Journal of Data Science 0 (0), 1–17
July 2024
DOI: 10.6339/24-JDS1134
Statistical Data Science

Rethinking Attention Weights as Bidirectional Coefficients

Yuxiang Huang[1], Hanfang Yang[1,*], and Xingrui Wang[2]

[1]*School of Statistics, Renmin University of China, Beijing, China*
[2]*Whiting School of Engineering, Johns Hopkins University, Baltimore, USA*

Abstract

Attention mechanism has become an almost ubiquitous model architecture in deep learning. One of its distinctive features is to compute non-negative probabilistic distribution to re-weight input representations. This work reconsiders attention weights as bidirectional coefficients instead of probabilistic measures for potential benefits in interpretability and representational capacity. After analyzing the iteration process of attention scores through backwards gradient propagation, we proposed a novel activation function, TanhMax, which possesses several favorable properties to satisfy the requirements of bidirectional attention. We conduct a battery of experiments to validate our analyses and advantages of proposed method on both text and image datasets. The results show that bidirectional attention is effective in revealing input unit's semantics, presenting more interpretable explanations and increasing the expressive power of attention-based model.

Keywords *attention mechanism; bidirectional coefficients; interpretability*

1 Introduction

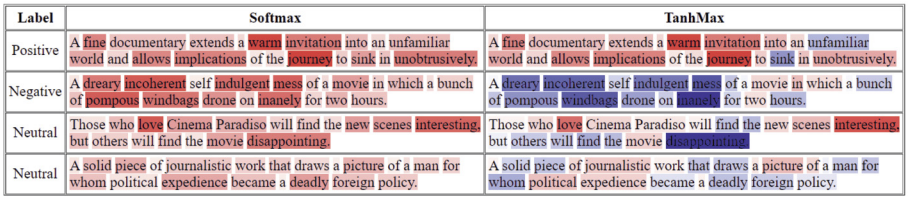
Attention mechanism has proved to be an effective component in deep learning. Considerable efforts have been put into research to take the best advantage of attention, including proposing efficient approximation to softmax (Martins and Astudillo, 2016; Shim et al., 2017; Choromanski et al., 2021; Titsias, 2016; Peng et al., 2021), breaking through softmax bottleneck (Lin, 2021; Kanai et al., 2018; Ganea et al., 2019; Yang et al., 2019) and reducing the quadratic computation and memory footprint of transformers (Zhen et al., 2022; Dehghani et al., 2019; Katharopoulos et al., 2020; Kitaev et al., 2020; Wang et al., 2020). The majority of existing modifications in this field retained the non-negative probabilistic distribution as one of attention's distinctive features and interpreted it as selecting input signals based on importance or relevance. From a novel standpoint, we rethink attention weights as quantitative metrics over input representation and investigate whether attention weights could act as bidirectional coefficients with meaningful positive-or-negative sign.

Our motivation of rethinking attention weights as bidirectional coefficients comes from the potential benefits in interpretability and representational power. The advantages are summarized as the following points:

(1) Finer depiction about input units. Robnik-Sikonja and Bohanec (2018) consider expressive power as a key property of machine learning explanations. As illustrated in Figure 1, both softmax and TanhMax attentions highlight embedded keywords in test samples like "warm" in the positive instance and "dreary" in the negative instance. However, non-negative softmax

*Corresponding author. Email: hyang@ruc.edu.cn.

© 2024 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.
Received November 4, 2023; Accepted April 16, 2024



Label	Softmax	TanhMax
Positive	A fine documentary extends a warm invitation into an unfamiliar world and allows implications of the journey to sink in unobtrusively.	A fine documentary extends a warm invitation into an unfamiliar world and allows implications of the journey to sink in unobtrusively.
Negative	A dreary incoherent self indulgent mess of a movie in which a bunch of pompous windbags drone on inanely for two hours.	A dreary incoherent self indulgent mess of a movie in which a bunch of pompous windbags drone on inanely for two hours.
Neutral	Those who love Cinema Paradiso will find the new scenes interesting, but others will find the movie disappointing.	Those who love Cinema Paradiso will find the new scenes interesting, but others will find the movie disappointing.
Neutral	A solid piece of journalistic work that draws a picture of a man for whom political expedience became a deadly foreign policy.	A solid piece of journalistic work that draws a picture of a man for whom political expedience became a deadly foreign policy.

Figure 1: Explaining the interpretability of bidirectional attention. We train two models on binary text classification datatset SST-2 with softmax or TanhMax activation. Attention weights are visualized on four representative test instances. From top to bottom are respectively positive instance, negative instance, neutral narrative instance and instance with semantic inflection. Red indicates positive weights and blue indicates negative weights with darkness proportional to the absolute value. Bidirectional coefficients could deliver more informative description, such as whether certain tokens suggest a particular label.

weights fail to distinguish difference of the two words in sentiments, only indicating they are relavant or important for model's prediction. In contrast, bidirectional TanhMax weights depict this subtlety with opposite sign, dividing these tokens into separate groups without prior knowledge about words' meanings.

(2) Better explanation about model prediction. An essential criterion for good explanations is to provide qualitative understanding between instance's components (e.g. words in text, patches in an image) and the response (Ribeiro et al., 2016a). In Figure 1, two neutral instances are selected as out of distribution test samples, on which both models trained with 0-1 binary cross entropy loss give dubious predictions with moderate output values around 0.5. It is natural for the last narrative sentence because there is no symbolic token in this instance and attention weights over input unit are relative small in magnitude. While in the third sentence with semantic inflection, softmax attention strongly underscores some keywords, which seems to be inconsistent with model's uncertain prediction. Because when enough tokens are picked out as supporting evidence, the model is supposed to be confident about its classification. On the contrary, TanhMax attention finds that words with disparate semantics coexist in a single sequence and presents a more reasonable explanation about model's uncertain prediction, i.e., a list of marked tokens with both positive signals (in red) and negative signals (in blue) neutralizing each other.

(3) Enhanced representational power. Softmax has long been accused of being a bottleneck of representational capacity of neural networks (Yang et al., 2018; Kanai et al., 2018; Dong et al., 2021). Kanai et al. (2018) revealed this deficiency occurs because softmax uses only exponential functions for nonlinearity. Dong et al. (2021), based on the shift-invariance property of softmax, proves the output of pure self-attention converges doubly exponentially to a rank-1 matrix. Our designed bidirectional attention gets rid of the limitations of log-linearity and shift-invariance from a novel perspective and allows deep models to have better expressive power. These properties will be further clarified in Section 4.

In this work, we focus on designing a well-behaved bidirectional attention and empirically demonstrate its benefits in interpretability on classification task and representational capacity on language modeling. We first analyze the iteration process of a generic attention paradigm with a gradient descent based learning framework (Sun and Lu, 2020) and figure out what role of each model component plays in the update of attention scores. Based on the theoretical result, we

propose *TanhMax* activation function and proves it meets several appealing properties. Through a battery of experiments, we validate our analyses and show the proper performance of proposed method.

2 Related Work

Multiple methods have been proposed to evaluate neural network explanations. One approach was to calculate the relevance based on partial derivatives of the model output. For example, Li et al. (2016a) demonstrated strategies for visualizing compositionality of neural models on NLP tasks and presented how much amount a unit contributes to the final composed meaning from first-order derivatives. Denil et al. (2014) considered dot product between prediction function gradient and word embedding as token relevence and proposed a novel evaluation technique that can be easily applied to labelled documents at scale. Another method to measure importance of single variable is to occlude them in the input and track the difference in the network's output. Li et al. (2016b) proposed a general methodology to interpret and analyze neural model's decision making process by quantifying the influence of erasing certain representation units, such as input word vectors and intermediate neurons. Other methods such as layer-wise relevance propagation (Bach et al., 2015) were also put forward and developed to determine input feature relevance.

Whether attention really helps model by attending input units remain a hot topic in research. Jain and Wallace (2019), Serrano and Smith (2019) doubted softmax distribution over attended-to sequence communicates the relative importance of input features. On the other hand, Vashishth et al. (2019) extended analysis to diverse NLP tasks and showed that attention weights are interpretable and correlate with feature importance measures when they are essential for final prediction and can not be reduced to a gating unit. We look into this issue from a novel perspective by rethinking attention coefficients as bidirectional metrics instead of probabilistic distribution. Nevertheless, we follow these works in experimental design and quantify the interpretability of bidirectional attention weights with similar measurements.

The softmax bottleneck was first revealed by Yang et al. (2018) in language modeling. The authors considered linear-softmax layers as the source of limited expressivity and proposed MoS (mixture of Softmaxes) to tackle this problem. As weighted mixtures of multiple softmax components, MoS improves on both perplexity and rank of output matrices. Kanai et al. (2018) identify the cause of softmax bottleneck by analyzing the output set of log-softmax and put forward sigsoftmax to solve the problem without introducing additional parameter. Several other works (Ganea et al., 2019; Yang et al., 2019) have also proposed lightweight alternatives to the computationally expensive MoS to overcome this bottleneck in representational capacity.

To our knowledge, existing research on bidirectional attention is quite limited. Wang et al. (2018) defined a generic non-local operation, which allow for negative coefficients when using dot-product similarity function. However, non-local operations defined in their paper does not possess required good properties of bidirectional attention described in Section 4 and may fall short to non-local operation of embedded Gaussian version, which is a softmax computation in essence. Zhen et al. (2022) hypothesized negative attention scores delivers negative-correlated contextual information and enforced non-negativity by passing features to a ReLU activation function before computing similarity scores. While our work shows that negative attention can also convey useful messages in semantics and be used to suppress irrelevant noise.

3 Preliminaries

For an input instance $x : \{x_1, x_2, \ldots, x_n\}$, x could be image, sequence or other features and i is the index that enumerates all possible input positions (in space or time). Query Q_i, key K_i and value V_i are obtained by an affine transformation on the x_i. We then define a generic attention operation as follows:

$$s_i = f(Q_i, K_i) \tag{1}$$

$$V = \sum_{\forall i} \frac{g(s_i)}{\sum_{\forall j} h(s_j)} V_i \tag{2}$$

Here f is the pairwise function to compute attention score s_i to represent relationship such as affinity or aversion. V is the output representation of the same size as V_i. The unary function g performs a non-linear transformation of the similarity signal s_i and is normalized by a factor $\mathcal{C}(s) = \sum_{\forall j} h(s_j)$. For numerical stability, the normalizing factor $\mathcal{C}(s)$ is required to be positive as a denominator.

Based on frequentist statistical theory, loss function is defined and estimated as follows:

$$R(\theta, y) = \int_X L(\theta, y(x)) dP_\theta(x) \approx \frac{1}{N} \sum_{k=1}^{N} L(\theta^{(k)}, y(x^{(k)})) \tag{3}$$

Here θ is a fixed but possibly unknown state of nature such as instance label and y is the output value of decision model taking attention operation Eq. (1) and Eq. (2) as its building blocks. $L(\theta, y(X))$ measures the loss caused by model's decisions. X is a vector of observations stochastically drawn from a population, dP_θ is a probability measure over the event space of X. The right-hand side term is an estimator with N denoting sample size and superscript (k) denoting sample order.

There already exist pairwise similarity functions having the potential to output bidirectional coefficients, such as scaled dot-product score function (Vaswani et al., 2017) and non-local operation (Wang et al., 2018). However, containing both positive and negative real numbers in function range does not necessarily means model would generate interpretable bidirectional weights. For example, Sun and Lu (2020) proved that the commonly-used scaled dot-product attention would yield positive attention scores and attention weights for tokens of opposite polarities. In order to find out the necessary conditions for bidirectional attention at first, we dive into the iteration process of attention coefficients and obtain the following Proposition 1.

Proposition 1. *For a simple model $y(x) = \sigma(V^T W)$ consisting of one single attention operation as defined by Eq. (1) and Eq. (2), the update formula of attention score s_i at time step τ is:*

$$\frac{ds_i}{d\tau} = -\underbrace{\{\frac{\partial s_i}{\partial Q_i}(\frac{\partial s_i}{\partial Q_i})^T + \frac{\partial s_i}{\partial K_i}(\frac{\partial s_i}{\partial K_i})^T\}}_{\Delta_I} \underbrace{\sum_{\forall j} \pi_j V_j^T W \sigma' \frac{\partial \hat{R}}{\partial y}}_{\Delta_{II}} \tag{4}$$

where $\pi_j = I[j = i] \cdot \frac{g'(s_i)}{\sum_k h(s_k)} - w_j \cdot \frac{h'(s_i)}{\sum_k h(s_k)}$ with $w_j = \frac{g(s_j)}{\sum_{\forall k} h(s_k)}$ being attention weight. σ is the activation function at output layer and σ' is its first-order derivative.

We leave the complete proof of Proposition 1 in the supplementary material. The update formula (4) described above help us figure out how each model component influences the iteration process of attention scores and could provide the following implications:

(1) Score alignment function f has little effect on the update direction. Taking query and key as inputs, f computes attention score s_i. In Δ_I of Eq. (4), the partial derivatives of s_i with respect Q_i and K_i exist in the form of self-dot-product, producing non-negative value. Thus, the update direction is not directly related to the specific form of score alignment function.

(2) Attention score s_i is updated in reference to the context. $o_i \triangleq V_i^T W$, defined as token-level polarity score in Sun and Lu (2020), measures token x_i's influence to the output value. In Δ_{II} of Eq. (4), the first term of π_j selects target token's polarity score by an indicative function and the second term of π_j computes instance-level polarity score by weighting o_j along the whole instance. Thus, attention module would renew target token x_i's score s_i by comparing its representation against context representation. This is in accordance with the long-range dependencies modeling of attention mechanism.

(3) The update of attention score s_i is task-specific. The types of final activation σ and loss measuring function L depend on what kind of task our model aims to solve. The multipliers σ' and $\frac{\partial \hat{R}}{\partial y}$ in Eq. (4) are basically determined when the downstream objective is given and would apply indiscriminate influence to every input unit of the same instance regardless of unit's polarity.

Based on above analyses, we find that the score alignment function and downstream task type do not play a significant role in the update process and loss function. Therefore, we have to turn to the non-linear function $g(\cdot)$ and normalizing factor $\mathcal{C}(s)$ defined in Eq. (2) in order to meet the purpose of bidirectional attention. The two operators combined can be regarded as an activation function $\phi(\cdot)$ over attention scores as follows:

$$[\phi(s)]_i \triangleq \frac{g(s_i)}{\mathcal{C}(s)} = \frac{g(s_i)}{\sum_{\forall k} h(s_k)} \tag{5}$$

4 Proposed Method

Based on the above derivation, it is required to properly design an activation function to allow for meaningful positive and negative coefficients. In order to achieve bidirectional coefficients and maintain model's expressive capacity, we propose four appealing properties that the activation function defined in Eq. (5) needs to satisfy as listed below. These favorable properties are proposed from the perspectives of interpretability and representational power based on previous analyses and existing research, which are specially suitable for bidirectional attention and may not be the prerequisites for activation functions aimed at solving other tasks.

(1) Monotonically increasing. $g(\cdot)$ should be monotonically increasing so that $\phi(\cdot)$ becomes a smoothed version of the argmax function (Bridle, 1989; Abramson et al., 1963). When $g(s_i)$ is monotonically increasing, we can obtain the value order of $\phi(s_i)$ by simply comparing the elements of s_i and easily figure out the relative importance of input units.

(2) Log-nonlinearity. According to Kanai et al. (2018), softmax can be the bottleneck of neural network's representational power because the exponential function in the numerator is a linear mapping after logarithmic transformation, which will cause the projected input vector space to have reduced dimensions. In order to overcome this drawback, $log(g(z))$ in Eq. (5) is supposed to be nonlinear.

(3) Origin-symmetry. It seems obvious for bidirectional attention to require $g(\cdot)$ to have origin-point symmetry so that generated coefficients could be either positive or negative. More importantly, this property not only extends function range to prevent post-transformation dimension reduction, but also facilitates interpretability by making positive coefficients and neg-

ative coefficients comparable in magnitude.

(4) Contextual normalization. The newly proposed activation function is supposed to compare the computed input signal $g(s_i)$ at the position i with the aggregated context information $\mathcal{C}(s)$ to select useful interdependency. Zhen et al. (2022) observed that models with re-weighting scheme converge faster and generalize better to downstream tasks. They explained it as normalization amplifies the correlated pairs, which might be helpful to identify useful patterns.

We propose a novel activation function, TanhMax, defined in Eq. (6) as a solution to bidirectional attention. We show that TanhMax meets all the above four characteristics in supplementary material and prove that TanhMax can enhance representational power by breaking softmax bottleneck in Proposition 2.

Definition 1. TanhMax is defined as:

$$[\phi_t(s)]_i = \frac{\exp(s_i) - \exp(-s_i)}{\sum_{\forall k}[\exp(s_k) + \exp(-s_k)]} \tag{6}$$

TanhMax could be regarded as a special case of Eq. (5) with $g(s_i) = e^{s_i} - e^{-s_i}$ and $h(s_i) = e^{s_i} + e^{-s_i}$. By plugging Eq. (6) into Eq. (4), the update equation of TanhMax attention scores can be obtained as follows:

$$\frac{ds_i}{d\tau} = \{\frac{\partial s_i}{\partial Q_i}(\frac{\partial s_i}{\partial Q_i})^T + \frac{\partial s_i}{\partial K_i}(\frac{\partial s_i}{\partial K_i})^T\} \underbrace{[\widetilde{w}_i o_i - \hat{w}_i \sum_{j=1}^{n} \hat{w}_j o_j]}_{\Delta'_{II}} \sigma' \frac{\partial \hat{R}}{\partial y} \tag{7}$$

here the two weighted factors respectively are:

$$\widetilde{w}_i = \frac{\exp(s_i) + \exp(-s_i)}{\sum_{k=1}^{n}[\exp(s_i) + \exp(-s_i)]}, \tag{8}$$

$$\hat{w}_i = \frac{\exp(s_i) - \exp(-s_i)}{\sum_{k=1}^{n}[\exp(s_i) + \exp(-s_i)]} \tag{9}$$

Based on Eq. (7), we could peek into the process about how TanhMax selects tokens with particular polarity. The first term of Δ'_{II} shares the same sign of o_i since $\widetilde{w}_i$ is always positive. During training, model gradually increases polarity score o_i for tokens with high frequency in positive instances and decreases o_i for those with high frequency in the negative in order to make correct classification. The second term of Δ'_{II} re-weight token-level polarity score o_i to get contextual representation, which would normally have relatively small value because most tokens in an instance are neutral and may cancel out each other in polarity. Thus, attention score s_i for tokens of different polarities would receive opposite changes every iteration because of opposite Δ'_{II}. For neutral tokens, the scores s_i will not have significant changes because corresponding Δ'_{II} and polarity score o_i remain close to zero.

The update equation (7) has showed that TanhMax attention could yield meaningful bidirectional scores with opposite signs indicating different polarities, which underlines the improved interpretability of proposed method. Meanwhile, the enhanced representational capacity is another advantage. We have argued that TanhMax could break softmax bottleneck by log-nonlinear transformation. We further demonstrate this property by examining the output range of a activation function and prove that the projected vector space of softmax is only a subset of TanhMax when certain conditions are met as described in Proposition 2.

Proposition 2. *Let $z \in \mathbb{S}$ be the input of TanhMax $\phi_t(\cdot)$ and softmax $\phi_s(\cdot)$. If the $\mathbb{S}$ is a d dimensional vector space and $\mathbf{1} \in \mathbb{S}$, the range of softmax is a subset of the range of TanhMax.*

$$\{\phi_s(z)|z \in \mathbb{S}\} \subset \{\phi_t(z)|z \in \mathbb{S}\} \tag{10}$$

Softmax has been revealed to be a bottleneck of representational power of neural networks (Yang et al., 2018). Proposition 2 shows that TanhMax could overcome this limitation by expanding the projected domain and indicates that TanhMax has the enhanced representational capacity compared with softmax. We leave the detailed proof of Proposition 2 in the supplementary material. Experiments in Section 5.3 could validate our analysis by showing that TanhMax could slow down the rank collapse phenomenon.

5 Experiments

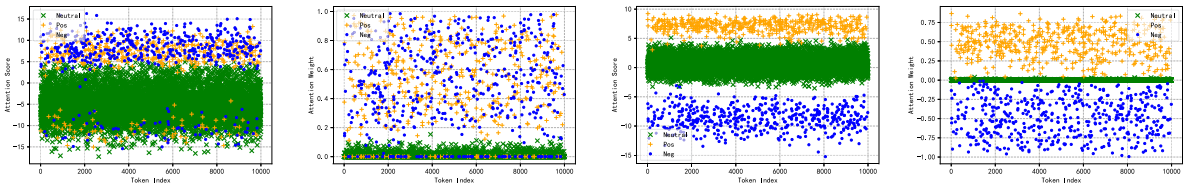
To validate our analyses and the favorable properties of proposed method, we conducted a battery of experiments on synthetic datasets and real datasets. The generation of synthetic data and the description of four text datasets and two images datasets could be found in the supplementary material.

5.1 Visualization of Attention Coefficients

In order to demonstrate that our proposed method could meet the requirements of bidirectional attention and deliver finer depictions about input units, we train two models with either Softmax or TanhMax activation function on binary text classification task and visualize the learned attention coefficients (i.e. attention scores or attention weights) in Figure 2. In this experiment, our model has one single attention module, adopts dropout (Srivastava et al., 2014) to prevent overfitting and uses sigmoid function $\sigma = \frac{\exp(\cdot)}{1+\exp(\cdot)}$ at the final output. During training, Adam optimizer (Kingma and Ba, 2015) is used for gradient descent to minimize binary cross entropy loss. All the parameters are learned from scratch to eliminate the disturbance of any prior information. For the same reason, we choose to initialize word embeddings with a uniform distribution from -0.1 to 0.1 instead of using pre-trained word embeddings. Other experimental settings can be found in the supplementary.

We adopt metric γ_e defined by Sun and Lu (2020) to divide tokens into three groups of polarities:

$$\gamma_e = \frac{f_e^+ - f_e^-}{f_e^+ + f_e^-} \tag{11}$$



(a) Softmax scores (b) Softmax weights (c) TanhMax scores (d) TanhMax weights

Figure 2: Scatter plots of attention coefficients. This experiment is conducted on synthetic dataset. Results on real datasets are presented in the supplementary material. Colors represent different token polarities with orange positive, blue negative and green neutral.

where f_e^+ and f_e^- refer to the frequencies in the positive and in the negative instances respectively. If $\gamma_e \in (0.5, 1)$ and $f_e^+ > 5$, the token will be regarded as a "positive token". If $\gamma_e \in (-1, -0.5)$ and $f_e^- > 5$, the token will be considered as a "negative token". If $\gamma_e \in (-0.1, 0.1)$ and $|f_e^+ - f_e^-| < 5$, the token will be treated as a "neutral token". The information from γ_e about the association between the token e and instance labels is not fed into model during training.

Figure 2 shows part of the results for learned attention coefficients. Neutral tokens basically have values around zero in either case. Softmax-based model gives both positive tokens and negative tokens relatively large coefficients, failing to capture more detailed difference in semantics. While, TanhMax-based model generally gives positive tokens positive coefficients and negative tokens negative coefficients, which shows TanhMax could meet the bidirectional requirements. We list selected tokens by two models in supplementary material to further demonstrate bidirectional interpretability.

5.2 Interpretability of Attention Weights

To better quantify the enhanced interpretability of proposed method, we empirically compare softmax and TanhMax attention on several explanation measurements (Jain and Wallace, 2019): (1) Kendall correlation τ_g between attention weights and gradient-based feature importance measures; (2) Kendall correlation τ_{loo} between attention weights and feature erasure or emphleave-one-out (LOO) measures; (3) median change in output value by randomly permuting original attention weights; (4) median change in output value by randomly sampling new attention weights. On feature occlusion measurements, we use two types of Total Variance Distance (TVD) to quantify the change between output distributions: (1) $\text{TVD}^{abs}(\hat{y}_1, \hat{y}_2) = \sum_{\forall k} |\hat{y}_{1k} - \hat{y}_{2k}|$ and (2) $\text{TVD}^{sgn}(\hat{y}_1, \hat{y}_2) = \sum_{\forall k} (\hat{y}_{1k} - \hat{y}_{2k})$ so that change in magnitude and direction can both be considered. The models we used here are the same to those introduced in Section 5.1.

For gradient-based experiments, we obtain importance scores by computing the partial derivatives of the target variable with respect to the input variable $\frac{\partial \hat{y}}{\partial x_i}$ (Samek et al., 2019). For occlusion-based experiments, we denote the input resulting from removing the unit at position i in x by x_{-i} and zero out the weight of removed unit so that its influence will not be considered in new prediction $\hat{y}(\boldsymbol{x}_{-i})$. For permutation and randomization, the median value is computed by one hundred independent tests for each instance.

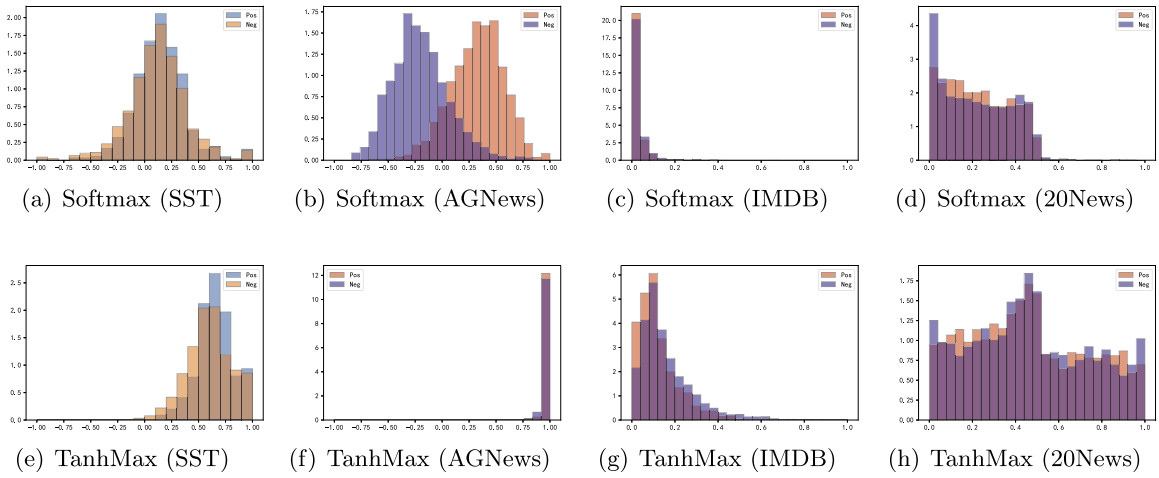
The results demonstrated in Table 1 and Figure 3 are all obtained on test datasets. Table 1 lists part of the statistics from our experiments on binary text classification tasks. Column **Mean±Std.** presents the mean and standard deviation of Kendall correlation. Column **Sig. Frac.** reports the fraction of instances for which this correlation is statistically significant with a significance level $\alpha = 0.01$. Results in Table 1 show that: (1) substituting softmax with TanhMax leads to a notable increment in Kendall correlation coefficients and the fraction of significantly correlated instances on most datasets, indicating that TanhMax attention weights are more representative to important input units; (2) sampling new TanhMax attention weights from distribution $U(-1, 1)$ leads to a larger change in output value than sampling new softmax attention weights from $U(0, 1)$, whereas randomly permuting TanhMax attention weights leads to equivalent or larger variation to that of softmax. Observation (2) seems reasonable for TanhMax weights are distributed in a larger interval, adding randomness to TanhMax weights would bring more significant disturbance to the final predictions. Figure 3 induces the same conclusions, which visualizes the statistics in Table 1.

Table 1: Experimental results on interpretability of attention weights. Column gradient τ_g and column occlusion τ_{loo}^{sgn} respectively present Kendall correlation on gradient-based and erasure-based importance measures with **Sig. Frac.** showing the fraction of instances for which correlation is statistically significant. Columns permutation and randomization reports the median change in output value caused by randomly permuting original weights and randomly sampling new attention weights.

	Gradient τ_g		Occlusion τ_{loo}^{sgn}		Permute	Random
Activation	**Mean±Std**	**Sig. Frac.**	**Mean±Std**	**Sig. Frac.**	**Mean±Std**	**Mean±Std**
SST dataset with class 0						
Softmax	0.34±0.26	0.42	−0.12±0.28	0.20	0.19±0.13	0.25±0.24
TanhMax	0.56±0.23	0.79	0.86±0.19	0.92	0.18±0.22	0.50±0.46
SST dataset with class 1						
Softmax	0.29±0.25	0.31	0.24±0.32	0.32	0.27±0.21	0.31±0.31
TanhMax	0.60±0.21	0.88	0.66±0.28	0.65	0.28±0.21	0.52±0.46
IMDB dataset with class 0						
Softmax	−0.18±0.14	0.12	−0.34±0.12	0.52	0.15±0.12	0.26±0.22
TanhMax	−0.13±0.21	0.18	0.98±0.02	1.0	0.18±0.14	0.48±0.29
IMDB dataset with class 1						
Softmax	−0.25±0.14	0.26	−0.36±0.11	0.60	0.13±0.10	0.24±0.18
TanhMax	−0.32±0.19	0.49	0.98±0.01	1.0	0.15±0.11	0.48±0.27
AGNews dataset with class 0						
Softmax	0.30±0.17	0.35	−0.28±0.27	0.40	0.21±0.13	0.11±0.20
TanhMax	0.74±0.10	1.0	0.83±0.10	1.0	0.18±0.13	0.48±0.30
AGNews dataset with class 1						
Softmax	0.27±0.17	0.25	0.25±0.25	0.33	0.25±0.13	0.18±0.28
TanhMax	0.65±0.12	1.0	0.84±0.11	1.0	0.12±0.10	0.44±0.28
20News dataset with class 0						
Softmax	0.77±0.09	1.0	−0.12±0.25	0.19	0.17±0.13	0.29±0.22
TanhMax	0.43±0.21	0.65	0.89±0.24	0.88	0.13±0.04	0.51±0.44
20News dataset with class 1						
Softmax	0.78±0.10	1.0	0.11±0.29	0.22	0.15±0.13	0.27±0.21
TanhMax	0.54±0.20	0.84	0.63±0.34	0.54	0.13±0.05	0.54±0.45

5.3 Representational Power

In Proposition 2, we show that TanhMax could prevent post-transformation dimension reduction and break softmax bottleneck. In this section, we examine the representational power of proposed method by empirically proving that TanhMax is effective in mitigating rank collapse of deep attention model and could achieve equivalent or better performance compared with other activation functions on classification and language modeling tasks.



(a) Softmax (SST) (b) Softmax (AGNews) (c) Softmax (IMDB) (d) Softmax (20News)

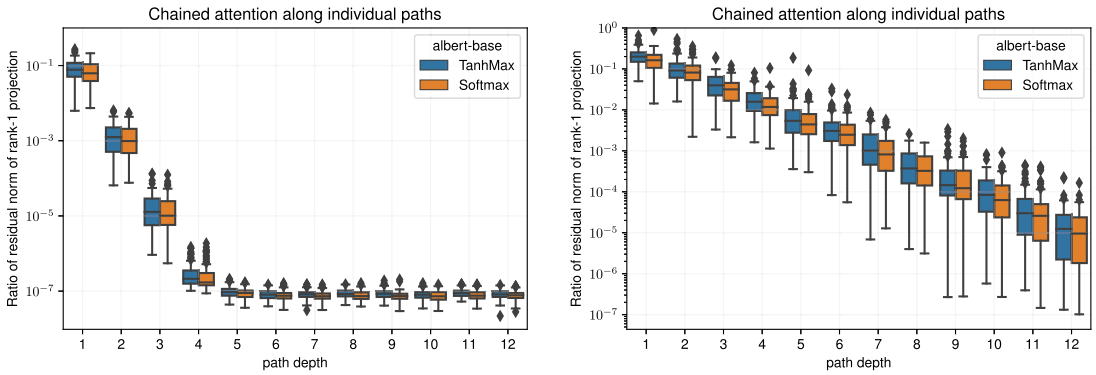
(e) TanhMax (SST) (f) TanhMax (AGNews) (g) TanhMax (IMDB) (h) TanhMax (20News)

Figure 3: Plots on interpretability measures of attention weights. Columns from left to right are respectively histogram of gradient-based Kendall correlation, histogram of emphleave-one-out Kendall correlation, histogram of median output change by permutation and violin plots of median output change by sampling. Captions beneath plots show the type of activation function and the dataset on which results are obtained from. Different colors represent different instance labels. Datasets are denoted parenthetically. More plots can be found in supplementary material.

Based on shift-invariant characteristic of softmax, Dong et al. (2021) proved that self-attention possesses a strong inductive bias towards "token uniformity" and the output converges doubly exponentially to a rank-1 matrix. We argue this rank collapse phenomenon is in essence consistent with softmax bottleneck, which maps input vectors into a lower dimensional space. TanhMax activation is not shift-invariant and would not leak information by disregarding terms that provide a constant contribution across rows, which could partly explain its enhanced representational power. To better illustrate this advantage, we follow Dong et al. (2021) and conduct experiments to compare the rank collapse phenomenon of softmax and TanhMax. The results are demonstrated in Figure 4, where the y-axis is the relative norm of residual defined in the referenced work and the x-axis indicate the depth of attention model. As we can see, softmax attention converges rapidly to a rank-1 matrix regardless of initialization or pretrained model, while TanhMax attention slows down this degeneration process. The result shows that the rank collapse of deep networks could be mitigated when replacing softmax with TanhMax.

Table 2 compares the prediction accuracy of softmax and TanhMax on classification task. The text classifiers used here are the same to those introduced in Section 5.1 and the results on image classification is obtained by training a ViT model (Dosovitskiy et al., 2021). Table 2 shows that TanhMax is on par with or outperforms softmax in classification precision.

To better demonstrate the enhanced representational power of proposed method, we run experiments of language modeling task on four text dataset. Table 4 lists the test accuracy of six-layer bert models (Devlin et al., 2019) to compare TanhMax with other activation functions. Sigsoftmax and sigmoid are other log-nonlinear activation functions suggested in (Kanai et al., 2018) to overcome softmax bottleneck. Dot product represents a specific form of non-local operation (Wang et al., 2018). Cosformer represents an architecture proposed in Zhen et al. (2022) as an improved variant of softmax. TanhMax is the only activation function that has the four good properties defined in Section 4 as illustrated in the comparative Table 3. The detailed



(a) initialization (b) pretrained

Figure 4: Comparison of softmax and TanhMax on the rank collapse phenomenon. Plots visualize the relative norm of the residual along depth for transformer.

Table 2: Comparison on prediction accuracy. We conduct experiments to compare the precision of softmax attention and TanhMax attention on six real datasets. SST, IMDB, AGNews and 20News are used for binary text classification. Cifar10 and Cifar100 are used for multi-label image classification. This table presents classification precision on test set.

	SST	**IMDB**	**AGNews**	**20News**	**Cifar10**	**Cifar100**
Softmax	0.834	0.732	0.892	0.843	0.902	0.668
TanhMax	0.872	0.747	0.905	0.850	0.887	0.662

Table 3: Comparative table to illustrate how different activation functions align with or diverge from the four identified properties.

Name	Monotonicity	Log-nonlinearity	Normalization	Origin-symmetry
Softmax	✓		✓	
ReLU	✓	✓		
Dot product	✓	✓		✓
Sigsoftmax	✓	✓	✓	
Cosformer	✓	✓	✓	
TanhMax	✓	✓	✓	✓

description about the activation functions mentioned above can be found in the supplementary material. Table 4 shows that none of softmax-based models reports the top performance, while TanhMax function achieved highest prediction precision on most experiments. Even if not optimal, accuracy of TanhMax-based model is among the top.

Based on these experiments, we conclude that TanhMax is more effective than softmax in eschewing model degeneration and extending model's representational capacity.

Table 4: Results of language modeling experiments. Statistics on validation dataset and test dataset are reported separately. The highest prediction accuracy are underscored in bold. Other activation functions are used as comparison.

dataset	Softmax	ReLU	Dot product	Sigsoftmax	Cosformer	TanhMax
Validation						
SST	76.1±0.10	76.2±0.29	75.1±0.20	75.8±0.17	75.7±0.18	**76.2±0.15**
IMDB	28.9±0.11	31.2±0.21	29.3±0.13	32.5±0.15	**35.2±0.19**	33.4±0.12
20News	35.4±0.12	41.2±0.19	34.2±0.14	38.8±0.16	41.5±0.15	**42.9±0.13**
AGNews	65.8±0.12	66.3±0.17	65.5±0.11	65.9±0.16	66.3±0.15	**67.5±0.15**
Test						
SST	73.6±0.14	73.4±0.26	73.5±0.21	73.8±0.15	74.8±0.23	**76.1±0.21**
IMDB	29.5±0.09	30.3±0.18	30.3±0.15	32.7±0.13	**34.6±0.21**	33.9±0.15
20News	37.8±0.11	40.5±0.23	32.8±0.19	39.2±0.12	41.1±0.16	**41.9±0.17**
AGNews	65.4±0.13	64.1±0.14	63.9±0.15	65.5±0.15	65.5±0.17	**66.2±0.20**

5.4 Extension: Explanation on Images

We also investigate our proposed method on static image recognition. We experiment with the ViT baseline (Dosovitskiy et al., 2021) on Cifar 100 datasets. To examine the interpretability of bidirectional attention, we plot TanhMax attention weights in heatmaps.[1] In Figure 5, positive weights and negative weights are respectively visualized in the middle row and the bottom row. The darkness is proportional to the magnitude.

Based on Figure 5, it can be seen that positive attention and negative attention focus on different regions of the pictures. In column one, positive attention highlights bird head, while the negative concentrates on the twig. In column two, positive attention covers head and body of the dog, while negative attention mainly stays in the top margin. In column three, negative weights stay mainly on the top of background while the positive focus on the bird's head and body. In column four, airframe and engines of the plane are highlighted by the positive weights, while the background is veiled by the negative. It seems that positive weights are searching for the evidence or distinctive features to correctly classify pictures and the negative weights are suppressing irrelevant signals to reduce noise.

6 Discussion

In this work, we focus on designing bidirectional attention and examine its advantages in interpretability and representational power. Our inspiration comes from: (1) linear regression model with bidirectional coefficients is commonly used in local interpretable model-agnostic explanations (Ribeiro et al., 2016b) and (2) Bert (Devlin et al., 2019) achieved state-of-the-art performance on a large suite of tasks by pretraining deep bidirectional representations on both left and right context. After analyzing the iteration process of attention scores, we find that the key to bidirectional property is to design a special activation function. Therefore, we propose TanhMax as an effective solution and prove that it have several appealing

[1] We visualized attention in the way described by https://github.com/jacobgil/vit-explain.

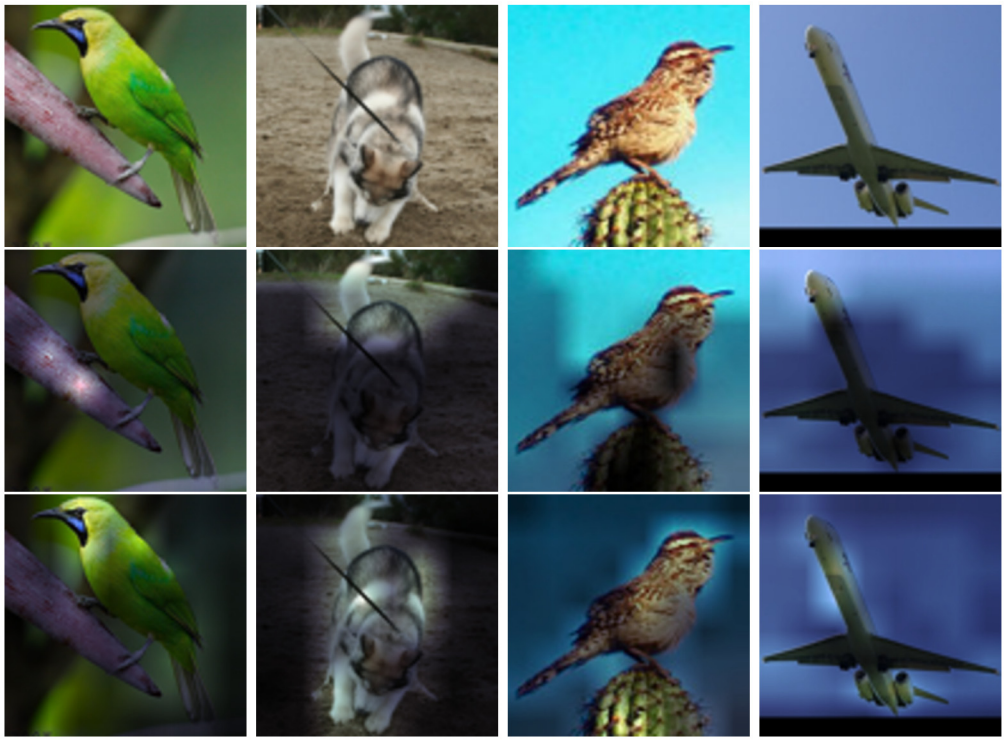


Figure 5: Visualization of TanhMax attention weights on Cifar100 images. Each columns represents one category, from left to right: bird, dog, boat and plane. Top row presents original images. Negative weights are visualized in the middle and positive weights in the bottom. The darkness is proportional to the absolute value of attention.

characteristics. Through a battery of experiments, we validate our analyses and demonstrate TanhMax's advantages in interpretability and representational capacity, especially compared with softmax.

One major purpose of our work is to nourish the development of structures or methods for improving the transparency and interpretability of transformer-based models, which is a long-standing concern in deep learning. Our proposed architecture works well for transformers with images and texts classification objectives, but there still remain potential improvements in interpretability on large datasets with low signal-to-noise ratio and its performance is not sufficiently inspected on other complicate tasks, such as distinguishing up-regulated and down-regulated genes in disease prediction relied on transcriptomics data. Taking into account the above concerns and other potential limitations that our methods may have, in the future, we will consider and study the following aspects: 1) the effect of our proposed methods on large language models like LLaMa (Touvron et al., 2023); 2) evaluation with other classical or realistic benchmark datasets with instances of different tasks; 3) better model explanation algorithm by integrating bidirectional coefficients with other information, such as gradient to the input unit (Selvaraju et al., 2017), instead of using attention coefficients alone.

Supplementary Material

The supplementary materials include: proof of propositions, description of activation functions, detailed experiment setting and additional experiment results. Our Python code in experiment section is also available on Github at https://github.com/BruceHYX/bidirectional_attention.

Acknowledgment

The authors are grateful to the Editor, Associate Editor, and two referees for many helpful comments.

Funding

This research was partially supported by the Major Project of the MOE (China) National Key Research Bases for Humanities and Social Sciences (22JJD910003).

References

Abramson NM, Braverman DJ, Sebestyen GS (1963). Pattern recognition and machine learning. *IEEE Transactions on Information Theory*, 9(4): 257–261. https://doi.org/10.1109/TIT.1963.1057854

Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7): e0130140. https://doi.org/10.1371/journal.pone.0130140

Bridle JS (1989). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In: *Neurocomputing – Algorithms, Architectures and Applications, Proceedings of the NATO Advanced Research Workshop on Neurocomputing Algorithms, Architectures and Applications*, Les Arcs, France, February 27–March 3, 1989 (F Fogelman-Soulié, J Hérault, eds.), volume 68 of *NATO ASI Series*. 227–236. Springer.

Choromanski KM, Likhosherstov V, Dohan D, Song X, Gane A, Sarlós T, et al. (2021). Rethinking attention with performers. In: *9th International Conference on Learning Representations, ICLR 2021*, Virtual Event, Austria, May 3–7, 2021 (S Mohamed, K Hofmann, A Oh, N Murray, I Titov, eds.), OpenReview.net.

Dehghani M, Gouws S, Vinyals O, Uszkoreit J, Kaiser L (2019). Universal transformers. In: *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 6–9, 2019 (T Sainath, A Rush, S Levine, K Livescu, S Mohamed, eds.), OpenReview.net.

Denil M, Demiraj A, De Freitas N (2014). Extraction of salient sentences from labelled documents. *arXiv preprint: https://arxiv.org/abs/1412.6815*.

Devlin J, Chang M, Lee K, Toutanova K (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers) (J Burstein, C Doran, T Solorio, eds.), 4171–4186. Association for Computational Linguistics.

Dong Y, Cordonnier J, Loukas A (2021). Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In: *Proceedings of the 38th International Conference*

on Machine Learning, ICML 2021, Virtual Event, July 18–24, 2021 (M Meila, T Zhang, eds.), volume 139 of *Proceedings of Machine Learning Research*, 2793–2803. PMLR.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In: *9th International Conference on Learning Representations, ICLR 2021*, Virtual Event, Austria, May 3–7, 2021 (S Mohamed, K Hofmann, A Oh, N Murray, I Titov, eds.), OpenReview.net.

Ganea O, Gelly S, Bécigneul G, Severyn A (2019). Breaking the softmax bottleneck via learnable monotonic pointwise non-linearities. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, Long Beach, California, USA, June 9–15, 2019 (K Chaudhuri, R Salakhutdinov, eds.), volume 97 of *Proceedings of Machine Learning Research*. 2073–2082. PMLR.

Jain S, Wallace BC (2019). Attention is not explanation. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers) (J Burstein, C Doran, T Solorio, eds.), 3543–3556. Association for Computational Linguistics.

Kanai S, Fujiwara Y, Yamanaka Y, Adachi S (2018). Sigsoftmax: Reanalysis of the softmax bottleneck. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, Montréal, Canada, December 3–8, 2018 (S Bengio, HM Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, eds.), 284–294.

Katharopoulos A, Vyas A, Pappas N, Fleuret F (2020). Transformers are rnns: Fast autoregressive transformers with linear attention. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, Virtual Event, July 13–18, 2020 (D Blei, H Daume, A Singh, eds.), volume 119 of *Proceedings of Machine Learning Research*, 5156–5165. PMLR.

Kingma DP, Ba J (2015). Adam: A method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 7–9, 2015 (Y Bengio, Y LeCun, eds.), *Conference Track Proceedings*.

Kitaev N, Kaiser L, Levskaya A (2020). Reformer: The efficient transformer. In: *8th International Conference on Learning Representations, ICLR 2020* (A Rush, S Mohamed, D Song, K Cho, M White, eds.), Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net.

Li J, Chen X, Hovy EH, Jurafsky D (2016a). Visualizing and understanding neural models in NLP. In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego California, USA, June 12–17, 2016 (K Knight, A Nenkova, O Rambow, eds.), 681–691. The Association for Computational Linguistics.

Li J, Monroe W, Jurafsky D (2016b). Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

Lin Y (2021). Breaking the softmax bottleneck for sequential recommender systems with dropout and decoupling. *CoRR*, abs/2110.05409.

Martins AFT, Astudillo RF (2016). From softmax to sparsemax: A sparse model of attention and multi-label classification. In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, New York City, NY, USA, June 19–24, 2016 (M Balcan, KQ Weinberger, eds.), volume 48 of *JMLR Workshop and Conference Proceedings*. 1614–1623. JMLR.org.

Peng H, Pappas N, Yogatama D, Schwartz R, Smith NA, Kong L (2021). Random feature

attention. In: *9th International Conference on Learning Representations, ICLR 2021*, Virtual Event, Austria, May 3–7, 2021 (S Mohamed, K Hofmann, A Oh, N Murray, I Titov, eds.), OpenReview.net.

Ribeiro MT, Singh S, Guestrin C (2016a). "why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August 13–17, 2016 (B Krishnapuram, M Shah, AJ Smola, CC Aggarwal, D Shen, R Rastogi, eds.), 1135–1144. ACM.

Ribeiro MT, Singh S, Guestrin C (2016b). "why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego California, USA, June 12–17, 2016 (K Knight, A Nenkova, O Rambow, eds.), 97–101. The Association for Computational Linguistics.

Robnik-Sikonja M, Bohanec M (2018). Perturbation-based explanations of prediction models. In: *Human and Machine Learning - Visible, Explainable, Trustworthy and Transparent* (J Zhou, F Chen, eds.), In: *Human-Computer Interaction Series*, 159–175. Springer.

Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K (Eds.) (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision* (K Ikeuchi, G Medioni, M Pelillo, eds.), 618–626.

Serrano S, Smith NA (2019). Is attention interpretable? In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, Florence, Italy, July 28–August 2, 2019, Volume 1: Long Papers (A Korhonen, DR Traum, L Màrquez, eds.), 2931–2951. Association for Computational Linguistics.

Shim K, Lee M, Choi I, Boo Y, Sung W (2017). Svd-softmax: Fast softmax approximation on large vocabulary neural networks. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Long Beach, CA, USA, December 4–9, 2017 (I Guyon, U von Luxburg, S Bengio, HM Wallach, R Fergus, SVN Vishwanathan, R Garnett, eds.), 5463–5473.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1): 1929–1958.

Sun X, Lu W (2020). Understanding attention for text classification. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, Online, July 5–10, 2020 (D Jurafsky, J Chai, N Schluter, JR Tetreault, eds.), 3418–3428. Association for Computational Linguistics.

Titsias MK (2016). One-vs-each approximation to softmax for scalable estimation of probabilities. In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, Barcelona, Spain, December 5–10, 2016 (DD Lee, M Sugiyama, U von Luxburg, I Guyon, R Garnett, eds.), 4161–4169.

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint: https://arxiv.org/abs/2302.13971*.

Vashishth S, Upadhyay S, Tomar GS, Faruqui M (2019). Attention interpretability across NLP

tasks. *CoRR*, abs/1909.11218.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017). Attention is all you need. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Long Beach, CA, USA, December 4–9, 2017 (I Guyon, U von Luxburg, S Bengio, HM Wallach, R Fergus, SVN Vishwanathan, R Garnett, eds.), 5998–6008.

Wang S, Li BZ, Khabsa M, Fang H, Ma H (2020). Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768.

Wang X, Girshick RB, Gupta A, He K (2018). Non-local neural networks. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18–22, 2018 (MS Brown, B Morse, S Peleg, eds.), 7794–7803. Computer Vision Foundation / IEEE Computer Society.

Yang Z, Dai Z, Salakhutdinov R, Cohen WW (2018). Breaking the softmax bottleneck: A high-rank RNN language model. In: *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, April 30–May 3, 2018 (Y Bengio, Y LeCun, T Sainath, eds.), Conference Track Proceedings. OpenReview.net.

Yang Z, Luong T, Salakhutdinov R, Le QV (2019). Mixtape: Breaking the softmax bottleneck efficiently. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019,* Vancouver, BC, Canada, December 8–14, 2019 (HM Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, EB Fox, R Garnett, eds.), 15922–15930.

Zhen Q, Sun W, Deng H, Li D, Wei Y, Lv B, et al. (2022). Cosformer: Rethinking softmax in attention. In: *International Conference on Learning Representations* (K Hofman, A Rush, Y Liu, C Finn, Y Choi, M Deisenroth, eds.).