Supplementary Material: Bidirectional Attention as a Mixture of Continuous Word Experts

A SUMMARY OF NOTATIONS

Below is a summary of commonly-used notations in Section 4.

Notation	Explanation
$\lvert V\rvert$	Vocabulary size
S	Sentence length
p	Embedding dimension
W^{LOV}	Center embedding matrix
C	Token (context) embedding matrix
$w_i^\top$	i-th row of W^{LOV}
$c_i^\top$	i-th row of C
P	Position encoding matrix
$\overline{X}$	One-hot encoding matrix of the masked sentence
$\overline{y}$	One-hot encoding of the target word
m	Position of the masked word
b	The masked word
e_j	A zero vector of length S with 1 on the j-th entry
$f_j(\cdot)$	The output generated by expert j
$\pi_j(\cdot)$	The contribution of expert j
a_s	The word on the s-th position of the masked sentence

B A SKETCH OF THE ATTENTION-BASED ARCHITECTURE

1. Let $X = \overline{X}C \in \mathbb{R}^{S\times p}$ be a matrix consisting of the token embeddings of each word in the masked sentence, and $X' = X + P \in \mathbb{R}^{S\times p}$.
2. Introduce attention weight matrices $W^V \in \mathbb{R}^{d\times p}$, $W^Q \in \mathbb{R}^{d_w\times p}$ and $W^K \in \mathbb{R}^{d_w\times p}$. Let $X^{\text{attn}} = \text{softmax}\left(\frac{X'(W^Q)^\top W^K(X')^\top}{\sqrt{d_w}}\right) X'(W^V)^\top \in \mathbb{R}^{S\times d}$, where the softmax is taken row-wise.
3. Let $W^O \in \mathbb{R}^{d\times p}$, and write $Z = X^{\text{attn}}W^O \in \mathbb{R}^{S\times p}$.
4. Introduce a residual connection, and write $Z' := X' + Z \in \mathbb{R}^{S\times p}$.
5. For each position $i \in [S]$, apply a linear layer $LIN_1(Z_i') = W'Z_i' \in \mathbb{R}^p$, where $W' \in \mathbb{R}^{p\times p}$.
6. Introduce another residual connection, and write $Z'' = Z' + LIN_1(Z') \in \mathbb{R}^{S\times p}$.
7. For each position $i \in [S]$, apply a linear layer $LIN_2(Z_i'') := W''Z_i'' \in \mathbb{R}^{|V|}$, where $W'' \in \mathbb{R}^{|V|\times p}$.
8. Perform the softmax operation and calculate the cross-entropy loss corresponding to predicting the masked word in the sentence.

C PROOF OF LEMMA 1

Proof. Recall that $m \in [S]$ and $b \in [|V|]$ represent the masked position and masked word, respectively. It is easy to see that $X^\top e_m = c_{|V|+1}$, where $e_m \in \{0,1\}^S$ is a zero vector with 1 on the m-th entry. Note that steps 1 to 4 of Appendix B give us

$$Z' = X + P + \text{softmax}\left(\frac{(X+P)(W^Q)^\top W^K(X+P)^\top}{\sqrt{d_w}}\right)(X+P)(W^V)^\top W^O \in \mathbb{R}^{S\times p}.$$

This is followed by steps 5 and 6, which yield $Z'' = Z' + LIN_1(Z')$ where the i-th row of Z'' is given by $(Z_i'')^\top$, where $Z_i'' = Z_i' + W'Z_i'$ for some $W' \in \mathbb{R}^{p\times p}$. Lastly, steps 7 and 8 result in $\alpha_m = \text{softmax}(W''Z_m'')$ for some $W'' \in \mathbb{R}^{|V|\times p}$, from which the loss is simply $-\log(e_b^\top \alpha_m)$, where $e_b \in \{0,1\}^{|V|}$ is a zero vector with 1 on the b-th entry. See that

$$\begin{aligned} W''Z_m'' &= (W'' + W''W')(Z')^\top e_m \\ &= W^\ell \left((X+P)^\top + (W^O)^\top W^V (X+P)^\top \text{softmax}\left(\frac{(X+P)(W^K)^\top W^Q (X+P)^\top}{\sqrt{d_w}} \right) \right) e_m, \end{aligned}$$

where $W^\ell = W'' + W''W' \in \mathbb{R}^{|V|\times p}$ and the softmax is taken column-wise. Writing $W^\ell c_{|V|+1} = g \in \mathbb{R}^{|V|}$, $W^\ell P^\top = D \in \mathbb{R}^{|V|\times S}$, $W^\ell (W^O)^\top W^V = W^{LOV} \in \mathbb{R}^{|V|\times p}$ and $(W^K)^\top W^Q = W^{KQ} \in \mathbb{R}^{p\times p}$, we obtain

$$\begin{aligned} W''Z_m'' &= g + De_m + \sum_{j=1}^{S} \frac{\exp\left(\frac{e_j^\top (X+P)W^{KQ}(c_{|V|+1}+P^\top e_m)}{\sqrt{d_w}}\right)}{\sum_{j=1}^{S}\exp\left(\frac{e_j^\top (X+P)W^{KQ}(c_{|V|+1}+P^\top e_m)}{\sqrt{d_w}}\right)} \left(W^{LOV}(X+P)^\top e_j\right) \\ &= \sum_{j=1}^{S} \frac{\exp\left(\frac{e_j^\top (\overline{X}C+P)W^{KQ}(c_{|V|+1}+P^\top e_m)}{\sqrt{d_w}}\right)}{\sum_{j=1}^{S}\exp\left(\frac{e_j^\top (\overline{X}C+P)W^{KQ}(c_{|V|+1}+P^\top e_m)}{\sqrt{d_w}}\right)} \left(W^{LOV}(\overline{X}C+P)^\top e_j + g + De_m\right), \end{aligned}$$

and the objective for this particular instance is

$$\begin{aligned} &-\sum_{j=1}^{S} \frac{\exp\left(\frac{e_j^\top (\overline{X}C+P)W^{KQ}(c_{|V|+1}+P^\top e_m)}{\sqrt{d_w}}\right)}{\sum_{j=1}^{S}\exp\left(\frac{e_j^\top (\overline{X}C+P)W^{KQ}(c_{|V|+1}+P^\top e_m)}{\sqrt{d_w}}\right)} \left(W^{LOV}(\overline{X}C+P)^\top e_j + g + De_m\right)_b \\ &+ \log\left(\sum_{k=1}^{|V|} \exp\left(\sum_{j=1}^{S} \frac{\exp\left(\frac{e_j^\top (\overline{X}C+P)W^{KQ}(c_{|V|+1}+P^\top e_m)}{\sqrt{d_w}}\right)}{\sum_{j=1}^{S}\exp\left(\frac{e_j^\top (\overline{X}C+P)W^{KQ}(c_{|V|+1}+P^\top e_m)}{\sqrt{d_w}}\right)} \left(W^{LOV}(\overline{X}C+P)^\top e_j + g + De_m\right)_k \right)\right), \end{aligned}$$

completing the proof.

D TABULAR DATA GENERATION PROCESS

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: (1) $n_c \in \{1, 5\}$, the number of features which generate Y; (2) noise $\in \{0, 0.5, 1.5\}$, where a larger value indicates a larger noise in the observed features; and (3) corr $\in \{0.1, 0.9\}$, where a larger value indicates a larger between-feature correlation in the training set as compared to the test set.

To simulate covariate shift, we introduce the parameter corr: the correlation of any covariate pair is $\pm$corr in the training set, and $1 -$ corr in the test set. We generate the responses as a linear combination of the covariates. Moreover, we add Gaussian noise to the covariates, mimicking settings where covariates are measured with error. Lastly, we bin each covariate and response into $C = 10$ categories based on their quantiles. This results in a 10-class classification problem with ordinal covariates and responses.

For a fixed $n_c \in \{1, 5\}$, noise $\in \{0, 0.5, 1.5\}$ and corr $\in \{0.1, 0.9\}$, our data generation process can be described as follows.

1. Let train_cov = corr $\cdot J_5 + (1 -$ corr$) \cdot I_5$ and test_cov $= (1 -$ corr$) \cdot J_5 +$ corr $\cdot I_5$. Here, J_5 represents a 5×5 matrix whose entries are all 1, and I_5 represents a 5×5 identity matrix.
2. Generate samples train_x_true and test_x_true from zero-mean multivariate normal distributions with covariance matrices train_cov and test_cov, respectively. Each sample is of size 2,000.
3. Introduce positively and negatively correlated covariates in the training samples by multiplying data in the first two features by -1.

4. Add Gaussian observation noises to the training and test samples. For the n_c features which generate the response, add $0.4 \cdot \text{noise} \cdot \mathcal{N}(0,1)$; otherwise, add $0.3 \cdot \text{noise} \cdot \mathcal{N}(0,1)$. Let the resulting samples be train_x and test_x.

5. Generate the true coefficient for each of the n_c features from $\mathcal{U}(0,10)$.

6. Generate the training response train_y, which is a linear combinations of the n_c features of train_x_true with the true coefficients as weights, plus a Gaussian noise from $\mathcal{N}(0,4)$. Generate the test response test_y in a similar manner.

7. Bin each feature and response of $(\text{train\_x}, \text{train\_y})$ and $(\text{test\_x}, \text{test\_y})$ into 10 quantile-based categories.

E IMPLEMENTATION AND HYPERAMETER TUNING PROCESS FOR COMPETING MODELS

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). We then evaluate the prediction accuracy (Acc) and mean squared error (MSE) on the corresponding test set. For each set of hyperparameters, we take the average of both metrics across the 20 generated data sets.

We implement the proposed extension of bidirectional attention (ATN) in Keras using a single-layer BERT [Devlin et al., 2018] with 5 heads, an embedding size of 20, and a feed-forward layer of dimension 5. We use the Adam optimizer with the default parameters, and a batch and epoch size of 128 and 200, respectively. For the competing methods, we use sklearn's implementation with hyperparameters chosen via 5-fold cross-validation in classification accuracy.

For each data set, the hyperparameters of the random forest (RF), gradient boosting (GB) and multilayer perceptron (MLP) models are chosen via 5-fold cross-validation based on the classification accuracy.

Random forest. We consider every combination of the following hyperparameters: (a) *criterion*: `gini` or `entropy`; (b) *n_estimators*: 50, 100 or 200; and (c) *max_depth*: 1, 3 or `None`.

Gradient boosting. We consider every combination of the following hyperparameters: (a) *learning_rate*: 0.01, 0.1 or 1; (b) *n_estimators*: 50, 100 or 200; and (c) *max_depth*: 1, 3 or 5.

Multilayer perceptron. We consider every combination of the following hyperparameters: (a) *hidden_layer_sizes*: (50,), (100,) or (100,50); (b) *alpha*: 0.0001, 0.001 or 0.01; and (c) *learning_rate*: `constant` or `adaptive`.

F DETAILS OF THE WORD ANALOGY EXPERIMENT

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; x_i represents the embedding of word i.

We only include a question when all four words involved are present in the vocabulary list of each model. Out of 19,544 questions in the data set, 9,522 (49%) of them satisfy this condition. Analyzing each category separately, we find that the condition is satisfied for 2,278 (26%) out of 8,869 semantic questions, and 7,244 (68%) out of 10,675 syntactic questions.

Models. We consider three models: (1) BERT base uncased, which is used in the original BERT paper [Devlin et al., 2018]; (2) GloVe trained on Wikipedia [Pennington et al., 2014]; (3) word2vec trained with CBOW [Mikolov et al., 2013]. The embedding dimensions of these models are 768, 300, 300 and 768, respectively, while the vocabulary size are around 30K, 400K, 3M and 30K, 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; we also exclude non-words (e.g., *[unused9]*, *## ?*) from the list of possible answers.

G DETAILED ANALYSIS OF EMBEDDINGS FOR continuous bag of words (CBOW) AND BIDIRECTIONAL ATTENTION

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

G.1 LINEAR WORD ANALOGIES IN CBOW EMBEDDINGS

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 to characterize their conditions required for linear word analogies.

What matrix does CBOW (approximately) factorize? Proposition 9 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 H.

Proposition 9. *Consider CBOW without negative sampling. Using the same notation as before, we have*

$$w_i^\top c_j \approx \log\left(\frac{p(w_i, c_j)}{p(c_j)}\right) + \log|V|.$$

From Proposition 9, we know that CBOW approximately factorizes M, a $|V| \times |V|$ matrix such that

$$M_{i,j} = \log\left(\frac{p(w_i, c_j)}{p(c_j)}\right) + \log|V|.$$

It is worth noting that this formula is similar to that for noise-contrastive estimation (NCE) as mentioned in Levy and Goldberg [2014], with $\log|V|$ replaced by $-\log k$. Also, observe that $w_i^\top c_j > w_k^\top c_j$ if and only if $p(w_i, c_j) > p(w_k, c_j)$.

We empirically verify Proposition 9 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 I. 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 $\log\left(p(w_i, c_j)/p(c_j)\right) + \log|V|$ 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.

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

Definition 10 (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 $\rho^{\mathcal{W},\mathcal{W}_*} \in \mathbb{R}^{|V|}$ 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 10, we observe that $\mathcal{W} \approx_P \mathcal{W}_*$ if and only if $\mathcal{W}_* \approx_P \mathcal{W}$. Also, we implicitly require 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, $M_i^\top$ denotes the i-th row of M. The proof is provided in Appendix J.

Lemma 11. *For any word sets $\mathcal{W}, \mathcal{W}_* \subseteq \mathcal{E}$ with the same cardinality, we have*

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

where

$$\sigma_j^{\mathcal{W}} = \log\left(\frac{p(\mathcal{W}|c_j)}{\prod_{w_i \in \mathcal{W}} p(w_i|c_j)}\right),$$

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

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

Proposition 12, which is equivalent to Corollary 2.3 of Allen and Hospedales [2019], follows from multiplying both sides of the equations in Lemma 11 by $C^\dagger = (CC^\top)^{-1}C$ (assuming C has full row rank) and setting $\mathcal{W} = \{w_b, w_{a^*}\}$ and $\mathcal{W}_* = \{w_{b^*}, w_a\}$.

Proposition 12. *Given any $w_a, w_{a^*}, w_b, w_{b^*} \in \mathcal{E}$, we have*

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

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

From Proposition 12, 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 mentioned above.

Alternatively, we can modify Definition 10 so that $\mathcal{W} \approx_P \mathcal{W}_*$ if and only if $\xi^{\mathcal{W},\mathcal{W}_*}$ (instead of $\rho^{\mathcal{W},\mathcal{W}_*}$) is element-wise small. Now, 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 $c_r + c_s \approx c_t + c_u$? Proposition 13, whose proof is provided in Appendix K, answers the question.

Proposition 13. *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 linear structures under some assumptions. We now extend this argument to MLM with self-attention, and show that the same conclusion holds under stronger assumptions.

G.2 LINEAR WORD ANALOGIES IN ATTENTION-BASED EMBEDDINGS

Similar to Section 4.2, we compute the matrix MLM with self-attention factorized and construct a paraphrasing argument to show linear structures in the learned embeddings.

What matrix does MLM with self-attention (approximately) factorize? To make calculations tractable, we exclude both residual connections and positional encodings. Let the masked sentence be $(a_1, \cdots, a_S)$. As before, let $m \in [S]$ and $b \in [|V|]$ denote the masked position and masked word, respectively. This means $a_i \in [|V|]$ for every $i \neq m$ and $a_m = |V| + 1$. From Lemma 1, the loss for this instance is given by

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

where $\tau_j = \exp\left(\frac{c_j^\top W^{KQ} c_{|V|+1}}{\sqrt{d_w}}\right)$. Proposition 14 approximates the matrix factorized by the attention objective, given all τ_j values for each $j \in [|V|+1]$. The proof is similar to that of Proposition 9, and therefore omitted.

Proposition 14. *Consider the attention objective as in Equation* (1)*. We have*

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

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

In other words, MLM with self-attention approximately factorizes a $|V| \times |V|$ matrix whose (i, j)-th entry is given by Equation (2). It is important to note that 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 $j \in [|V|+1]$, our problem approximately reduces to a vanilla CBOW. In particular, we always have $\gamma_j^d \approx 1/S$, whence Proposition 14 yields $w_i^\top c_j \approx \frac{p(w_i, c_j)}{p(c_j)} \cdot |V| - 1 \approx \log\left(\frac{p(w_i, c_j)}{p(c_j)}\right) + \log |V|$. Using Proposition 12, we argue that the resulting embeddings approximately form a linear relationship, up to some error terms.

The paraphrasing argument for MLM with self-attention. We first define

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

for every $j \in [|V|+1]$. This means

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

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

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

where $\bar{p}(c_j) = \sum_i \bar{p}(w_i, c_j)$. Note that 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 Lemma 15, which is a version of Lemma 11 for the matrix N, where

$$N_{i,j} = \log\left(\frac{\bar{p}(w_i, c_j)}{\bar{p}(c_j)}\right) + \log |V|.$$

Here, $N_i^\top$ denotes the i-th row of N. The proof is analogous to that of Lemma 11 and is thus omitted.

Lemma 15. *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

$$\bar{\sigma}_j^{\mathcal{W}} = \log\left(\frac{\bar{p}(\mathcal{W} | c_j)}{\prod_{w_i \in \mathcal{W}} \bar{p}(w_i | c_j)}\right),$$

$$\bar{\sigma}_j^{\mathcal{W}_*} = \log\left(\frac{\bar{p}(\mathcal{W}_*|c_j)}{\prod_{w_i \in \mathcal{W}_*} \bar{p}(w_i|c_j)}\right),$$

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

Propositions 16 and 17 are the attention versions of Propositions 12 and 13. The proof of Proposition 16 follows from multiplying both sides of the equations in Lemma 15 by $\tilde{C}^\dagger = (\tilde{C}\tilde{C}^\top)^{-1}C$ (assuming $\tilde{C}$ has full row rank) and setting $\mathcal{W} = \{w_b, w_{a^*}\}$ and $\mathcal{W}_* = \{w_{b^*}, w_a\}$. The proof of Proposition 17 can be found in Appendix L.

Proposition 16. *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 $\mathcal{W} = \{w_b, w_{a^*}\}$ and $\mathcal{W}_* = \{w_{b^*}, w_a\}$.

Proposition 17. *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 $\tilde{c}_r + \tilde{c}_s \approx \tilde{c}_t + \tilde{c}_u$.*

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 this is satisfied, the linear relationships hold for the transformed embeddings $\tilde{c}_i$'s instead of the token embeddings c_i's. 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 approximately have linear relationships for the token embeddings c_i's.

Remarks. *It is easy to see that our result can technically be extended to incorporate positional encodings by considering each (word, position) pair as a unit. In particular, analogies are drawn between (word, position) units.*

H PROOF OF PROPOSITION 9

Proposition 9. *Consider CBOW without negative sampling. Using the same notation as before, we have*

$$w_i^\top c_j \approx \log\left(\frac{p(w_i, c_j)}{p(c_j)}\right) + \log|V|.$$

Proof. For simplicity, we assume that the window size is always $2m$. Consider an instance with i as the center word and

$j \in J$ as the context words. The loss for this instance can be approximated as

$$\begin{aligned}
&-\frac{\sum_{j\in J} w_i^\top c_j}{2m} + \log\left(\sum_{k=1}^{|V|} \exp\left(\frac{\sum_{j\in J} w_k^\top c_j}{2m}\right)\right) \\
&\approx -\frac{\sum_{j\in J} w_i^\top c_j}{2m} + \log\left(\sum_{k=1}^{|V|}\left(1 + \frac{\sum_{j\in J} w_k^\top c_j}{2m} + \frac{(\sum_{j\in J} w_k^\top c_j)^2}{8m^2}\right)\right) \\
&= -\frac{\sum_{j\in J} w_i^\top c_j}{2m} + \log|V| + \log\left(1 + \frac{\sum_{k=1}^{|V|}\left(\sum_{j\in J} w_k^\top c_j\right)}{2m|V|} + \frac{\sum_{k=1}^{|V|}\left(\sum_{j\in J} w_k^\top c_j\right)^2}{8m^2|V|}\right) \\
&\approx -\frac{\sum_{j\in J} w_i^\top c_j}{2m} + \log|V| + \frac{\sum_{k=1}^{|V|}\left(\sum_{j\in J} w_k^\top c_j\right)}{2m|V|} + \frac{\sum_{k=1}^{|V|}\left(\sum_{j\in J} w_k^\top c_j\right)^2}{8m^2|V|} \\
&\leq -\frac{\sum_{j\in J} w_i^\top c_j}{2m} + \log|V| + \frac{\sum_{k=1}^{|V|}\left(\sum_{j\in J} w_k^\top c_j\right)}{2m|V|} + \frac{\sum_{k=1}^{|V|}\left(\sum_{j\in J}(w_k^\top c_j)^2\right)}{4m|V|},
\end{aligned}$$

where we used the Taylor expansions $\exp(x) \approx 1 + x + x^2/2$ and $\log(1+x) \approx x$, as well as the Cauchy-Schwarz inequality. Ignoring the constant $\log|V|$ and multiplying by $2m|V|$, the approximate loss can be written as

$$-|V|\sum_{j\in J} w_i^\top c_j + \sum_{k=1}^{|V|}\left(\sum_{j\in J} w_k^\top c_j\right) + \frac{1}{2}\sum_{k=1}^{|V|}\left(\sum_{j\in J}(w_k^\top c_j)^2\right).$$

Summing this over all instances and only extracting terms which depend on $w_i^\top c_j$, we have the following loss which we want to minimize:

$$\ell(i,j) = -|V| \cdot \#(w_i, c_j) w_i^\top c_j + \#(c_j) w_i^\top c_j + \frac{1}{2}\#(c_j)(w_i^\top c_j)^2.$$

Taking derivative with respect to $w_i^\top c_j$ and setting it to 0 yields

$$w_i^\top c_j = \left(\frac{\#(w_i, c_j)}{\#(c_j)}\right) \cdot |V| - 1 = \left(\frac{p(w_i, c_j)}{p(c_j)} \cdot |V|\right) - 1.$$

The approximation $x \approx \log(1+x)$ completes the proof.

I CORPUS GENERATION PROCESS

1. Consider four subjects (mathematics, statistics, sociology and history) and four adjectives (fun, boring, easy and difficult). Assign scores to each subject which represents the level of each adjective:
   (a) mathematics: (4, 2, 4, 2).
   (b) statistics: (6, 0, 5, 1).
   (c) sociology: (1, 5, 2, 4).
   (d) history: (0, 6, 0, 6).
2. Consider three types of sentence:
   (a) Type 1: I like `subj1` and `subj2`, where `subj1` and `subj2` are independently chosen from the list of subjects with probability $(4/11, 5/11, 1/11, 1/11)$.
   (b) Type 2: `subj1` and `subj2` is `adj`, where `subj1` and `subj2` are independently chosen from the list of subjects with uniform probability, and `adj` is chosen from the list of adjectives with probability proportional to the sum of the scores of `subj1` and `subj2`.
   (c) Type 3: `subj` is `adj1` and `adj2`, where `subj` is chosen from the list of subjects with uniform probability, and `adj1` and `adj2` are independently chosen from the list of adjectives with probability proportional to the score of `subj`.
3. To generate each sentence, we first randomly choose the sentence type with uniform probability. We then form the sentence following the process above.

J PROOF OF LEMMA 11

Lemma 11. *For any word sets $\mathcal{W}, \mathcal{W}_* \subseteq \mathcal{E}$ with the same cardinality, we have*

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

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

Proof. Observe that $p(c_j|\mathcal{W}_*) = \frac{p(\mathcal{W}_*|c_j)p(c_j)}{p(\mathcal{W}_*)}$ and $p(c_j|\mathcal{W}) = \frac{p(\mathcal{W}|c_j)p(c_j)}{p(\mathcal{W})}$, whence $\rho_j^{\mathcal{W},\mathcal{W}_*} = \log\left(\frac{p(c_j|\mathcal{W}_*)}{p(c_j|\mathcal{W})}\right) = \log\left(\frac{p(\mathcal{W}_*|c_j)}{p(\mathcal{W}|c_j)}\right) + \log\left(\frac{p(\mathcal{W})}{p(\mathcal{W}_*)}\right)$. We have

$$\begin{aligned}&\sum_{w_i \in \mathcal{W}_*} M_i - \sum_{w_i \in \mathcal{W}} M_i \\ &= \sum_{w_i \in \mathcal{W}_*} \log\left(\frac{p(w_i, c_j)}{p(c_j)}\right) - \sum_{w_i \in \mathcal{W}} \log\left(\frac{p(w_i, c_j)}{p(c_j)}\right) \\ &= \log \prod_{w_i \in \mathcal{W}_*} p(w_i|c_j) - \log \prod_{w_i \in \mathcal{W}} p(w_i|c_j) \\ &= \log\left(\frac{\prod_{w_i \in \mathcal{W}_*} p(w_i|c_j)}{\prod_{w_i \in \mathcal{W}} p(w_i|c_j)}\right) + \log\left(\frac{p(\mathcal{W}_*)}{p(\mathcal{W}_*)}\right) + \log\left(\frac{p(\mathcal{W})}{p(\mathcal{W})}\right) + \log\left(\frac{p(\mathcal{W}_*|c_j)}{p(\mathcal{W}_*|c_j)}\right) + \log\left(\frac{p(\mathcal{W}|c_j)}{p(\mathcal{W}|c_j)}\right) \\ &= \log\left(\frac{p(\mathcal{W}_*|c_j)}{p(\mathcal{W}|c_j)}\right) + \log\left(\frac{p(\mathcal{W})}{p(\mathcal{W}_*)}\right) + \log\left(\frac{p(\mathcal{W}|c_j)}{\prod_{w_i \in \mathcal{W}} p(w_i|c_j)}\right) - \log\left(\frac{p(\mathcal{W}_*|c_j)}{\prod_{w_i \in \mathcal{W}_*} p(w_i|c_j)}\right) + \log\left(\frac{p(\mathcal{W}_*)}{p(\mathcal{W})}\right) \\ &= \rho_j^{\mathcal{W},\mathcal{W}_*} + \sigma_j^{\mathcal{W}} - \sigma_j^{\mathcal{W}_*} + \delta_j^{\mathcal{W},\mathcal{W}_*}.\end{aligned}$$

Also,

$$\begin{aligned}\rho_j^{\mathcal{W},\mathcal{W}_*} + \delta_j^{\mathcal{W},\mathcal{W}_*} &= \log\left(\frac{p(\mathcal{W}_*|c_j)}{p(\mathcal{W}|c_j)}\right) + \log\left(\frac{p(\mathcal{W})}{p(\mathcal{W}_*)}\right) + \log\left(\frac{p(\mathcal{W}_*)}{p(\mathcal{W})}\right) \\ &= \xi_j^{\mathcal{W},\mathcal{W}_*},\end{aligned}$$

which completes the proof.

K PROOF OF PROPOSITION 13

Proposition 13. *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 $c_r + c_s \approx c_t + c_u$.*

Proof. For any $c_v \in \mathcal{E}$, we have $(w_r + w_s)^\top c_v \approx (w_t + w_u)^\top c_v$. From Proposition 3, this expression can be simplified as $\log p(w_r, c_v) + \log p(w_s, c_v) \approx \log p(w_t, c_v) + \log p(w_u, c_v)$. This implies $\log p(w_v, c_r) + \log p(w_v, c_s) \approx \log p(w_v, c_t) + \log p(w_v, c_u)$. Observe that

$$\begin{aligned}&w_v^\top (c_r + c_s - c_t - c_u) \\ &= (\log p(w_v, c_r) + \log p(w_v, c_s) - \log p(w_v, c_t) - \log p(w_v, c_u)) + \log\left(\frac{p(c_t)p(c_u)}{p(c_r)p(c_s)}\right) \\ &\approx 0 + \log\left(\frac{p(\mathcal{W}_*)}{p(\mathcal{W})}\right) \\ &\approx 0.\end{aligned}$$

Since this holds for every v and W has full row rank, we conclude that $c_r + c_s \approx c_t + c_u$, completing the proof.

L PROOF OF PROPOSITION 16

Proposition 16. 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 $\tilde{c}_r + \tilde{c}_s \approx \tilde{c}_t + \tilde{c}_u$.

Proof. For any $\tilde{c}_v \in \mathcal{E}$, we have $(w_r + w_s)^\top \tilde{c}_v = (w_t + w_u)^\top \tilde{c}_v$. From Appendix G.2, this expression can be simplified as $\log \bar{p}(w_r, c_v) + \log \bar{p}(w_s, c_v) \approx \log \bar{p}(w_t, c_v) + \log \bar{p}(w_u, c_v)$. By the assumption that $\bar{p}(w_i, c_j) \approx \bar{p}(w_j, c_i)$, this implies $\log \bar{p}(w_v, c_r) + \log \bar{p}(w_v, c_s) \approx \log \bar{p}(w_v, c_t) + \log \bar{p}(w_v, c_u)$. Observe that

$$\begin{aligned}
&w_v^\top (\tilde{c}_r + \tilde{c}_s - \tilde{c}_t - \tilde{c}_u) \\
&= (\log \bar{p}(w_v, c_r) + \log \bar{p}(w_v, c_s) - \log \bar{p}(w_v, c_t) - \log \bar{p}(w_v, c_u)) + \log \left(\frac{\bar{p}(c_t)\bar{p}(c_u)}{\bar{p}(c_r)\bar{p}(c_s)} \right) \\
&\approx 0 + \log \left(\frac{\bar{p}(\mathcal{W}_*)}{\bar{p}(\mathcal{W})} \right) \\
&\approx 0.
\end{aligned}$$

Since this holds for every v and W has full row rank, we conclude that $\tilde{c}_r + \tilde{c}_s \approx \tilde{c}_t + \tilde{c}_u$, completing the proof.

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.