Towards Hierarchical Importance Attribution: Explaining Compositional Semantics for Neural Sequence Models

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks have achieved impressive performance in handling complicated semantics in natural language, while mostly treated as black boxes. To explain how the model handles compositional semantics of words and phrases, we study the hierarchical explanation problem. We highlight the key challenge is to compute non-additive and context-independent importance for individual words and phrases. We show some prior efforts on hierarchical explanations, e.g. contextual decomposition, do not satisfy the desired properties mathematically, leading to inconsistent explanation quality in different models. In this paper, we propose a formal way to quantify the importance of each word or phrase to generate hierarchical explanations. We modify contextual decomposition algorithms according to our formulation, and propose a model-agnostic explanation algorithm with competitive performance. Human evaluation and automatic metrics evaluation on both LSTM models and fine-tuned BERT Transformer models on multiple datasets show that our algorithms robustly outperform prior works on hierarchical explanations. We show our algorithms help explain compositionality of semantics, extract classification rules, and improve human trust of models.

1 INTRODUCTION

Recent advances in deep neural networks have led to impressive results on a range of natural language processing (NLP) tasks, by learning latent, compositional vector representations of text data (Peters et al., 2018; Devlin et al., 2018; Liu et al., 2019b). However, interpretability of the predictions given by these complex, "black box" models has always been a limiting factor for use cases that require explanations of the features involved in modeling (e.g., words and phrases) (Guidotti et al., 2018; Ribeiro et al., 2016). Prior efforts on enhancing model interpretability have focused on either constructing models with intrinsically interpretable structures (Bahdanau et al., 2015; Liu et al., 2019a), or developing post-hoc explanation algorithms which can explain model predictions without elucidating the mechanisms by which model works (Mohseni et al., 2018; Guidotti et al., 2018). Among these work, post-hoc explanation has come to the fore as they can operate over a variety of trained models while not affecting predictive performance of models.

Towards post-hoc explanation, a major line of work, additive feature attribution methods (Lundberg & Lee, 2017; Ribeiro et al., 2016; Binder et al., 2016; Shrikumar et al., 2017), explain a model prediction by assigning importance scores to individual input variables. However, these methods may not work for explaining compositional semantics in natural language (e.g., phrases or clauses), as the importance of a phrase often is non-linear combination of the importance of the words in the phrase. Contextual decomposition (CD) (Murdoch et al., 2018) and its hierarchical extension (Singh et al., 2019) go beyond the additive assumption and compute the contribution solely made by a word/phrase to the model prediction (i.e., individual contribution), by decomposing the output variables of the neural network at each layer. Using the individual contribution scores so derived, these algorithms generate hierarchical explanation on how the model captures compositional semantics (e.g., stress or negation) in making predictions (see Figure 1 for example).

1 Code and data have been uploaded and will be published upon acceptance of the paper.
Under review as a conference paper at ICLR 2020

Figure 1: Different score attribution algorithms. (a) Input occlusion assigns a negative score for the word “interesting”, as the sentiment of the phrase becomes less negative after removing “interesting” from the original sentence. (b) Additive attributions assign importance scores for words “not” and “interesting” by linearly distributing contribution score of “not interesting”, exemplified with Shapley Values (Shapley, 1997). Intuitively, only (c) Hierarchical explanations highlight the negative compositional effect between the words “not” and “interesting”.

However, despite contextual decomposition methods have achieved good results in practice, what reveals extra importance that emerge from combining two phrases has not been well studied. As a result, prior lines of work on contextual decomposition have focused on exploring model-specific decompositions based on their performance on visualizations. We identify the extra importance from combining two phrases can be quantified by studying how the importance of the combined phrase differs from the sum of the importance of the two component phrases on its own. Similar strategies have been studied in game theory for quantifying the surplus from combining two groups of players (Driessen, 2013). Following the definition above, the key challenge is to formulate the importance of a phrase on its own, i.e., context independent importance of a phrase. However, while contextual decomposition algorithms try to decompose the individual contributions from given phrases for explanation, we show neither of them satisfy this context independence property mathematically.

To this end, we propose a formal way to quantify the importance of each individual word/phrase, and develop effective algorithms for generating hierarchical explanations based on the new formulation. To mathematically formalize and efficiently approximate context independent importance, we formulate \( N \)-context independent importance of a phrase, defined as the difference of model output after masking out the phrase, marginalized over all possible \( N \) words surrounding the phrase in the sentence. We propose two explanation algorithms according to our formulation, namely the Sampling and Contextual Decomposition algorithm (SCD), which overcomes the weakness of contextual decomposition algorithms, and the Sampling and OCclusion algorithm (SOC), which is simple, model-agnostic, and performs competitively against prior lines of algorithms. We experiment with both LSTM and fine-tuned Transformer models to evaluate the proposed methods. Quantitative studies involving automatic metrics and human evaluation on sentiment analysis and relation extraction tasks show that our algorithms consistently outperform competitors in the quality of explanations. Our algorithms manage to provide hierarchical visualization of compositional semantics captured by models, extract classification rules from models, and help users to trust neural networks predictions.

In summary, our work makes the following contributions: (1) we identify the key challenges in generating post-hoc hierarchical explanations and propose a mathematically sound way to quantify context independent importance of words and phrases for generating hierarchical explanations; (2) we extend previous post-hoc explanation algorithm based on the new formulation of \( N \)-context independent importance and develop two effective hierarchical explanation algorithms; and (3) both experiments using automatic evaluation metrics and human evaluation demonstrate that the proposed explanation algorithms consistently outperform the compared methods (with both LSTM and Transformer as base models) over several datasets.

2 Preliminaries

2.1 Post-hoc Explanations of Neural Sequence Models

We consider a sequence of low-dimensional word embeddings \( x_{1:T} := (x_1, x_2, ..., x_T) \), or denoted as \( x \) for brevity, as the input to a neural sequence model, such as standard RNNs, LSTM (Hochreiter & Schmidhuber, 1997) and Transformers (Vaswani et al., 2017). These neural models extract latent,
compositional representations $h_{1:T}$ (i.e., hidden states) from the input sequence $x$, and feed these hidden state vectors to a prediction layer to generate output in the label space (e.g., sentiment polarity of a sentence). For LSTM, we use the last hidden state $h_T$ to give unnormalized prediction scores $s(x) \in \mathbb{R}^{d_c}$ over $d_c$ label classes as follows.

$$s(x) = W_f h_T,$$

where $W_f \in \mathbb{R}^{d_c \times d_h}$ is a trainable weight matrix. For Transformers, the representation corresponding to the “[CLS]” token at the final layer is fed to the prediction layer to generate scores $s(x)$.

Towards post-hoc explanation of $s(x)$, a notable line of work, **additive feature attribution methods** (Ribeiro et al., 2016; Shrikumar et al., 2017; Sundararajan et al., 2017), measure word-level importance to the model prediction $s(x)$ by attributing a importance score $\phi(x_i, x)$ to each word in the input sequence $x_i \in x$. Such additive attribution methods are related to Shapley Values (Shapley, 1953) and thus can be proven to enjoy good properties, including that it has unique solution of a “fair” attribution (Lundberg & Lee, 2017). However, the additive assumption hinders these methods from explaining the complex interactions between words and compositional semantics in a sentence (e.g., modeling negation, transition, and emphasis in sentiment classification), as shown in Figure 1.

### 2.2 Hierarchical Explanations via Contextual Decomposition

To capture non-linear compositional semantics, the line of work on contextual decomposition (CD) (Murdock et al., 2018) designs non-additive measures of importance from individual words/phrases to the model predictions, and further extend to agglomerative contextual decomposition (ACD) algorithm (Singh et al., 2019) for generating hierarchical explanations.

Given a phrase $p = x_{i:j}$ in the input sequence $x$, contextual decomposition (CD) attributes a score $\phi(p, x)$ as the contribution **solely** from $p$ to the model’s prediction $s(x)$. Note that $\phi(p, x)$ does not equal to the sum of the scores of each word in the phrase, i.e., $\phi(p, x) \neq \sum_{x_i \in p} \phi(x_i, x)$. Starting from the input layer, CD iteratively decomposes each hidden state $h$ of the model into the contribution solely made by $p$, denoted as $\beta$, and the contributions involving the words outside the phrase $p$, denoted as $\gamma$, with the relation $h = \beta + \gamma$ holds. Note that the algorithm also keeps the contribution from the bias term, denoted as $\zeta$.

For a linear layer $h = W_i x_i + b$, with input $x_i$, the contribution solely from $p$ to $h$ is defined as $\beta = W_i x_i$ when $x_i$ is part of the phrase (i.e., $x_i \in p$), and the contribution involving other words in the sentences (denoted as $x \setminus p$) is defined as $\gamma = 0$. The contribution of the bias term $\zeta$ is thus $b$. When $x_i$ lies outside of the phrase in the sentence (i.e., $x_i \notin p$), $\gamma$ is quantified as $W_i x_i$ and $\beta$ is 0. In the cases when CD encounters element-wise multiplication operations $h = h_a * h_b$ (e.g., in LSTMs), it eliminates the multiplicative interaction terms which involve the information outside of the phrase $p$. Specifically, suppose that $h_a$ and $h_b$ have been decomposed as $h_a = \beta^a + \gamma^a + \zeta^a$ and $h_b = \beta^b + \gamma^b + \zeta^b$, CD computes the $\beta$ term for $h_a * h_b$ as $\beta = \beta^a * \beta^b + \beta^a * \beta^b + \zeta^a * \zeta^b$.

When dealing with non-linear activation $h' = \sigma(h)$, CD computes the contribution solely from the phrase $p$ as the average activation differences caused by $\beta$ supposing $\gamma$ is present or absent,

$$\beta' = \frac{1}{2} [\sigma(\beta + \gamma + \zeta) - \sigma(\gamma + \zeta)] + \frac{1}{2} [\sigma(\beta + \zeta) - \sigma(\zeta)].$$

Following the three strategies introduced above, CD decomposes all the intermediate outputs starting from the input layer, until reaching the final output of the model $h_T = \beta + \gamma$. The logit score $W_f \beta$ is treated as the contribution of the given phrase $p$ to the final prediction $s(x)$.

As a follow-up study, Singh et al. (2019) extends CD algorithm to other families of neural network architectures, and proposes agglomerative contextual decomposition algorithm (ACD). The decomposition of activation functions is modified as $\beta' = \sigma(\beta')$. For the linear layer $h' = W h + b$ with its decomposition $h = \beta + \gamma$, the bias term $b$ is decomposed proportionally and merged into the $\beta'$ term of $h'$, based on $\beta' = W \beta + |W| \beta / (|W| \beta + |W| \gamma) \cdot a$.

### 3 Methodology

In this section, we start by identifying desired properties of phrase-level importance score attribution for hierarchical explanations. We propose a measure of context-independent importance of a phrase and introduce two explanation algorithms instantiated from the proposed formulation.
Non-additivity: Importance of a phrase $\phi(p, x)$ should be quantified by a non-linear function over the importance scores of all the component words $x_i \in p$, i.e., $\phi(p, x) \neq \sum_{x_i \in p} \phi(x_i, x)$, in contrast to the family of additive feature attribution methods.

Context Independence: For deep neural networks, when two phrases combine, their importance to predicting a class may greatly change. The surplus by combining two phrases can be quantified by the difference between the importance of the combined phrase and the sum of the importance of two phrases on its own. It follows how the surplus of combining two groups of players can be quantified in game theory (Driessen, 2013; Fujimoto et al., 2006). According to the definition, the importance of two component phrases should be evaluated independently of each other. Formally, if we are interested in how combining two phrases $p_1$ and $p_2$ contribute to a specific prediction for an input $x$, we expect for input sentences $\hat{x}$ where only $p_2$ is replaced to another phrase, the importance attribution for $p_1$ remains the same, i.e., $\phi(p_1, \hat{x}) = \phi(p_1, x)$. In the hierarchical explanation setting, we are interested in how combining a phrase and any other contextual words or phrases in the input $x$ changes the prediction for the input $x$. Therefore, we expect $\phi(p, x) = \phi(p, \hat{x})$ given the phrase $p$ in two different contexts $x$ and $\hat{x}$.

Limitations of CD and ACD: Unfortunately, while CD and ACD try to construct decomposition so that $\beta$ terms represent the contributions solely from a given phrase, the assigned importance scores by these algorithms do not satisfy the context independence property mathematically. For CD, we see the computation of $\beta$ involves the $\gamma$ term of a specific input sentence in Eq. 2 (see Figure 2(a) for visualization). Similarly, for ACD, the decomposition of the bias term involves the $\gamma$ terms of a specific input sentence. As a result, the $\beta$ terms computed by both algorithms depend on the context of the phrase $p$. Regarding the decomposition of activation functions, the decomposition $\beta' = \sigma(\beta)$ in ACD seems plausible, which does not involve $\gamma$ for computing $\beta$ terms. However, in case every activation is decomposed in this way and suppose the bias terms are merged into $\beta$, the algorithm is equivalent to feeding only the phrase $p$ into the classifier with all other input masked as zero. Empirical results show unreliable explanation quality of both algorithms in some models.

Given the limitation of prior works, we start by formulating a importance measure of phrases that satisfies both non-additivity and context independence property.

3.2 Context-Independent Importance

Given a phrase $p := x_{i:j}$ appearing in a specific input $x_{1:T}$, we first relax our setting and define the importance of a phrase independent of all the possible $N$-word contexts adjacent to it. The
The process of evaluating context independent importance is formally written as

$$\phi(p, x) = E_{x_δ}[s(x_{-δ}; x_δ) - s(x_{-δ}\setminus p; x_δ)],$$

(3)

where \(x_{-δ}\) denotes the resulting sequence after masking out an \(N\)-word context surrounding the phrase \(p\) from the input \(x\). Here, \(x_δ\) is a \(N\)-word sequence sampled from a distribution \(p(x_δ|x_{-δ})\), which is conditioned on the phrase \(p\) as well as other words in the sentence \(x\). Details on the sampling process will be elaborated later in the section. Accordingly, we use \(s(x_{-δ}; x_δ)\) to denote the model prediction score after replacing the masked-out context \(x_{-δ}\) with a sampled \(N\)-word sequence \(x_δ\). We use \(x\setminus p\) to denote the operation of masking out the phrase \(p\) from the input sentence \(x\). The specific implementation of this masking out operation varies across different explanation algorithms and is instantiated from their formulation.

Following the notion of \(N\)-context independent importance, we define context-independent importance of a phrase \(p\) by increasing the size of the context \(N\) to sufficiently large (e.g., length of the sentence). The context independent importance can be equivalently written as follows.

$$\phi^\gamma(p) = E_x[s(x) - s(x\setminus p)|p \subseteq x].$$

(4)

Computing Eqs. 3 and 4 are intractable as it requires integrating over a large number of variants of \(x_δ\) as replacements (i.e., number of variants for \(x_δ\) is exponential to the size of \(N\)). While it is possible to approximate the expectations in Eqs. 3 and 4 by sampling from the training text corpus, we find it common that a phrase occurs sparsely in the corpus. Therefore, we approximate the expectation by sampling from a language model pre-trained using the training corpus. The language model helps model a smoothed distribution of \(p(x_δ|x_{-δ})\). In practice, all our explanation algorithms implement \(N\)-context independent importance following Eq. 3 where the size of the neighborhood \(N\) is a parameter to be specified.

### 3.3 Sampling and Contextual Decomposition Algorithm

In contextual decomposition algorithm, the desirable context independence property is compromised when computing decomposition of activation functions, as discussed in Section 2.3. Following the new formulation on context-independent importance introduced in Section 3.2, we present a simple modification of the contextual decomposition algorithm, and develop a new sampling and contextual decomposition (SCD) algorithm for effective generation of hierarchical explanations.

SCD only modifies the way to decompose activation functions in CD. Specifically, given the output \(h = s(l)(x)\) at an intermediate layer \(l\) with the decomposition \(h = \beta + \gamma\), we decompose the activation value \(\sigma(h)\) into \(\beta' + \gamma'\), with the following definition:

$$\beta' = E_{\gamma}[\sigma(\beta + \gamma) - \sigma(\gamma)] = E_h[\sigma(h) - \sigma(h - \beta)],$$

(5)

i.e., \(\beta'\) is defined as the expected difference between the activation values when the \(\beta\) term is present or absent. \(h\) is computed for different input sequences \(x\) with the contexts of the phrase \(p\) sampled from the distribution \(p(x_δ|x_{-δ})\). Eq. 5 is a layer wise application of Eq. 4, where the masking operation is implemented with calculating \(\sigma(h - \beta)\). Figure 2(b) provides a visualization for the decomposition.

To perform sampling, we first pretrain a LSTM language model from two directions on the training data. For sampling, we mask the words that are not conditioned in \(p(x_δ|x_{-δ})\). Some other sampling options include performing Gibbs sampling from a masked language model (Wang et al., 2019). The algorithm then obtain a set of samples \(S\) by sampling with the trained language model. For each sample in \(S\), the algorithm records the input of the \(i\)-th non-linear activation function to obtain a sample set \(\phi_h^{(i)}\). During the explanation, the decomposition of the \(i\)-th non-linear activation function
is calculated as,
\[ \beta' = \frac{1}{|S^i_h|} \sum_{h \in S^i_h} [\sigma(h) - \sigma(h - \beta)]. \]

(6)

Some neural models such as Transformers involve operations that normalize over different dimensions of a vectors, e.g. softmax functions and layer normalization operations. We observe improved performance by not decomposing the normalizer of these terms when the phrase \( p \) is shorter than a threshold, assuming that the impact of \( p \) to the normalizer can be ignored. Besides, for element-wise multiplication in LSTM models, we treat them in the same way as other non-linear operations and decompose them as Eq. 5, where the decomposition of \( h_1 h_2 \) is written as
\[ \beta' = \mathbb{E}_{\gamma_1, \gamma_2} [(\beta_1 + \gamma_1)(\beta_2 + \gamma_2) - \gamma_1 \gamma_2] \]

3.4 SAMPLING AND OCCLUSION ALGORITHM

We show it is possible to fit input occlusion (Li et al., 2016) algorithms into our formulation. Input occlusion algorithms calculate the importance of \( p \) specific to an input example \( x \) by observing the prediction difference after replacing the phrase \( p \) with padding tokens, noted as \( 0_p \).

\[ \phi(p, x) = s(x) - s(x_\delta; x_\delta; 0_p) \]

(7)

It is obvious that importance score by the input occlusion algorithm is dependent on all the context words of \( p \) in \( x \). To eliminate the dependence, we perform sampling around the phrase \( p \). This leads to the Sampling and Occlusion (SOC) algorithm, which computes the importance of phrases as the average prediction difference after masking the phrase for each replacement of neighboring words in the input example.

Similar to SCD, SOC samples neighboring words \( x_\delta \) from a trained language model \( p(x_\delta | x_{-\delta}) \) and obtain a set of neighboring word replacement \( S \). For each replacement \( x_\delta \in S \), the algorithm computes the model prediction differences after replacing the phrase \( p \) with padding tokens. The importance \( \phi(p, x) \) is then calculated as the average prediction differences. Formally, the algorithm calculates,
\[ \phi(p, x) = \frac{1}{|S|} \sum_{x_\delta \in S} [s(x_{-\delta}; x_\delta) - s(x_{-\delta}; x_\delta; 0_p)]. \]

(8)

Sampling and Occlusion is advantageous in that it is model-agnostic and easy to implement. The input occlusion algorithm inside Eq. 8 can also be replaced with other measure of phrase importance, such as Shapley values (Shapley, 1997), with the phrase \( p \) and other input words considered as players. We expect it is helpful for longer sequences when there are multiple evidences outside the context region saturating the prediction.

4 EXPERIMENTS

We evaluate explanation algorithms on both shallow LSTM models and deep fine-tuned BERT Transformer (Devlin et al., 2018) models. We use two sentiment analysis datasets, namely the Stanford Sentiment Treebank-2 (SST-2) dataset (Socher et al., 2013) and the Yelp Sentiment Polarity dataset (Zhang et al., 2015), as well as a relation extraction dataset, namely the TACRED dataset (Zhang et al., 2017). The two tasks are modeled as binary and multi-class classification tasks respectively. For the SST-2 dataset, while it provides sentiment polarity scores for all the phrases on the nodes of the constituency parsing trees, we do not train our model on these phrases, and use these scores as the evaluation for the phrase level explanations. Our Transformer model is fine-tuned from pretrained BERT (Devlin et al., 2018) model. See Appendix A for model details.

Compared Methods We compare our explanation algorithm with following baselines: Input occlusion (Li et al., 2016) and Integrated Gradient+SHAP (GradSHAP) (Lundberg & Lee, 2017); two algorithms applied for hierarchical explanations, namely Contextual Decomposition (CD) (Murdoch et al., 2018), and Agglomerative Contextual Decomposition (ACD) (Singh et al., 2019). We also compare with a naive however neglected baseline in prior literature, which directly feed the given
Table 1: Correlation between word & phrase importance attribution and linear model coefficients & SST-2 human annotations, achieved by baselines and our explanation algorithms.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Dataset</th>
<th>BERT word ρ</th>
<th>BERT phrase ρ</th>
<th>LSTM word ρ</th>
<th>LSTM phrase ρ</th>
<th>Yelp Polarity word ρ</th>
<th>Yelp Polarity phrase ρ</th>
<th>TACRED word ρ</th>
<th>TACRED phrase ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Occlusion</td>
<td>SST-2</td>
<td>0.2229</td>
<td>0.4081</td>
<td>0.6489</td>
<td>0.4899</td>
<td>0.3781</td>
<td>0.6935</td>
<td>0.7646</td>
<td>0.5756</td>
</tr>
<tr>
<td>Direct Feed</td>
<td>SST-2</td>
<td>0.2005</td>
<td>0.4889</td>
<td>0.6798</td>
<td>0.5588</td>
<td>0.3875</td>
<td>0.7905</td>
<td>0.1986</td>
<td>0.5771</td>
</tr>
<tr>
<td>GradSHAP</td>
<td>SST-2</td>
<td>0.5073</td>
<td>0.5991</td>
<td>0.7024</td>
<td>0.5402</td>
<td>0.5791</td>
<td>0.7388</td>
<td>0.2965</td>
<td>0.6651</td>
</tr>
<tr>
<td>GradSHAP</td>
<td>Yelp Polarity</td>
<td>0.5223</td>
<td>0.4741</td>
<td>0.7271</td>
<td>0.4959</td>
<td>0.7294</td>
<td>0.9094</td>
<td>0.5324</td>
<td>0.7662</td>
</tr>
<tr>
<td>GradSHAP</td>
<td>TACRED</td>
<td>0.5481</td>
<td>0.6015</td>
<td>0.7151</td>
<td>0.5664</td>
<td>0.7180</td>
<td>0.7793</td>
<td>0.7980</td>
<td>0.6823</td>
</tr>
<tr>
<td>GradSHAP</td>
<td>TACRED</td>
<td>0.6265</td>
<td>0.6628</td>
<td>0.7226</td>
<td>0.5649</td>
<td>0.6971</td>
<td>0.7683</td>
<td>0.7982</td>
<td>0.7354</td>
</tr>
</tbody>
</table>

Table 1 shows word ρ and phrase ρ achieved by our algorithms and competitors. Generally, explanation algorithms that follow our formulations achieve highest word ρ and phrase ρ for all the datasets and models. SOC and SCD perform robustly on the deep Transformer model, achieving higher word ρ and phrase ρ than input occlusion and contextual decomposition algorithms by a large margin. We see the simple Direct Feed method provide promising results on shallow LSTM networks, but fail in deeper Transformer models. The statistic based approximation of the context independent impor-
Figure 4: Extracting phrase-level classification patterns from LSTM relation extraction model with SCD. Red indicate evidence for predicting the class, and blue indicate distractor for predicting the class. By applying the agglomerative clustering algorithm and defining a threshold score, we effectively extract “a sister, O-Person” as a classification rule for the relation per:siblings. However, we see CD fails in this example.

4.2 Explanation as Classification Pattern Extraction from Models

We show our explanation algorithm is a nature fit for extracting phrase level classification rules from neural classifiers. With the agglomerative clustering algorithm in [Singh et al., 2019], our explanation effectively identify phrase-level classification patterns without evaluating all possible phrases in the sentence even when a predefined hierarchy does not exist. Figure 4 show an example of automatically constructed hierarchy and extracted classification rules in an example in the TACRED dataset.

4.3 Enhancing Human Trust of Models

We follow the human evaluation protocol in [Singh et al., 2019] and study whether our explanations help subjects to better trust model predictions. We ask subjects to rank the provided visualizations based on how they would like to trust the model. For the SST-2 dataset, we show subjects the predictions of the fine-tuned BERT model, and the explanations generated by SOC, SCD, ACD and GradSHAP algorithms for phrases. The phrase polarities are visualized in a hierarchy with the provided parsing tree of each sentence in the dataset. For the TACRED dataset, we show the explanations provided by SOC, SCD, CD and Direct Feed algorithms on the LSTM model. We binarilize the importance of a phrase by calculating the difference between its importance to the predicted class and its top importance to other classes, and the hierarchies are constructed automatically with agglomerative clustering (Singh et al., 2019). Figure 5 shows average ranking of explanations, where 4 notes the best, and 1 notes the worst. On the SST-2 dataset, SOC achieve significantly higher ranking than ACD and GradSHAP, showing a $p$-value less than 0.05 and 0.001 respectively. On the TACRED dataset, SCD achieve the best ranking, showing significantly better ranking than CD and Direct Feed with $p$-value less than $10^{-6}$.

4.4 Parameter Analysis

Both SOC and SCD algorithms require specifying the size of the context region $N$ and the number of samples $K$. In Figure 6 (also Figure 7 in Appendix) we show the impact of these parameters. We also plot the performance curves when we pad the contexts instead of sampling. We see sampling...
the context achieves much better performance than padding the context. We also see word \( \rho \) and phrase \( \rho \) increase as the number of samples \( K \) increases. The overall performance also increases as the size of the context region \( N \) increases at the early stage, and saturates when \( N \) grows large, which implies words or phrases usually do not interact with the words that are far away them in the input. The saturation also implies the performance of trained language models can be a bottleneck of the performance.

\section{Related Works}

Interpretability of neural networks has been studied with various techniques, including probing learned features with auxiliary tasks [Tenney et al., 2019], or designing models with inherent interpretability [Bahdanau et al., 2015; Lei et al., 2016]. A major line of work, local explanation algorithms, explains predictions by assigning importance scores for input features. This line of work include input occlusion [Kadár et al., 2017], gradient based algorithms [Simonyan et al., 2013; Hechtlinger, 2016; Ancona et al., 2017], additive feature attribution methods [Ribeiro et al., 2016; Shrikumar et al., 2017; Sundararajan et al., 2017], among which Shapley value based approaches [Lundberg & Lee, 2017] have been studied intensively because of its good mathematical properties. Researchers also study how to efficiently marginalize over alternative input features to be explained [Zintgraf et al., 2017; Chang et al., 2019] for input occlusion algorithms, while our research show extra focus could be placed on marginalizing over contexts. Regarding explanations of models with structured inputs, [Chen et al., 2019] propose L-Shapley and C-Shapley for efficient approximation of Shapley values, with a similar hypothesis with us that the importance of a word is usually strongly dependent on its neighboring contexts. [Chen et al., 2018] propose a feature selection based approach for explanation in an information theoretic perspective.

On the other hand, global explanation algorithms [Guidotti et al., 2018] have also been studied for identifying generally important features, such as Feature Importance Ranking Measure [Zien et al., 2009], Accumulated Local Effects [Apley, 2016]. We note that the context independence property in our proposed methods implies we study hierarchical explanation as a global explanation problem [Guidotti et al., 2018]. Compared with local explanation algorithms, global explanation
algorithms are less studied for explaining individual predictions (Poerner et al., 2018), because they reveal the average behavior of models. However, with a hierarchical organization, we show global explanations are also powerful at explaining individual predictions, achieving better human evaluation scores and could explain compositional semantics where local explanation algorithms such as additive feature attribution algorithms totally fail. Moreover, we note that the use of explanation algorithms is not exclusive; we may apply explanation algorithms of different categories to make a more holistic explanation of model predictions.

6 CONCLUSION

In this work, we identify two desirable properties for informative hierarchical explanations of predictions, namely the non-additivity and context-independence. We propose a formulation to quantify context independent importance of words and phrases that satisfies the properties above. We revisit the prior line of works on contextual decomposition algorithms, and propose Sampling and Contextual Decomposition (SCD) algorithm. We also propose a simple and model agnostic explanation algorithm, namely the Sampling and Occlusion algorithm (SOC). Experiments on multiple datasets and models show that our explanation algorithms generate informative hierarchical explanations, help to extract classification rules from models, and enhance human trust of models.

REFERENCES


Figure 7: Word $\rho$ and phrase $\rho$ curve as the size of the context region $N$ and the number of samples $K$ change on LSTM trained on the SST-2 dataset. Dash line notes for the performance of padding the context words instead of sampling.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Label</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST-2</td>
<td>Positive</td>
<td>frighteningly evocative; insight and honesty</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>neither funny nor provocative; kill the suspense</td>
</tr>
<tr>
<td>TACRED</td>
<td>person:age</td>
<td>[PERSON], [NUMBER], was; a [NUMBER] man</td>
</tr>
<tr>
<td></td>
<td>organization:top-member</td>
<td>chief engineer of the [ORGANIZATION]</td>
</tr>
<tr>
<td></td>
<td>person:origin</td>
<td>[Nationality] citizen; tribal member from [COUNTRY]</td>
</tr>
</tbody>
</table>

Table 2: Phrase-level classification patterns extracted from models. We show the results of SCD and SOC respectively for the SST-2 and the TACRED dataset.

A IMPLEMENTATION DETAILS

Our LSTM classifiers use 1 layer unidirectional LSTM and the number of hidden units is set to 128, 500, and 300 for SST-2, Yelp, and TACRED dataset respectively. For all models, we load the pretrained 300-dimensional Glove word vectors (Pennington et al., 2014). The language model sampler is also built on LSTM and have the same parameter settings as the classifiers. Our Transformer models are fine-tuned from pretrained BERT models (Devlin et al., 2018), which have 12 layers and 768 hidden units of per representation. On three datasets, LSTM models achieve 82% accuracy, 95% accuracy, and 0.64 F1 score on average. The fine-tuned BERT models achieve 92% accuracy, 96% accuracy, and 0.68 F1 score on average. We use the same parameter settings between LSTM classifiers and language models on three datasets.

B PERFORMANCE ON ADVERSARIAL MODELS

For computing context independent importance of a phrase, an intuitive and simple alternative approach, which is nevertheless neglected in prior literature, is to only feed the input to the model and treat the prediction score as the explanation. In Table 1 while the score of the Direct Feed is lower than that of the best performing algorithms, the score is rather competitive.

The potential risk of this explanation is that it assumes model performs reasonably on incomplete sentence fragments that are significantly out of the data distribution. As a result, the explanation of short phrases can be misleading. To simulate the situation, we train a LSTM model on inversed labels on isolate words, in addition to the origin training sentences. The model could achieve the same accuracy as the original LSTM model. However, the word $\rho$ and the phrase $\rho$ of Direct Feed drop by a large margin, showing a word $\rho$ of -0.38 and 0.09. SOC and SCD are still robust on the adverse LSTM model, both showing a word $\rho$ and phrase $\rho$ of more than 0.60 and 0.55.

The masking operation could also cause performance drop because the masked sentence can be out of data distribution when explaining long phrases. For SOC, the risk can be resolved by implementing the masking operation of the phrase $p$ by another round of sampling from a language model conditioned on its context $x_{-p}$, but we do not find empirical evidence showing that it improves performance.
Figure 8: More examples about rule extraction from LSTM models trained on TACRED relation extraction dataset with SOC. Red indicate evidence for predicting the class, and blue indicate distractor for predicting the class. By applying the agglomerative clustering algorithm and defining a threshold score, we effectively extract classification rules from LSTM models. The ground truth label noted on the top.

C EXPLANATION HEATMAPS
Under review as a conference paper at ICLR 2020

It delivers some chills and sustained unease, but flounders in its quest for Deeper Meaning.

(a) SOC

(b) SCD

(c) CD

(d) GradSHAP

Figure 9: Explanation heatmaps generated by SOC, SCD, CD, and GradSHAP on a negatively predicted sentence by BERT Transformer model in SST-2 dataset. Only SOC and SCD captures adversarial conjunction connected by “but”.

15