will be drawn between (word, position) pairs.

5 DISCUSSION

In this paper, we prove that a single-head single-layer bidirectional attention is equivalent to a continuous bag of words (CBOW) model with mixture-of-experts (MOE) weights, given appropriate reparameterization. This statistical perspective reveals the distinct use of MOE in bidirectional attention, supporting the empirical observations that bidirectional attention excels in capturing heterogeneous pattern. This connection further suggests immediate extensions of attention to tabular data, leading to improved out-of-distribution (OOD) generalizations as compared to existing approaches. It also allows us characterize the conditions under which embeddings from bidirectional attention and CBOWexhibit linear structure. These analyses show that bidirectional attention requires much stronger assumptions than its non-attention predecessors to exhibit linear word analogies.

One limitation of this work is that the linear word analogy argument in Section 4 ignores residual connections. In addition, we only consider bidirectional attention architectures that use linear linears, as opposed to feed-forward layers used in Devlin et al. [2018]. Beyond addressing these limitations, exploring the statistical properties of bidirectional attention is an interesting avenue for future work. It will also be interesting to provide theoretical justifications for the observed robustness of bidirectional attention to covariate shifts, and to understand the fundamental differences between static and contextual word embeddings in relation to their abilities to form linear analogies.

ACKNOWLEDGEMENTS

This work was supported in part by the Office of Naval Research under grant number N00014-23-1-2590 and the National Science Foundation under Grant No. 2231174 and No. 2310831. We thank Sasha Rush for suggesting the name "bidirectional attention."

References

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

S. O. Arik and T. Pfister. Tabnet: Attentive interpretable tabular learning, 2020.

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.

Y. Bai, F. Chen, H. Wang, C. Xiong, and S. Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection, 2023.

A. Bietti, V. Cabannes, D. Bouchacourt, H. Jegou, and L. Bottou. Birth of a transformer: A memory viewpoint, 2023.

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. URL 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.

C. Han, Z. Wang, H. Zhao, and H. Ji. In-context learning of large language models explained as kernel regression, 2023.

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.

M. Joseph. Pytorch tabular: A framework for deep learning with tabular data, 2021.

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. Li, Y. Li, and A. Risteski. How do transformers learn topic structure: Towards a mechanistic understanding, 2023.

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.

W. Song, C. Shi, Z. Xiao, Z. Duan, Y. Xu, M. Zhang, and J. Tang. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1161–1170, New York, NY, USA, 2019. Association for

Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3357925. URL https://doi.org/10.1145/3357384.3357925.

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

M. Sugiyama and A. J. Storkey. Mixture regression for covariate shift. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/a74c3bae3e13616104c1b25f9da1f11f-Paper.pdf.

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. Ushio, L. Espinosa Anke, S. Schockaert, and J. Camacho-Collados. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.280. URL https://aclanthology.org/2021.acl-long.280.

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.

S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. An explanation of in-context learning as implicit bayesian inference, 2022.