

Grad-SAM: Explaining Transformers via Gradient Self-Attention Maps

Oren Barkan*
The Open University
Israel

Ori Katz
Microsoft & Technion
Israel

Edan Hauon*
IDC Herzeliya
Israel

Itzik Malkiel
Tel Aviv University
Israel

Avi Caciularu*
Bar-Ilan University
Israel

Omri Armstrong
Tel Aviv University
Israel

Noam Koenigstein
Microsoft & Tel Aviv University
Israel

ABSTRACT

Transformer-based language models significantly advanced the state-of-the-art in many linguistic tasks. As this revolution continues, the ability to explain model predictions has become a major area of interest for the NLP community. In this work, we present Gradient Self-Attention Maps (Grad-SAM) - a novel gradient-based method that analyzes self-attention units and identifies the input elements that explain the model's prediction the best. Extensive evaluations on various benchmarks show that Grad-SAM obtains significant improvements over state-of-the-art alternatives.

CCS CONCEPTS

• Computing methodologies → Machine Learning, Natural Language Processing.

KEYWORDS

Explainable & Interpretable AI, NLP, Deep Learning, BERT, Transformers, Self-Attention, Transparent Machine Learning

ACM Reference Format:

Oren Barkan, Edan Hauon, Avi Caciularu, Ori Katz, Itzik Malkiel, Omri Armstrong, and Noam Koenigstein. 2021. Grad-SAM: Explaining Transformers via Gradient Self-Attention Maps. In *Proceedings of the 30th ACM Int'l Conf. on Information and Knowledge Management (CIKM '21)*, November 1–5, 2021, Virtual Event, Australia. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3459637.3482126>

*Authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CIKM '21, November 1–5, 2021, Virtual Event, Australia.

© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8446-9/21/11...\$15.00
<https://doi.org/10.1145/3459637.3482126>

1 INTRODUCTION

Deep contextualized language models have significantly advanced the state-of-the-art in various linguistic tasks such as question answering [54], coreference resolution [28], and other NLP benchmarks [50, 51]. These models provide an efficient way to learn representations in a fully self-supervised manner from text corpora, solely using co-occurrence statistics. Empirically, deep contextualized language models that rely on the Transformer architecture [47], were shown to achieve state-of-the-art results, when are finetuned on supervised tasks [6, 13, 19, 38, 40].

Unlike traditional feature-based machine learning models that assign and optimize weights to interpretable explicit features, Transformer based architectures such as BERT [21] rely on a stack of multi-head self-attention layers, composed of hundreds of millions of parameters. These models are much more complex and computational heavier than models that learn non-contextualized representations [2, 4, 7, 9–11, 14, 33, 37].

At inference, Transformers-based models compute pairwise interactions of the resulting vector representations, making it particularly challenging to explain which part of the input contributed to the final prediction. Recently, significant efforts were put towards interpreting these models, mostly by applying *white-box* analysis [23, 49]. In this case, the goal is to probe the models' performance through lower-level components in the neural model.

Nonetheless, a central line of works attempted to study the types of linguistic knowledge encoded in such deep language models. Recent studies discovered that the BERT model [22] was shown to rely on surface structures (word, order, specific sequences, or co-occurrences) during pre-training [18, 41]. However, how and where exactly this information is stored, as well as retrieved during inference time, is still an open question that yet to be explored.

Gradient-based methods [45] yield decent ad-hoc explanations for predictions by highlighting which parts of the input correspond with the model's predictions. Moreover, recent works [1, 43, 52] showed that the input's sequence gradients have a high correlation with importance scores provided by human annotators, providing better interpretations than the scores produced by the raw token attentions. They observed that gradient-based ranking of attention scores better explains the model prediction than their magnitudes.

In this paper, we propose Gradient Self-Attention Maps (Grad-SAM) - a novel gradient-based interpretation method that probes BERT’s predictions. We demonstrate the effectiveness of Grad-SAM as a ranking machinery that identifies the input elements that contribute to the model prediction the most. We present both quantitative and qualitative results, indicating that Grad-SAM significantly outperforms state-of-the-art alternatives.

2 RELATED WORK

Recent methods for explaining predictions made by deep learning models considered explanations computed through convolutional layers [3, 12, 35, 36, 42], and attention based architectures [5, 8, 13, 18, 23–25, 32, 39, 46, 46, 49, 52]. While the authors in [26] argue that attention scores sometimes does not interpret model predictions faithfully, other works show that attention scores do offer plausible and meaningful interpretations that are often sufficient and correct [39, 48, 52].

Recently, a framework attempting to rationalize predictions named FRESH [27] was proposed. FRESH equips BERT predictions with faithfulness by construction – their goal was to focus on extracting rationales by introducing an additional extractor model trained to predict snippets. Then, a classifier is trained over the snippets and is expected to output faithful explanations.

In parallel, several methods were suggested to produce model interpretations via gradients [49]. The vast majority of these works utilized the gradients of the prediction w.r.t. the input for computing the importance of each token in the input sequence. In some scenarios, gradient-based approaches were shown to provide more faithful explanations than attention-based methods [15]. This family of gradient-based explainability methods have been applied [16, 17, 30], yet in a task-specific manner, to different downstream tasks.

Unlike the aforementioned works, our proposed Grad-SAM method integrates the information from attention scores together with their gradients in a finetuned BERT model. Furthermore, Grad-SAM is generic (not task-specific) in the sense it relies on the analysis of a given finetuned model only, and does not require the training an additional extractor network (in contrast to FRESH). Yet, our evaluation shows that Grad-SAM provides more faithful rationales than the ones produced by FRESH, across multiple linguistic tasks. Finally, it is worth noting that Grad-SAM, in its essence and purpose, differs from [42] by several aspects: First, it focuses on the NLP domain. Second, it operates on a completely different architecture (BERT). Third, it analyzes self-attention units from multiple layers in the model and not just the last one. Lastly, Grad-SAM treats negative gradient differently (Sec. 3).

3 GRADIENT SELF-ATTENTION MAPS

We begin by defining the problem setup and several notations. Then, we describe and explain Grad-SAM in detail. While our focus is on providing explanations for BERT, Grad-SAM is applicable for any architecture based on self-attention (SA) units [47].

Let $\mathcal{T} = \{t_i\}_{i=1}^{N_{\mathcal{T}}}$ be a vocabulary of supported tokens. Let Ω be a set containing all sentences of length N that can be composed from \mathcal{T} (shorter sentences are padded by a reserved token [PAD]), where each sentence starts and ends with the special tokens [CLS] and [SEP], respectively. BERT [22] is a parametric function $s : \Omega \rightarrow$

\mathbb{R}^n that receives a sequence (sentence) of N tokens $x = (x_i)_{i=1}^N$ ($x_i \in \mathcal{T}$) and outputs a n -dimensional vector of scores $s_x := s(x)$. In general, BERT is optimized via a two-phase process: In the *pre-training* phase, BERT is optimized w.r.t. to the Masked Language Model task together with the Next Sentence Prediction task. In the second phase, BERT is *finetuned* w.r.t. a specific downstream task e.g., multiclass/binary classification, or a regression task. Hence, n stands for the number of classes/output dimension and changes w.r.t. the specific downstream task at hand.

s is implemented as a cascade of L encoder layers. Given a sentence $x \in \Omega$, each token x_i in the sentence is mapped to a d -dimensional vector (embedding) to form a matrix $U_x^0 \in \mathbb{R}^{d \times N}$. In practice, this embedding is a summation of the token, positional, and segment embeddings. Then, U_x^0 is passed through a stack of L encoder layers. The l -th encoder layer ($1 \leq l \leq L$) receives the intermediate representations $U_x^{l-1} \in \mathbb{R}^{d \times N}$ (produced by the $(l-1)$ -th layer), and outputs the new representations U_x^l . Finally, $u_{[\text{CLS}]}^L$ (the first column in U_x^L , which corresponds to the [CLS] token) is used as input to a subsequent fully connected layer that outputs s_x .

Each encoder layer employs M SA heads that are applied in parallel to U_x^{l-1} , producing M different attention matrices

$$A_x^{lm} = \text{softmax} \left(\frac{(W_q^{lm} U_x^{l-1})^T W_k^{lm} U_x^{l-1}}{\sqrt{d_a}} \right), \quad (1)$$

where $W_q^{lm}, W_k^{lm} \in \mathbb{R}^{d_a \times d}$, are the query and key mappings, and $1 \leq m \leq M$. Each entry $[A_x^{lm}]_{ij}$ quantifies the amount of attention x_i receives by x_j , according to the attention head m in the layer l . Then, the encoder output U_x^l is obtained by a subsequent set of operations that involves the M attention matrices and *value* mappings as detailed in [47]. We refer to [22] for a detailed description of BERT.

Our goal is to explain the predictions made by BERT. To this end, we propose to utilize the attention matrices A_x^{lm} together with their gradients in order to produce a ranking over the tokens in x s.t. tokens that affect the model prediction the most, are ranked higher.

Given a sentence $x \in \Omega$ and a finetuned BERT model s , we compute the prediction $s_x \in \mathbb{R}^n$. In this work, our focus is on classification tasks. Specifically, for binary classification, $n = 1$, and we set s_x as the logit score. However, for multiclass classification, $n > 1$ (depending on the number of distinct classes) we focus on a specific entry in s_x which is associated with the ground truth class to be explained. For the sake of brevity, from here onwards, s_x represents the logit score in binary classification, or the logit score associated with the ground truth class s_x (in the case of multiclass classification) and disambiguation should be clear from the context.

Our goal is to quantify the importance of each token $x_i \in x$ w.r.t. s . In other words, we wish to identify tokens in x that contribute to s the most, hence explaining the prediction made by the model. To this end, we propose the following explanatory scheme: First, we pass x through BERT to compute s_x . Then, the *importance* of

the token x_i w.r.t. the prediction s_x is computed by:

$$r_{x_i} = \frac{1}{LMN} \sum_{l=1}^L \sum_{m=1}^M \sum_{j=1}^N [H_x^{lm}]_{ij}, \quad (2)$$

with

$$H_x^{lm} = A_x^{lm} \circ \text{ReLU}(G_x^{lm}), \quad (3)$$

where $G_x^{lm} := \frac{\partial s_x}{\partial A_x^{lm}}$ are the element-wise gradients of s_x w.r.t. to A_x^{lm} , and \circ stands for the Hadamard product. Eq. 2 scores the importance of each token $x_i \in x$ w.r.t. s_x , enabling ranking the tokens in x according to their importance. Higher values of r_{x_i} indicate higher importance of x_i , hence a better explanation of the prediction score s_x . In practice, for $x_i \in \{[\text{CLS}], [\text{SEP}], [\text{PAD}]\}$, we set $r_{x_i} = -\infty$, as these tokens cannot provide for good explanations.

The motivation behind Eqs. 2 and 3 is as follows: We are willing to identify tokens in x for which 1) High attention is received from other tokens in x (information from the *attention activations*), and 2) Further increase in the amount of the received attention will increase s_x the most (information from *gradient* of the attention activations). Eq. 3 ensures that these two conditions are met, since if $[G_x^{lm}]_{ij} \leq 0$, then $[H_x^{lm}]_{ij} = 0$, and if $[A_x^{lm}]_{ij}$ is small, then $[H_x^{lm}]_{ij}$ is close to zero (recall that $[A_x^{lm}]_{ij} \geq 0$, as it is the result of softmax). Finally, Eq. 2 aggregates the overall contribution of the attention scores and the positive gradients from all SA heads across all encoder layers, w.r.t. $x_i \in x$.

We wish to re-emphasize the following important point: Zeroing the negative gradients in Eq. 3 enables the preservation of the positive values of H_x^{lm} (associated with positive gradients), which otherwise may be cancelled out by a large accumulated negative value in the summation in Eq. 2. The activations in the i -th row within an attention matrix A_x^{lm} quantify the importance of the token x_i w.r.t. the other tokens in x . In addition, if $[G_x^{lm}]_{ij} > 0$, then an increase in the activation $[A_x^{lm}]_{ij}$ should lead to an increase in the model’s output score. Therefore, the importance of the token x_i (according to the attention head m in the encoder layer l) is determined by the summation over the i -th row in H_x^{lm} , and the contribution to this sum come from elements for which both the activation and its gradient are positive. Finally, the overall importance of x_i is accumulated from the M SA heads in L layers according to Eq. 2.

In regular BERT-base models, there are 144 SA heads ($M = 12, L = 12$) that act as filters. However, in practice, we observed that only a few attention entries are activated. Specifically, we found out that there is a large number of activations that are close to zero, but associated with negative gradients. The accumulated effect of this negative sum leads to a suppression (or even complete cancellation) of the small number of activations with positive gradients (which hold the actual information we are wish to preserve). Hence, we zero those negative gradients (using ReLU). The necessity of the negative gradient trimming, prior to the summation, along with the complementary contribution from the attention activations and their gradients, are validated in the ablation study presented in Tabs. 1 and 2.

4 EXPERIMENTAL SETUP AND RESULTS

4.1 Datasets and Downstream Tasks

In all of the experiments, we use a pre-trained BERT-base-uncased model, taken from Huggingface’s Transformers library [53], associated with a standard tokenizer. Then, we finetune BERT on five downstream tasks (in separate):

- The Stanford Sentiment Treebank (SST) [44]: A sentiment analysis task (binary classification).
- AgNews (AGN) [20]: A multiclass classification task, where news articles are categorized into *science*, *sports*, *business*, *world*.
- IMDB [31]: A sentiment analysis task (binary classification of movie reviews).
- MultiRC (MRC) [29]: A binary classification task. The same processing from [27] was followed to produce True / False labels w.r.t. a given snippet.

4.2 Evaluated Methods

We compare several methods for ranking the importance of tokens in a sentence x :

- (1) **Gradient**: This is the ‘Gradient’ method from [27].
- (2) **[CLS] Att**: This is the ‘[CLS] Attn’ method from [27].
- (3) **Att**: Setting $H_x^{lm} = A_x^{lm}$ and using Eq. 2.
- (4) **Att-Grad**: Setting $H_x^{lm} = G_x^{lm}$ and using Eq. 2.
- (5) **Att-Grad-R**: Setting $H_x^{lm} = \text{ReLU}(G_x^{lm})$ and using Eq. 2.
- (6) **Att \times Att-Grad**: Setting $H_x^{lm} = A_x^{lm} \circ G_x^{lm}$ and using Eq. 2.
- (7) **Grad-SAM**: Using Eq. 2 (our proposed method).

Note that methods 3-6 are ablated versions of Grad-SAM.

4.3 Quantitative Evaluations

Our first evaluation follows the protocol from [27]: For each sentence x in the test set, we used each method (Sec. 4.2) to produce a different ranking over x ’s tokens. Then, we preserved the top $k\%$ ranked tokens and masked the rest. We used the same values from [27]: $k = 20\%$ for SST, AGN, and MRC, and $k = 30\%$ for IMDB. Finally, we compute the mean Macro F1 score for each combination of method and task. We further include the original *Full text* results obtained on x without masking.

Table 1 depicts the results for each combination of explanation method and task. First, we see that Grad-SAM significantly outperforms both the Gradient and the [CLS] Att methods from [27]. It is worth noting that the methods from [27] require a simultaneous training of an auxiliary model that *learns to mask*, during the BERT’s finetuning phase, while our Grad-SAM method eliminates this need. The ablation study reveals that the ReLU operation over the gradient-attention is crucial (Att-Grad-R > Att-Grad). This indicates that by trimming the negative gradients, we avoid the unwanted suppression of positive gradients (if exist) across the summation in Eq. 2. Moreover, Grad-SAM, which combines the attention scores together with ReLUed gradients, performs the best across all tasks.

Our second evaluation is based on the Area Over the Perturbation Curve (AOPC) [34] metric that is designed to assess the faithfulness of explanations produced by Grad-SAM and the other methods. AOPC calculates the average change of accuracy over test data by masking the top $k\%$ tokens in the sentence x (the tokens are ranked by the explanation method). Hence, the larger the

Method	SST	AGN	IMDB	MRC
<i>Full text</i>	.904	.942	.957	.682
Gradient	.682	.863	.933	.654
[CLS] Att	.812	.911	.941	.639
Att	.801	.855	.837	.632
Att-Grad	.706	.792	.715	.634
Att-Grad-R	<u>.819</u>	<u>.911</u>	<u>.946</u>	<u>.657</u>
Att×Att-Grad	.810	.778	.743	.636
Grad-SAM	.823	.921	.949	.662

Table 1: Model predictive performances across datasets. We report the mean-macro F1 scores on the test sets. The top row (*Full text*) corresponds to passing the sentence, without masking (upper-bound on performance).

Method	SST	AGN	IMDB	MRC
<i>Full text</i>	.904	.942	.957	.682
Gradient	.16	.101	.05	.09
[CLS] Att	.165	.177	.055	.082
Att	.113	.118	.047	.093
Att-Grad	.072	.109	.031	.113
Att-Grad-R	<u>.179</u>	.132	<u>.059</u>	<u>.12</u>
Att×Att-Grad	.152	.12	.052	.103
Grad-SAM	.195	.14	.065	.122

Table 2: AOPC evaluation. Note that the *Full Text* row is presented for reference, reporting the mean-macro F1 scores on test sets without any word filtration. The other rows report the AOPC for each combination of method and dataset.

value of AOPC, the better the explanations of the models. Table 2 depicts the results with $k = 20\%$ (the top 20% of words ranked by each method). The *Full Text* row is presented for reference, reporting the mean-macro F1 scores on the original sentences from the test sets without any masking. The other rows report the AOPC for each combination of method and dataset. Again, we compare Grad-SAM to the same baselines from Sec. 4.2 (and perform an ablation study).

The results in Tab. 2 indicate that: 1) Grad-SAM outperforms the other methods, hence is capable of identifying the words in the input sequence that contribute the most to the (correct) model prediction. For example, for a BERT model that was finetuned on the SST dataset, we observe that by masking the top 20% words proposed by [CLS] Att, the accuracy drops to 16.5 points, whereas in the case of Grad-SAM, the accuracy drops to 19.5 points. 2) BERT is sensitive to the context; omitting important words hinder the semantics in the sentence and significantly affects the model’s predictions. Overall, this AOPC evaluation provides another evidence that Grad-SAM is a state-of-the-art machinery that generates faithful explanations.

4.4 Qualitative Results

In this section, we provide qualitative examples produced by our Grad-SAM method and the [CLS] Att methods from [27]. We follow the same procedure described in Sec. 4.3: Namely, we applied both Grad-SAM and [CLS] Att to rank the tokens according to their importance and considered the top $k = 20\%$ tokens in the list produced by each method. Finally, we masked all the tokens in the sentences besides the top $k = 20\%$ selected tokens, fed the masked sentence to BERT, and performed the prediction.

From the AGNews test set, we randomly picked 4 examples associated with several ground truth labels. From the SST test set, we randomly picked 3 positive and 3 negative sentences (according to the ground truth labels). For all examples, the original prediction made by BERT (without masking) is correct (matches the ground truth label).

Table 3 presents the results for both datasets. For AGNews, we observe that Grad-SAM based masking (fifth column) does not lead to a change in the model’s predictions, while [CLS] Att based masking (last column) does change the model’s prediction, and to an incorrect one (recall that the original prediction made by the model matches the ground truth label). Finally, Grad-SAM identifies tokens that better explain the prediction made by BERT.

5 CONCLUSION

This work joins a growing effort to better interpreting deep contextualized language models. To this end, we present Grad-SAM, a novel gradient-based method for explaining predictions made by a finetuned BERT model. Extensive evaluations show that Grad-SAM outperforms other state-of-the-art methods across various datasets, tasks, and evaluation metrics.

REFERENCES

- [1] Samira Abnar and Willem Zuidema. 2020. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4190–4197. <https://doi.org/10.18653/v1/2020.acl-main.385>
- [2] Oren Barkan. 2017. Bayesian neural word embedding. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- [3] Oren Barkan, Omri Armstrong, Amir Hertz, Avi Caciularu, Ori Katz, Itzik Malkiel, and Noam Koenigstein. 2021. GAM: Explainable Visual Similarity and Classification via Gradient Activation Maps. In *Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM)*.
- [4] Oren Barkan, Avi Caciularu, and Ido Dagan. 2020. Within-Between Lexical Relation Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 3521–3527. <https://doi.org/10.18653/v1/2020.emnlp-main.284>
- [5] Oren Barkan, Avi Caciularu, Ori Katz, and Noam Koenigstein. 2020. Attentive item2vec: Neural attentive user representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- [6] Oren Barkan, Avi Caciularu, Idan Rejwan, Ori Katz, Jonathan Weill, Itzik Malkiel, and Noam Koenigstein. 2020. Cold item recommendations via hierarchical item2vec. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 912–917.
- [7] Oren Barkan, Avi Caciularu, Idan Rejwan, Ori Katz, Jonathan Weill, Itzik Malkiel, and Noam Koenigstein. 2021. Representation Learning via Variational Bayesian Networks. In *Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM)*.
- [8] Oren Barkan, Yonatan Fuchs, Avi Caciularu, and Noam Koenigstein. 2020. Explainable recommendations via attentive multi-persona collaborative filtering. In *ACM Conference on Recommender Systems (RecSys)*.
- [9] Oren Barkan, Roy Hirsch, Ori Katz, Avi Caciularu, and Noam Koenigstein. 2021. Anchor-based Collaborative Filtering. In *Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM)*.

Document / Sentence	Tokens highlighted by Grad-SAM	Tokens highlighted by [CLS] Att	Original Prediction	Prediction (Grad-SAM Masking)	Prediction ([CLS] Att Masking)
AGNews					
The federal agency that protects private sector pension plans announced yesterday that the maximum annual benefit for plans taken over in 2005 will be \$45,614 for workers who wait until age 65 to retire.	federal, pension, yesterday, plans, 2005, \$45,614, workers	annual, workers, wait, until, age, 65, retire	Business	Business	World
South Korea's key allies play down a shock admission its scientists experimented to enrich uranium.	Korea, allies, uranium	s, scientists, enrich, uranium	World	World	Sci/Tech
OTTAWA – A local firm that says it can help shrink backup times at large data centers is growing its business thanks to an alliance with Sun Microsystems Inc.	OTTAWA, –, local, firm, centers, business	it, backup, data, centers, business, Microsystems	Business	Business	Sci/Tech
Tokyo share prices fell steeply Friday, led by technology stocks after a disappointing report from US chip giant Intel. The US dollar was up against the Japanese yen.	Tokyo, share, prices, Friday, stocks, US	prices, steeply, Friday, chip, Intel, yen	Business	Business	Sci/Tech
SST					
A great idea becomes a not great movie	not, great	becomes, great	Negative	Negative	Positive
Flashy pretentious and as impenetrable as morvern's thick working class scottish accent	flashy, pretentious, impenetrable	flashy, and, impenetrable	Negative	Negative	Positive
A strong first quarter slightly less so second quarter and average second half	strong, less, average	strong, and, average	Negative	Negative	Positive
An impressive if flawed effort that indicates real talent	impressive, flawed	an, flawed	Positive	Positive	Negative
This road movie gives you emotional whiplash and you'll be glad you went along for the ride	gives, emotional, whiplash, glad	this, emotional, whiplash, and	Positive	Positive	Negative
It never fails to engage us	never, fails	it, never	Positive	Positive	Negative

Table 3: Top $k = 20\%$ ranked tokens for the AGNews dataset followed by SST dataset. The tokens are ordered according to their scores in a descending order. Original Predicted stands for the prediction made by BERT on the original input (without masking). The last two columns present the prediction made by BERT after applying the masking produced by Grad-SAM and [CLS] Att [27].

- [10] Oren Barkan, Roy Hirsch, Ori Katz, Avi Caciularu, Yoni Weill, and Noam Koenigstein. 2021. Cold Start Revisited: A Deep Hybrid Recommender with Cold-Warm Item Harmonization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3260–3264.
- [11] Oren Barkan, Roy Hirsch, Ori Katz, Jonathan Weill, and Noam Koenigstein. 2021. Cold Item Integration in Deep Hybrid Recommenders via Tunable Stochastic Gates. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*.
- [12] Oren Barkan, Ori Katz, and Noam Koenigstein. 2020. Neural Attentive Multi-view Machines. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- [13] Oren Barkan, Noam Razin, Itzik Malkiel, Ori Katz, Avi Caciularu, and Noam Koenigstein. 2020. Scalable attentive sentence pair modeling via distilled sentence embedding. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*.

- [14] Oren Barkan, Idan Rejwan, Avi Caciularu, and Noam Koenigstein. 2020. Bayesian Hierarchical Words Representation Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3871–3877. <https://doi.org/10.18653/v1/2020.acl-main.356>
- [15] Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Online, 149–155. <https://doi.org/10.18653/v1/2020.blackboxnlp-1.14>
- [16] Aaron Chan, Soumya Sanyal, Boyuan Long, Jiahu Xu, Tanishq Gupta, and Xi-ang Ren. 2021. SalKG: Learning From Knowledge Graph Explanations for Commonsense Reasoning. *arXiv preprint arXiv:2104.08793* (2021).
- [17] George Chrysostomou and Nikolaos Aletras. 2021. Variable Instance-Level Explainability for Text Classification. *arXiv preprint arXiv:2104.08219* (2021).
- [18] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. In *Black-BoxNLP@ACL*.
- [19] Andrew M Dai and Quoc V Le. 2015. Semi-supervised Sequence Learning. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc., 3079–3087.
- [20] Gianna M Del Corso, Antonio Gulli, and Francesco Romani. 2005. Ranking a stream of news. In *Proceedings of the international conference on World Wide Web (WWW)*.
- [21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805* [cs.CL]
- [23] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 31–36.
- [24] Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, and Noam Koenigstein. 2021. Self-Supervised Document Similarity Ranking via Contextualized Language Models and Hierarchical Inference. *arXiv preprint arXiv:2106.01186* (2021).
- [25] Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do Attention Heads in BERT Track Syntactic Dependencies? *arXiv preprint arXiv:1911.12246* (2019).
- [26] Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).
- [27] Sarthak Jain, Sarah Wiegrefe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to Faithfully Rationalize by Construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4459–4473. <https://doi.org/10.18653/v1/2020.acl-main.409>
- [28] Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for Coreference Resolution: Baselines and Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5803–5808. <https://doi.org/10.18653/v1/D19-1588>
- [29] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 252–262. <https://doi.org/10.18653/v1/N18-1023>
- [30] Luoqi Li, Xiang Chen, Ningyu Zhang, Shumin Deng, Xin Xie, Chuanqi Tan, Moshu Chen, Fei Huang, and Huajun Chen. 2021. Normal vs. Adversarial: Saliency-based Analysis of Adversarial Samples for Relation Extraction. *arXiv preprint arXiv:2104.00312* (2021).
- [31] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*.
- [32] Itzik Malkiel, Oren Barkan, Avi Caciularu, Noam Razin, Ori Katz, and Noam Koenigstein. 2020. Optimizing BERT for Unlabeled Text-Based Items Similarity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1704–1714.
- [33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In

- Advances in neural information processing systems*. 3111–3119.
- [34] Dong Nguyen. 2018. Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1069–1078. <https://doi.org/10.18653/v1/N18-1097>
 - [35] Badri Patro, Vinay Namboodiri, et al. 2020. Explanation vs attention: A two-player game to obtain attention for VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11848–11855.
 - [36] Badri N Patro, Mayank Lunayach, Shivansh Patel, and Vinay P Namboodiri. 2019. U-cam: Visual explanation using uncertainty based class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7444–7453.
 - [37] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
 - [38] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. <https://doi.org/10.18653/v1/N18-1202>
 - [39] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2019. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913* (2019).
 - [40] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. [n.d.]. Improving Language Understanding by Generative Pre-Training. ([n. d.]).
 - [41] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327* (2020).
 - [42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
 - [43] Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2931–2951. <https://doi.org/10.18653/v1/P19-1282>
 - [44] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. <https://www.aclweb.org/anthology/D13-1170>
 - [45] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.
 - [46] Wen Tai, HT Kung, Xin Luna Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1433–1439.
 - [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
 - [48] Jesse Vig. 2019. Visualizing attention in transformer-based language representation models. *arXiv preprint arXiv:1904.02679* (2019).
 - [49] Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. In *Empirical Methods in Natural Language Processing*.
 - [50] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGlue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*. 3266–3280.
 - [51] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 353–355. <https://doi.org/10.18653/v1/W18-5446>
 - [52] Sarah Wiegrefe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 11–20. <https://doi.org/10.18653/v1/D19-1002>
 - [53] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
 - [54] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.