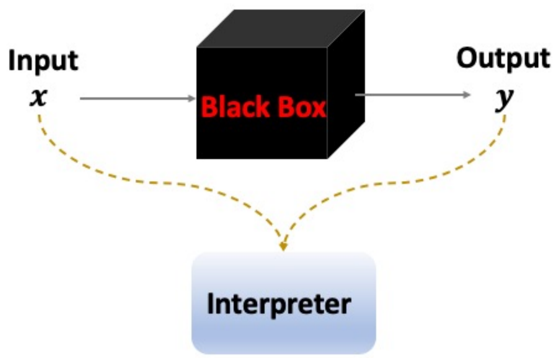
CS 4501/6501 Interpretable Machine Learning

Post-hoc explanations: beyond feature-level

Hanjie Chen, Yangfeng Ji
Department of Computer Science
University of Virginia
{hc9mx, yangfeng}@virginia.edu

Explaining Black-box Model

Model-agnostic



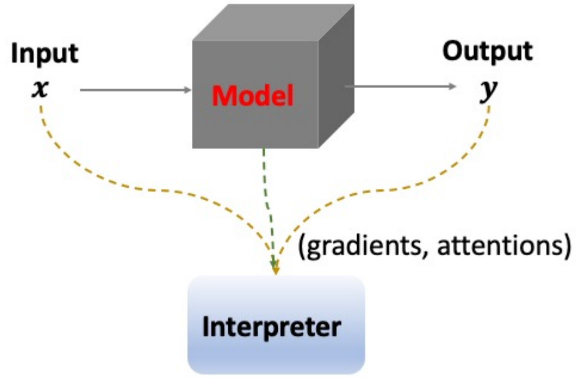
LIME

- Applicable to any black-box models
- Computational complexity
- Work well on traditional models (e.g., CNN), but not on complex DNN

SHAP

- Applicable to any black-box models
- Computational complexity
- Best performance (empirically)

Model-dependent



IG

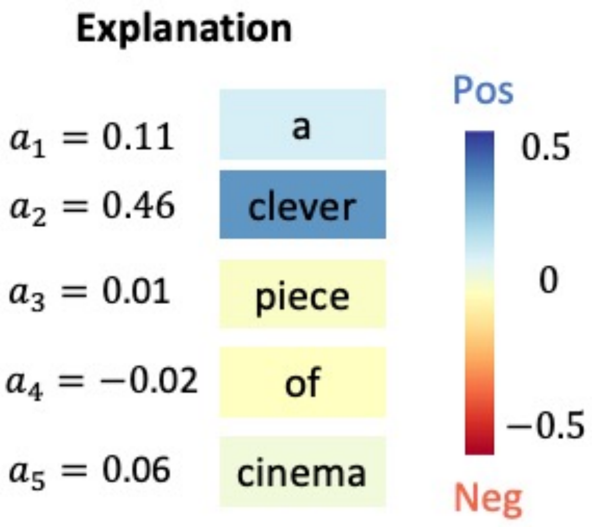
- Require access to model gradients
- Simple, fast
- Work well on both traditional and DNN models

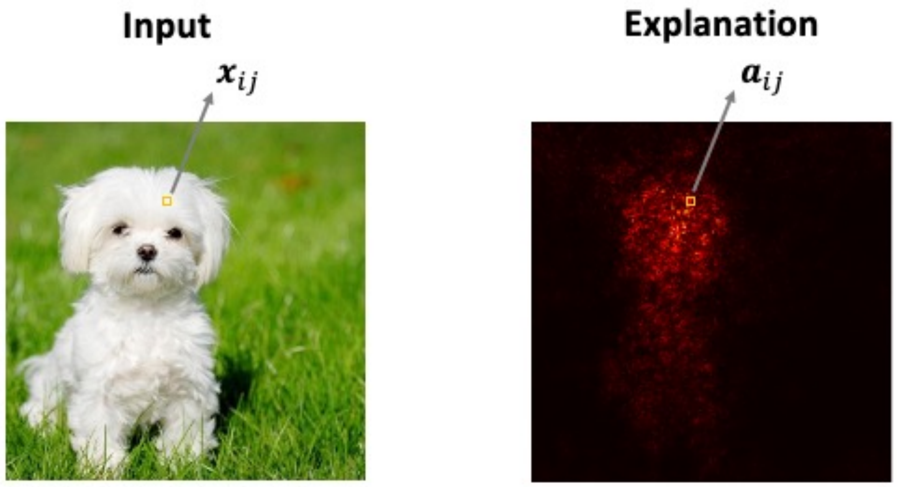
Attention

- Rely on attention mechanism
- Simple, fast (no additional computation)
- Not clear (much debate)

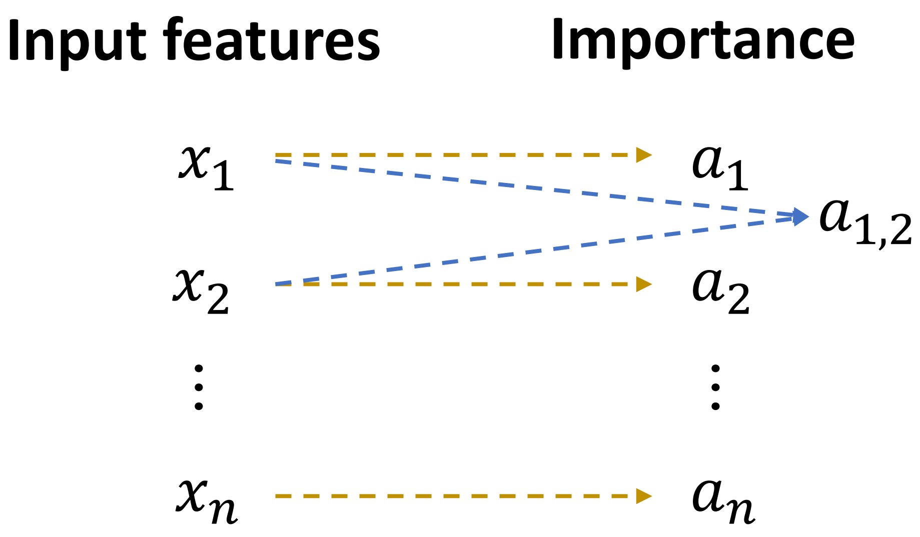
Single Feature-level Explanation

Input features	Importance
x_1	a_1
x_2	a_2
$\vdots$	$\vdots$
x_n	a_n





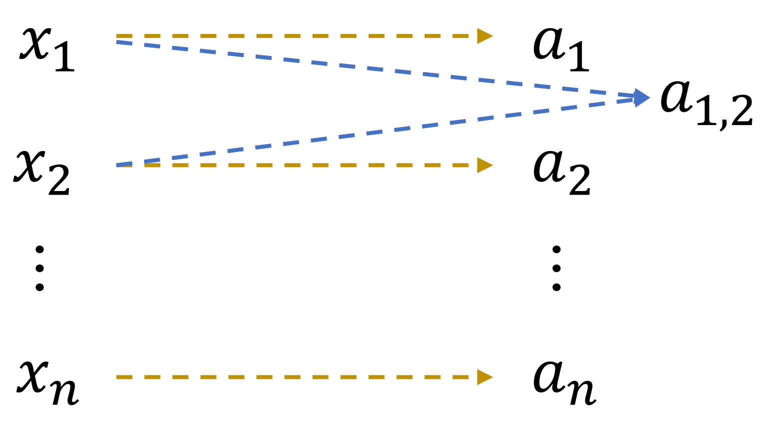
Composite Feature-level Explanation



When single features have interactions, it is critical to know the importance of the composite feature composed with these single features

Composite Feature-level Explanation

Input features **Importance**



Example 1

Input
(prediction: positive)

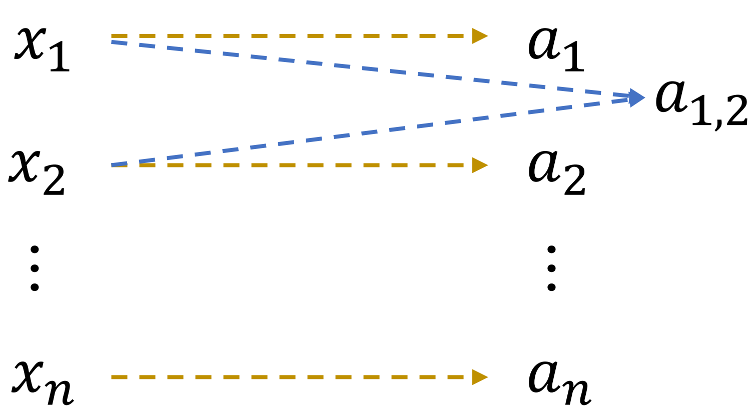
not
a
bad
journey

Explanation

Why does the model think "journey" as positive?

Composite Feature-level Explanation

Input features **Importance**



Example 1

Input
(prediction: positive)

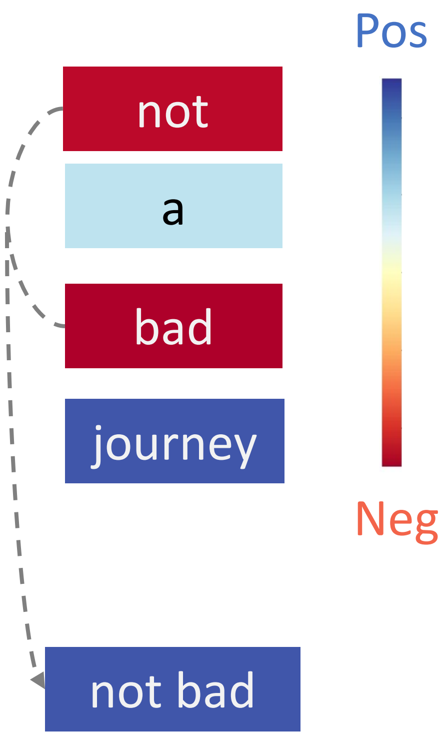
not

a

bad

journey

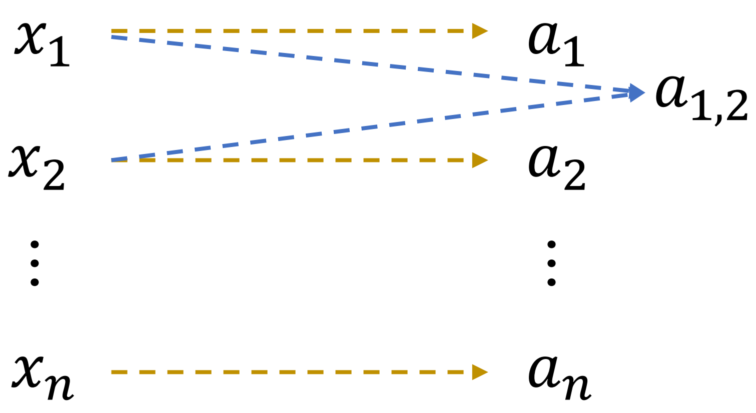
Explanation



Composite feature importance

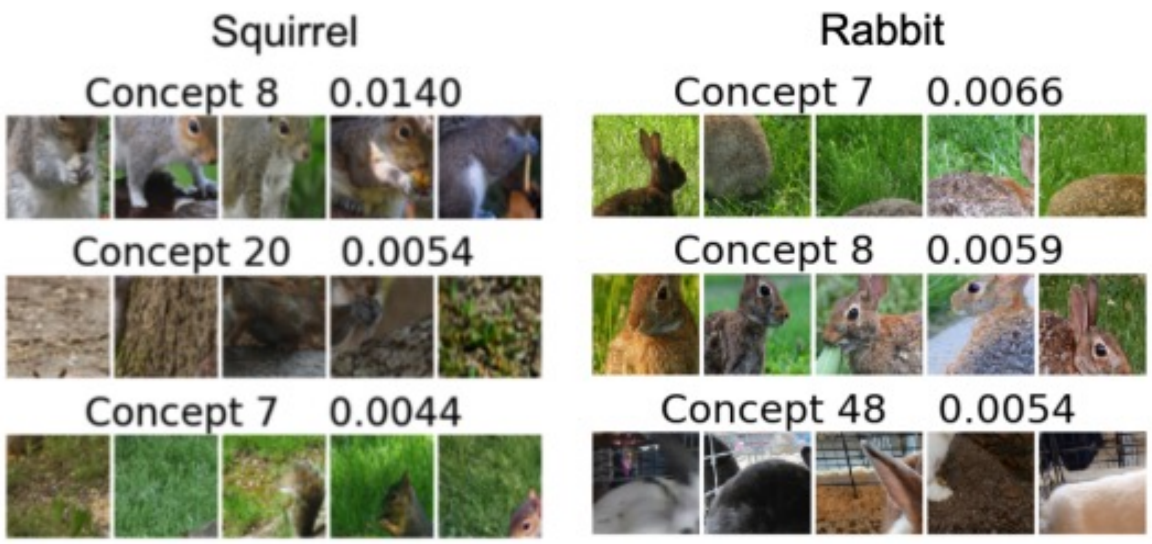
Composite Feature-level Explanation

Input features **Importance**



Example 2

Composite feature importance is more intuitive for human understanding



(Yeh et al., 2020)

Beyond Feature Attribution

- Contextual Decomposition (CD)
- Hierarchical Explanation via Divisive Generation (HEDGE)

CD

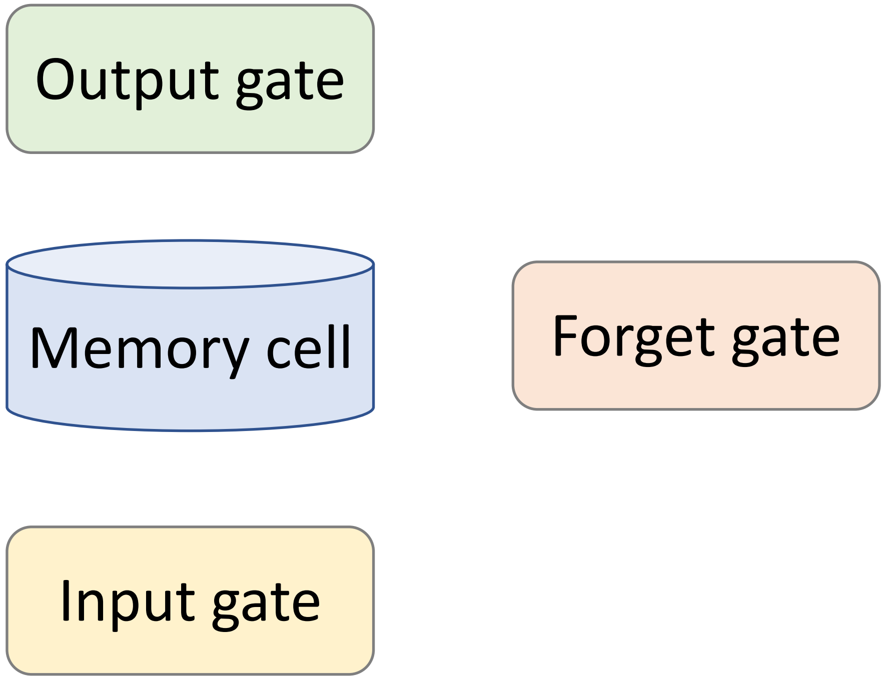
Beyond Word Importance: Contextual Decomposition to Extract Interactions From LSTMs

W. James Murdoch, Peter J. Liu, Bin Yu

(ICLR, 2018)

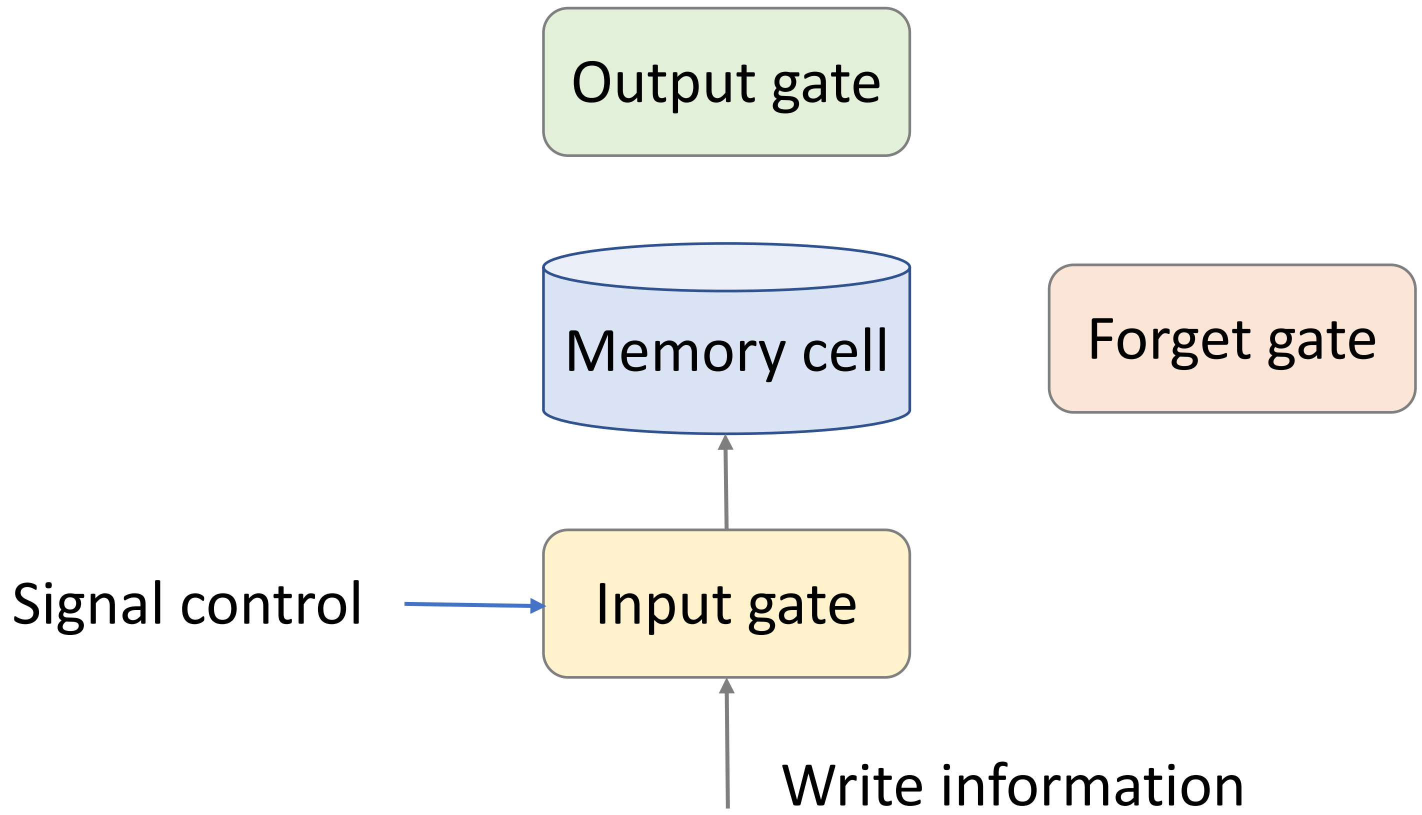
LSTM

- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]



LSTM

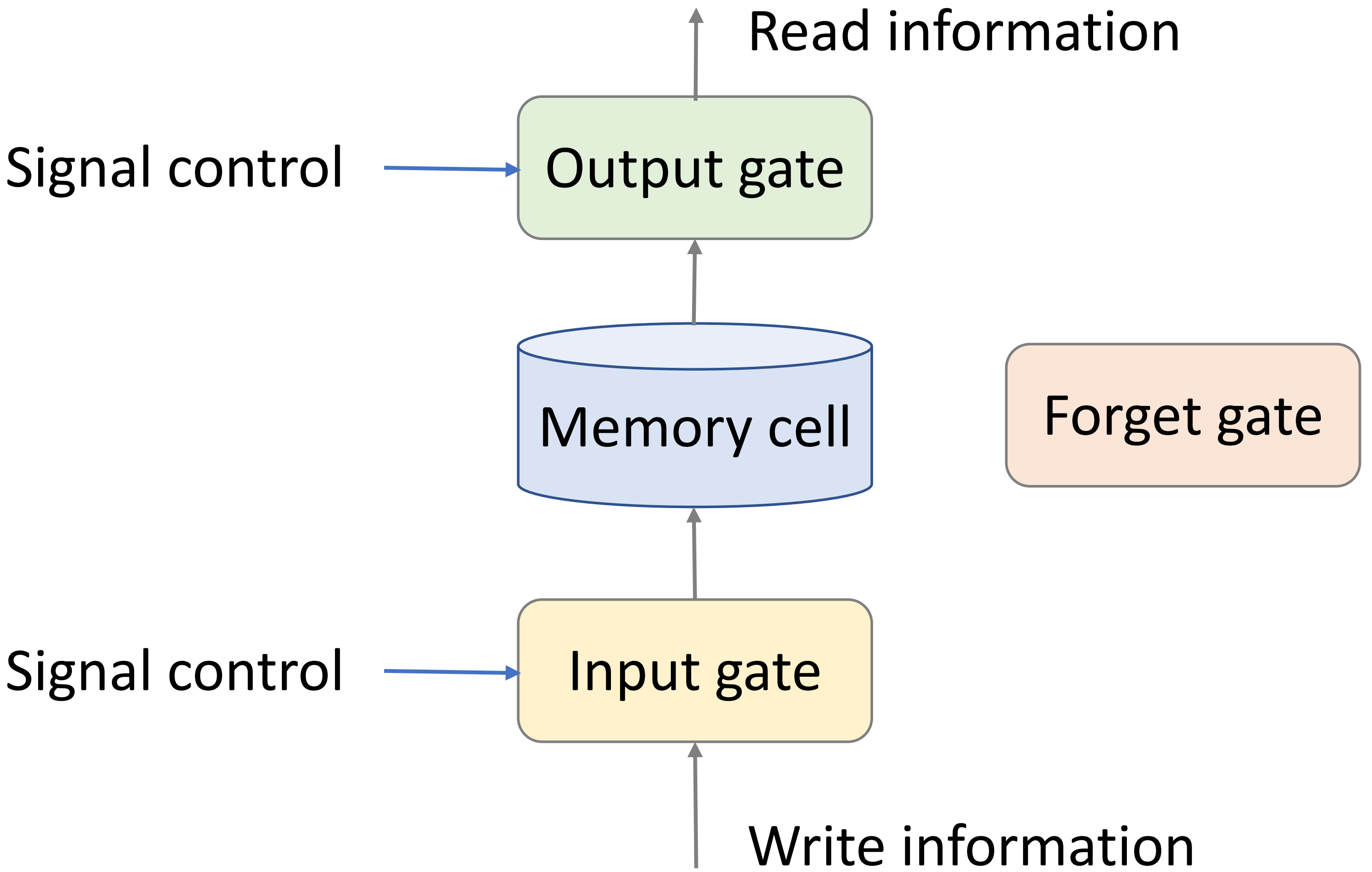
- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]



Source: https://www.youtube.com/watch?v=xCGidAeyS4M

LSTM

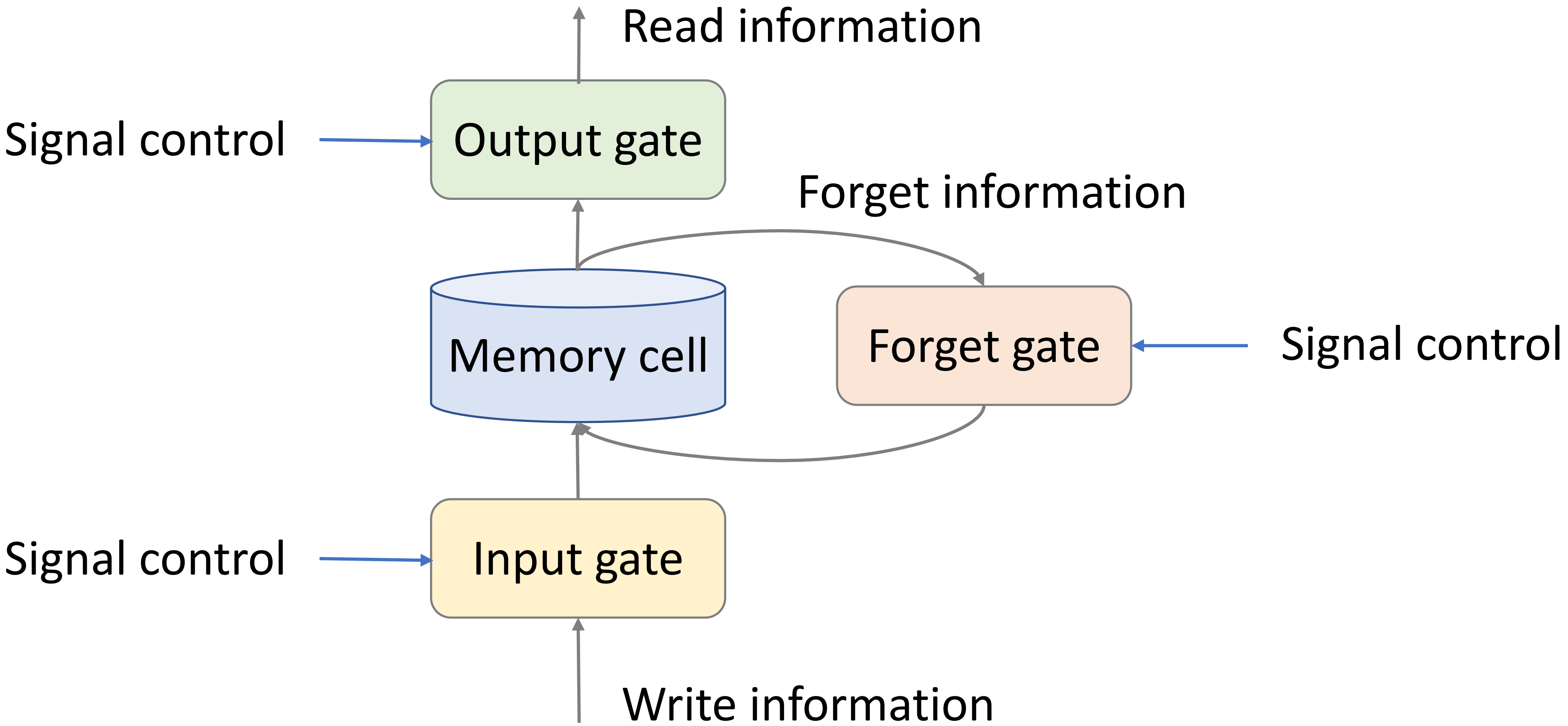
- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]



Source: https://www.youtube.com/watch?v=xCGidAeyS4M

LSTM

- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]



Source: https://www.youtube.com/watch?v=xCGidAeyS4M

LSTM

- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]

Input word embeddings: $x_1, \cdots, x_T \in \mathbb{R}^{d_1}$

Cell: $c_t \in \mathbb{R}^{d_2}$

State vector: $h_t \in \mathbb{R}^{d_2}$

$(h_0 = c_0 = 0)$

LSTM

- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]

Input word embeddings: $x_1, \cdots, x_T \in \mathbb{R}^{d_1}$

Cell: $c_t \in \mathbb{R}^{d_2}$

State vector: $h_t \in \mathbb{R}^{d_2}$

$(h_0 = c_0 = 0)$

(Output gate) $o_t = \sigma(W_o \boxed{x_t} + V_o \boxed{h_{t-1}} + b_o)$

(Forget gate) $f_t = \sigma\left(W_f \boxed{x_t} + V_f \boxed{h_{t-1}} + b_f\right)$

(Input gate) $i_t = \sigma(W_i \boxed{x_t} + V_i \boxed{h_{t-1}} + b_i)$

x_t: current input

h_{t-1}: previous output

LSTM

- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]

Input word embeddings: $x_1, \cdots, x_T \in \mathbb{R}^{d_1}$

Cell: $c_t \in \mathbb{R}^{d_2}$

$(h_0 = c_0 = 0)$

State vector: $h_t \in \mathbb{R}^{d_2}$

(Output gate) $o_t = \sigma(\boxed{W_o} x_t + \boxed{V_o} h_{t-1} + \boxed{b_o})$

(Forget gate) $f_t = \sigma\left(\boxed{W_f} x_t + \boxed{V_f} h_{t-1} + \boxed{b_f}\right)$

(Input gate) $i_t = \sigma(\boxed{W_i} x_t + \boxed{V_i} h_{t-1} + \boxed{b_i})$

W, V, b are model parameters

LSTM

- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]

Input word embeddings: $x_1, \cdots, x_T \in \mathbb{R}^{d_1}$

Cell: $c_t \in \mathbb{R}^{d_2}$

State vector: $h_t \in \mathbb{R}^{d_2}$

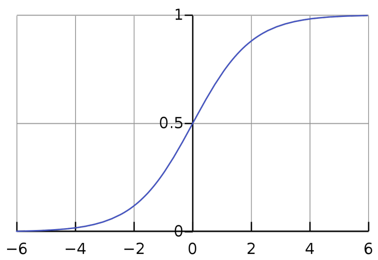
$(h_0 = c_0 = 0)$

(Output gate) $o_t = \sigma(W_o x_t + V_o h_{t-1} + b_o)$

(Forget gate) $f_t = \sigma\left(W_f x_t + V_f h_{t-1} + b_f\right)$

(Input gate) $i_t = \sigma(W_i x_t + V_i h_{t-1} + b_i)$

$\sigma(\cdot)$: sigmoid function



Range: 0 to 1

LSTM

- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]

Input word embeddings: $x_1, \cdots, x_T \in \mathbb{R}^{d_1}$

Cell: $c_t \in \mathbb{R}^{d_2}$

State vector: $h_t \in \mathbb{R}^{d_2}$

$(h_0 = c_0 = 0)$

(Output gate) $o_t = \sigma(W_o x_t + V_o h_{t-1} + b_o)$

(Forget gate) $f_t = \sigma\left(W_f x_t + V_f h_{t-1} + b_f\right)$

(Input gate) $i_t = \sigma(W_i x_t + V_i h_{t-1} + b_i)$

$$g_t = tanh\left(W_g x_t + V_g h_{t-1} + b_g\right)$$

$$c_t = f_t \odot c_{t-1} + \boxed{i_t \odot g_t}$$ Information written into the cell

$$h_t = o_t \odot tanh(c_t)$$

LSTM

- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]

Input word embeddings: $x_1, \cdots, x_T \in \mathbb{R}^{d_1}$

Cell: $c_t \in \mathbb{R}^{d_2}$ $(h_0 = c_0 = 0)$

State vector: $h_t \in \mathbb{R}^{d_2}$

(Output gate) $o_t = \sigma(W_o x_t + V_o h_{t-1} + b_o)$

(Forget gate) $f_t = \sigma\left(W_f x_t + V_f h_{t-1} + b_f\right)$

(Input gate) $i_t = \sigma(W_i x_t + V_i h_{t-1} + b_i)$

$$g_t = tanh\left(W_g x_t + V_g h_{t-1} + b_g\right)$$

$$c_t = \boxed{f_t \odot c_{t-1}} + i_t \odot g_t$$

Information left in the cell after forgetting

$$h_t = o_t \odot tanh(c_t)$$

LSTM

- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]

Input word embeddings: $x_1, \cdots, x_T \in \mathbb{R}^{d_1}$

Cell: $c_t \in \mathbb{R}^{d_2}$

State vector: $h_t \in \mathbb{R}^{d_2}$

$(h_0 = c_0 = 0)$

(Output gate) $o_t = \sigma(W_o x_t + V_o h_{t-1} + b_o)$

(Forget gate) $f_t = \sigma\left(W_f x_t + V_f h_{t-1} + b_f\right)$

(Input gate) $i_t = \sigma(W_i x_t + V_i h_{t-1} + b_i)$

$$g_t = tanh\left(W_g x_t + V_g h_{t-1} + b_g\right)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$h_t = o_t \odot tanh(c_t)$ Current output

LSTM

- Long Short-term Memory Network (LSTM) [Hochreiter and Schmidhuber, 1997]

Input word embeddings: $x_1, \cdots, x_T \in \mathbb{R}^{d_1}$
Cell: $c_t \in \mathbb{R}^{d_2}$
State vector: $h_t \in \mathbb{R}^{d_2}$

$(h_0 = c_0 = 0)$

(Output gate) $o_t = \sigma(W_o x_t + V_o h_{t-1} + b_o)$

(Forget gate) $f_t = \sigma\left(W_f x_t + V_f h_{t-1} + b_f\right)$

(Input gate) $i_t = \sigma(W_i x_t + V_i h_{t-1} + b_i)$

$g_t = tanh\left(W_g x_t + V_g h_{t-1} + b_g\right)$

$c_t = f_t \odot c_{t-1} + i_t \odot g_t$

$h_t = o_t \odot tanh(c_t)$

$\xrightarrow{t = 1, \cdots, T}$

Probability distribution

$p = softmax(W h_T)$

Question?

CD

- Contextual Decomposition

An arbitrary phrase: $x_q, \cdots, x_r$ $(1 \leq q \leq r \leq T)$

Decompose each c_t **and** h_t **into a sum of two contributions**

Goal: compute the contribution of the phrase to model prediction

$$h_t = \beta_t + \gamma_t$$

$$c_t = \beta_t^c + \gamma_t^c$$

β_t, β_t^c : contributions made solely by the given phrase

γ_t, γ_t^c : contributions involving elements outside of the phrase

CD

- Contextual Decomposition

An arbitrary phrase: $x_q, \cdots, x_r$ $(1 \leq q \leq r \leq T)$

Decompose each c_t and h_t into a sum of two contributions

Goal: compute the contribution of the phrase to model prediction

$$h_t = \beta_t + \gamma_t$$

$$c_t = \beta_t^c + \gamma_t^c$$

β_t, β_t^c : contributions made solely by the given phrase

γ_t, γ_t^c : contributions involving elements outside of the phrase

$$p = softmax(Wh_T) \longrightarrow p = softmax(\underline{W\beta_T} + W\gamma_T)$$

the phrase's contribution to the LSTM's prediction

CD

- Contextual Decomposition

An arbitrary phrase: $x_q, \cdots, x_r \ (1 \le q \le r \le T)$

Decompose each c_t **and** h_t **into a sum of two contributions**

Goal: compute the contribution of the phrase to model prediction

$$h_t = \beta_t + \gamma_t$$

$$c_t = \beta_t^c + \gamma_t^c$$

β_t, β_t^c : contributions made solely by the given phrase

γ_t, γ_t^c : contributions involving elements outside of the phrase

$$p = softmax(Wh_T) \longrightarrow p = softmax(\underline{W\beta_T} + W\gamma_T)$$

the phrase's contribution to the LSTM's prediction

How to compute β_t, γ_t?

CD

- Contextual Decomposition

Disambiguating interactions between gates

$$i_t = \sigma(W_i x_t + V_i h_{t-1} + b_i)$$

$$= L_\sigma(W_i x_t) + L_\sigma(V_i h_{t-1}) + L_\sigma(b_i)$$

$$g_t = tanh\left(W_g x_t + V_g h_{t-1} + b_g\right)$$

$$= L_{tanh}\left(W_g x_t\right) + L_{tanh}\left(V_g h_{t-1}\right) + L_{tanh}\left(b_g\right)$$

Assume we have a way of linearizing the gates

CD

- Contextual Decomposition

Disambiguating interactions between gates

$$i_t \odot g_t$$

$$= \left(L_\sigma(W_i x_t) + L_\sigma(V_i h_{t-1}) + L_\sigma(b_i)\right) \odot \left(L_{tanh}(W_g x_t) + L_{tanh}(V_g h_{t-1}) + L_{tanh}(b_g)\right)$$

$$= \left(L_\sigma(W_i x_t) + L_\sigma(V_i \beta_{t-1}) + L_\sigma(V_i \gamma_{t-1}) + L_\sigma(b_i)\right) \odot \left(L_{tanh}(W_g x_t) + L_{tanh}(V_g \beta_{t-1}) + L_{tanh}(V_g \gamma_{t-1}) + L_{tanh}(b_g)\right)$$

CD

- Contextual Decomposition

Disambiguating interactions between gates

$$i_t \odot g_t$$

$$= \left(L_\sigma(W_i x_t) + L_\sigma(V_i h_{t-1}) + L_\sigma(b_i)\right) \odot \left(L_{tanh}(W_g x_t) + L_{tanh}(V_g h_{t-1}) + L_{tanh}(b_g)\right)$$

$$= \left(L_\sigma(W_i x_t) + L_\sigma(V_i \beta_{t-1}) + L_\sigma(V_i \gamma_{t-1}) + L_\sigma(b_i)\right) \odot \left(L_{tanh}(W_g x_t) + L_{tanh}(V_g \beta_{t-1}) + L_{tanh}(V_g \gamma_{t-1}) + L_{tanh}(b_g)\right)$$

Cross-terms:

- ❑ solely from the phrase, e.g., $L_\sigma(V_i \beta_{t-1}) \odot L_{tanh}(V_g \beta_{t-1})$
- ❑ from some interaction between the phrase and other factors , e.g., $L_\sigma(V_i \beta_{t-1}) \odot L_{tanh}(V_g \gamma_{t-1})$
- ❑ purely from other factors , e.g., $L_\sigma(b_i) \odot L_{tanh}(V_g \gamma_{t-1})$

CD

- Contextual Decomposition

Disambiguating interactions between gates

$$i_t \odot g_t$$

$$= \left(L_\sigma(W_i x_t) + L_\sigma(V_i h_{t-1}) + L_\sigma(b_i)\right) \odot \left(L_{tanh}(W_g x_t) + L_{tanh}(V_g h_{t-1}) + L_{tanh}(b_g)\right)$$

$$= \left(L_\sigma(W_i x_t) + L_\sigma(V_i \beta_{t-1}) + L_\sigma(V_i \gamma_{t-1}) + L_\sigma(b_i)\right) \odot \left(L_{tanh}(W_g x_t) + L_{tanh}(V_g \beta_{t-1}) + L_{tanh}(V_g \gamma_{t-1}) + L_{tanh}(b_g)\right)$$

Cross-terms:

- ☑ solely from the phrase, e.g., $L_\sigma(V_i \beta_{t-1}) \odot L_{tanh}(V_g \beta_{t-1})$ β_t^u
- ☐ from some interaction between the phrase and other factors , e.g., $L_\sigma(V_i \beta_{t-1}) \odot L_{tanh}(V_g \gamma_{t-1})$
- ☐ purely from other factors , e.g., $L_\sigma(b_i) \odot L_{tanh}(V_g \gamma_{t-1})$

CD

- Contextual Decomposition

Disambiguating interactions between gates

$$i_t \odot g_t$$

$$= \left(L_\sigma(W_i x_t) + L_\sigma(V_i h_{t-1}) + L_\sigma(b_i)\right) \odot \left(L_{tanh}(W_g x_t) + L_{tanh}(V_g h_{t-1}) + L_{tanh}(b_g)\right)$$

$$= \left(L_\sigma(W_i x_t) + L_\sigma(V_i \beta_{t-1}) + L_\sigma(V_i \gamma_{t-1}) + L_\sigma(b_i)\right) \odot \left(L_{tanh}(W_g x_t) + L_{tanh}(V_g \beta_{t-1}) + L_{tanh}(V_g \gamma_{t-1}) + L_{tanh}(b_g)\right)$$

Cross-terms:

- ❑ solely from the phrase, e.g., $L_\sigma(V_i \beta_{t-1}) \odot L_{tanh}(V_g \beta_{t-1})$
- ☑ from some interaction between the phrase and other factors , e.g., $L_\sigma(V_i \beta_{t-1}) \odot L_{tanh}(V_g \gamma_{t-1})$
- ☑ purely from other factors , e.g., $L_\sigma(b_i) \odot L_{tanh}(V_g \gamma_{t-1})$ γ_t^u

CD

- Contextual Decomposition

Disambiguating interactions between gates

$$i_t \odot g_t = \beta_t^u + \gamma_t^u$$

$$f_t \odot c_{t-1} = \beta_t^f + \gamma_t^f$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$= \beta_t^u + \gamma_t^u + \beta_t^f + \gamma_t^f$$

$$= \beta_t^c + \gamma_t^c$$

CD

- Contextual Decomposition

Disambiguating interactions between gates

$$i_t \odot g_t = \beta_t^u + \gamma_t^u$$

$$f_t \odot c_{t-1} = \beta_t^f + \gamma_t^f$$

$$\begin{aligned} c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ &= \beta_t^u + \gamma_t^u + \beta_t^f + \gamma_t^f \\ &= \beta_t^c + \gamma_t^c \end{aligned}$$

$$\begin{aligned} h_t &= o_t \odot tanh(c_t) \\ &= o_t \odot tanh(\beta_t^c + \gamma_t^c) \\ &= o_t \odot \left(L_{tanh}(\beta_t^c) + L_{tanh}(\gamma_t^c)\right) \\ &= o_t \odot L_{tanh}(\beta_t^c) + o_t \odot L_{tanh}(\gamma_t^c) \\ &= \beta_t + \gamma_t \end{aligned}$$

CD

- Contextual Decomposition

Disambiguating interactions between gates

$$i_t \odot g_t = \beta_t^u + \gamma_t^u$$

$$f_t \odot c_{t-1} = \beta_t^f + \gamma_t^f$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$= \beta_t^u + \gamma_t^u + \beta_t^f + \gamma_t^f$$

$$= \beta_t^c + \gamma_t^c$$

$$h_t = o_t \odot tanh(c_t)$$

$$= o_t \odot tanh(\beta_t^c + \gamma_t^c)$$

$$= o_t \odot \left(L_{tanh}(\beta_t^c) + L_{tanh}(\gamma_t^c)\right)$$

$$= o_t \odot L_{tanh}(\beta_t^c) + o_t \odot L_{tanh}(\gamma_t^c)$$

$$= \beta_t + \gamma_t$$

Iteratively decomposing until we get
$h_T = \beta_T + \gamma_T$
$\beta_0 = \gamma_0 = 0$
β_t: $x_{t\ (q \le t \le r)}$ γ_t: $x_{t\ (t>r, t<q)}$

CD

- Contextual Decomposition

Disambiguating interactions between gates

$$i_t \odot g_t = \beta_t^u + \gamma_t^u$$

$$f_t \odot c_{t-1} = \beta_t^f + \gamma_t^f$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$= \beta_t^u + \gamma_t^u + \beta_t^f + \gamma_t^f$$

$$= \beta_t^c + \gamma_t^c$$

$$h_t = o_t \odot tanh(c_t)$$

$$= o_t \odot tanh(\beta_t^c + \gamma_t^c)$$

$$= o_t \odot \left(L_{tanh}(\beta_t^c) + L_{tanh}(\gamma_t^c)\right)$$

$$= o_t \odot L_{tanh}(\beta_t^c) + o_t \odot L_{tanh}(\gamma_t^c)$$

$$= \beta_t + \gamma_t$$

What are the linearizing functions L_σ, L_{tanh}?

CD

- Contextual Decomposition

Linearizing activation functions (L_{σ}, L_{tanh})

$$tanh\left(\sum_{i=1}^{N} y_i\right) = \sum_{i=1}^{N} L_{tanh}(y_i) \qquad (N \leq 4)$$

CD

- Contextual Decomposition

Linearizing activation functions (L_σ, L_{tanh})

$$tanh\left(\sum_{i=1}^{N} y_i\right) = \sum_{i=1}^{N} L_{tanh}(y_i) \qquad (N \leq 4)$$

Telescoping sum (given a natural ordering to $\{y_i\}$)

$$L_{tanh}(y_k) = tanh\left(\sum_{j=1}^{k} y_j\right) - tanh\left(\sum_{j=1}^{k-1} y_j\right)$$

CD

- Contextual Decomposition

Linearizing activation functions (L_σ, L_{tanh})

$$tanh\left(\sum_{i=1}^{N} y_i\right) = \sum_{i=1}^{N} L_{tanh}(y_i) \qquad (N \leq 4)$$

Telescoping sum (given a natural ordering to $\{y_i\}$)

$$L_{tanh}(y_k) = tanh\left(\sum_{j=1}^{k} y_j\right) - tanh\left(\sum_{j=1}^{k-1} y_j\right)$$

$$\sum_{i=1}^{N} L_{tanh}(y_i) = tanh\left(\sum_{j=1}^{N} y_j\right) - tanh\left(\sum_{j=1}^{N-1} y_j\right) + tanh\left(\sum_{j=1}^{N-1} y_j\right) - tanh\left(\sum_{j=1}^{N-2} y_j\right) + \cdots + tanh\left(\sum_{j=1}^{2} y_j\right) - tanh\left(\sum_{j=1}^{1} y_j\right) + tanh\left(\sum_{j=1}^{1} y_j\right)$$

CD

- Contextual Decomposition

Linearizing activation functions (L_σ, L_{tanh})

$$tanh\left(\sum_{i=1}^{N} y_i\right) = \sum_{i=1}^{N} L_{tanh}(y_i) \qquad (N \leq 4)$$

Telescoping sum (given a natural ordering to $\{y_i\}$)

$$L_{tanh}(y_k) = tanh\left(\sum_{j=1}^{k} y_j\right) - tanh\left(\sum_{j=1}^{k-1} y_j\right)$$

$$\sum_{i=1}^{N} L_{tanh}(y_i) = tanh\left(\sum_{j=1}^{N} y_j\right) - \cancel{tanh\left(\sum_{j=1}^{N-1} y_j\right)} + \cancel{tanh\left(\sum_{j=1}^{N-1} y_j\right)} - \cancel{tanh\left(\sum_{j=1}^{N-2} y_j\right)} + \cdots + \cancel{tanh\left(\sum_{j=1}^{2} y_j\right)} - \cancel{tanh\left(\sum_{j=1}^{1} y_j\right)} + \cancel{tanh\left(\sum_{j=1}^{1} y_j\right)}$$

CD

- Contextual Decomposition

Linearizing activation functions (L_σ, L_{tanh})

$$tanh\left(\sum_{i=1}^{N} y_i\right) = \sum_{i=1}^{N} L_{tanh}(y_i) \qquad (N \leq 4)$$

Telescoping sum (given a natural ordering to $\{y_i\}$)

$$L_{tanh}(y_k) = tanh\left(\sum_{j=1}^{k} y_j\right) - tanh\left(\sum_{j=1}^{k-1} y_j\right)$$

$\{\beta_{t-1}, \gamma_{t-1}, x_t\}$ have no clear ordering

CD

- Contextual Decomposition

Linearizing activation functions (L_σ, L_{tanh})

$$tanh\left(\sum_{i=1}^{N} y_i\right) = \sum_{i=1}^{N} L_{tanh}(y_i) \qquad (N \leq 4)$$

Telescoping sum (given a natural ordering to $\{y_i\}$)

$$L_{tanh}(y_k) = tanh\left(\sum_{j=1}^{k} y_j\right) - tanh\left(\sum_{j=1}^{k-1} y_j\right)$$

All permutations: $\pi_1, \cdots, \pi_{M_N}$

$\pi_i^{-1}(k)$: the position of y_k in π_i

Average over all orderings

$$L_{tanh}(y_k) = \frac{1}{M_N}\sum_{i=1}^{M_N}\left[tanh\left(\sum_{j=1}^{\pi_i^{-1}(k)} y_{\pi_i(j)}\right) - tanh\left(\sum_{j=1}^{\pi_i^{-1}(k)-1} y_{\pi_i(j)}\right)\right]$$

Summary

- **LSTM**

  Input word embeddings: $x_1, \cdots, x_T \in \mathbb{R}^{d_1}$

$$
\begin{aligned}
o_t &= \sigma(W_o x_t + V_o h_{t-1} + b_o) \\
f_t &= \sigma\left(W_f x_t + V_f h_{t-1} + b_f\right) \\
i_t &= \sigma(W_i x_t + V_i h_{t-1} + b_i) \\
g_t &= tanh\left(W_g x_t + V_g h_{t-1} + b_g\right) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot tanh(c_t)
\end{aligned}
$$

- **Contribution of an arbitrary phrase:**

  $x_q, \cdots, x_r \ (1 \leq q \leq r \leq T)$

$$
\sigma(\cdot) = \sum L_\sigma, tanh(\cdot) = \sum L_{tanh}
$$

$$
h_t = \beta_t + \gamma_t
$$

$$
\xrightarrow{t = 1, \cdots, T}
$$

$$
p = softmax(W\underline{\beta_T} + W\gamma_T)
$$

the phrase's contribution
to the LSTM's prediction

Question?

Visualizations

Text
"used to be my favorite" (negative)
"not worth the time" (negative)

Attribution Method	Heat Map	
Gradient	used to be my favorite	not worth the time
Leave One Out (Li et al., 2016)	used to be my favorite	not worth the time
Cell decomposition (Murdoch & Szlam, 2017)	used to be my favorite	not worth the time
Integrated gradients (Sundararajan et al., 2017)	used to be my favorite	not worth the time
Contextual decomposition	used to be my favorite	not worth the time

Legend Very Negative Negative Neutral Positive Very Positive

Visualizations

The first phrase is positive, but the second one is negative

CD is the only method that accurately captures this dynamic

Attribution Method	Heat Map
Gradient	It's easy to love Robin Tunney – she's pretty and she can act – but it gets harder and harder to understand her choices.
Leave one out (Li et al., 2016)	It's easy to love Robin Tunney – she's pretty and she can act – but it gets harder and harder to understand her choices.
Cell decomposition (Murdoch & Szlam, 2017)	It's easy to love Robin Tunney – she's pretty and she can act – but it gets harder and harder to understand her choices.
Integrated gradients (Sundararajan et al., 2017)	It's easy to love Robin Tunney – she's pretty and she can act – but it gets harder and harder to understand her choices.
Contextual decomposition	It's easy to love Robin Tunney – she's pretty and she can act – but it gets harder and harder to understand her choices.

Legend: Very Negative | Negative | Neutral | Positive | Very Positive

Discussion

- CD is model-dependent
- Decomposing complex DNN (e.g., transformer) is not trivial

Beyond Feature Attribution

- Contextual Decomposition (CD)

- Hierarchical Explanation via Divisive Generation (HEDGE)

HEDGE

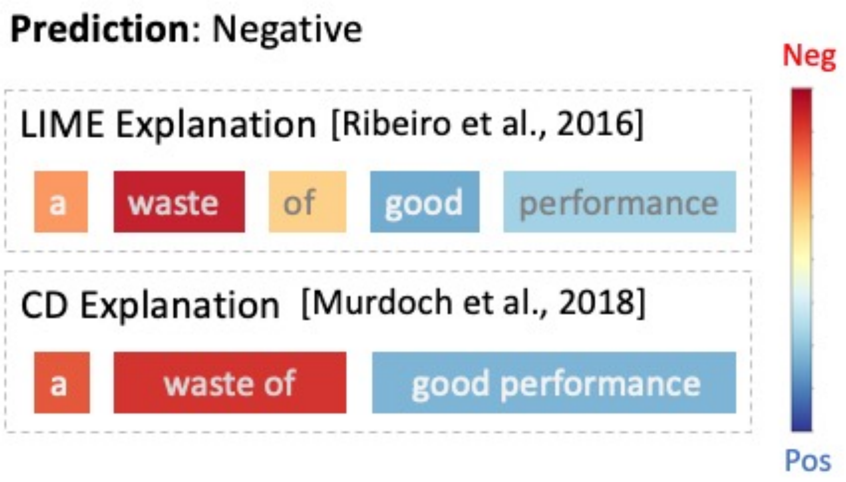
Generating Hierarchical Explanations on Text Classification
via Feature Interaction Detection

Hanjie Chen, Guangtao Zheng, Yangfeng Ji

(ACL, 2020)

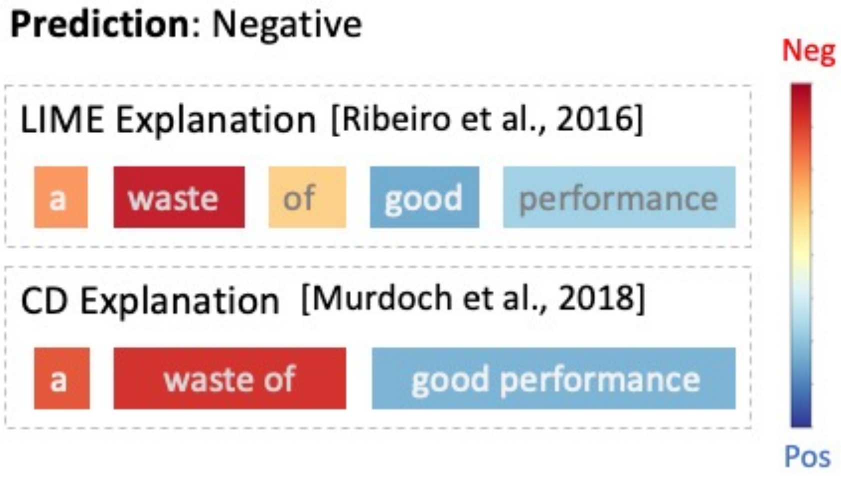
HEDGE

Why we need hierarchical explanations?



HEDGE

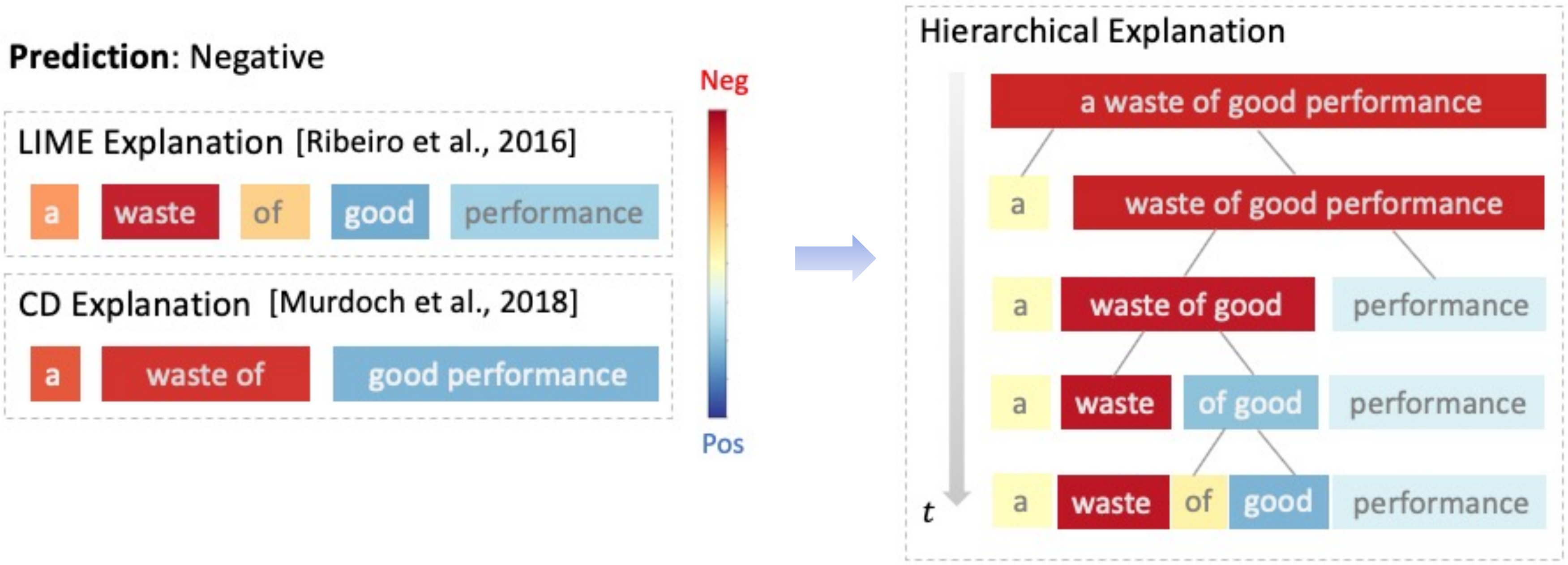
Why we need hierarchical explanations?



How the word “good” or a phrase related to “good” contributes to the model prediction?

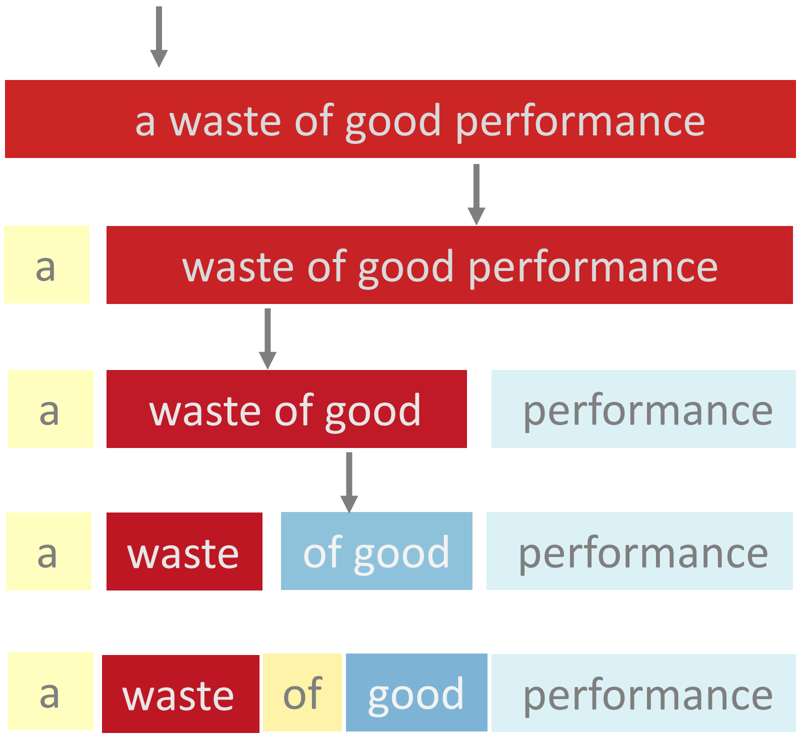
HEDGE

Why we need hierarchical explanations?



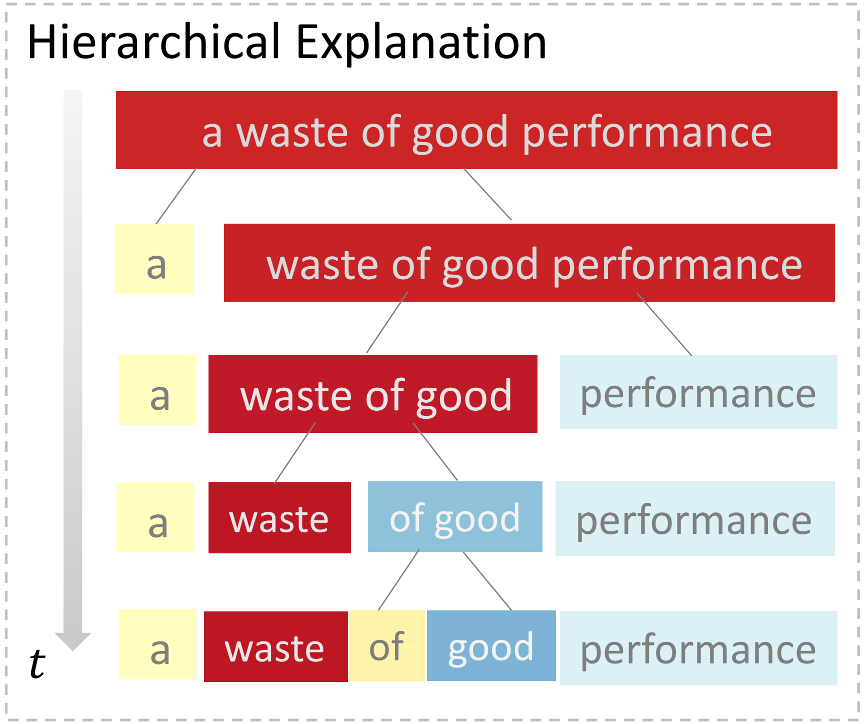
Hierarchical explanation via divisive generation (HEDGE)

- Where is the dividing point?
- Which text segment should be split?
- How to quantify feature importance?



HEDGE

- Definition
  - A text with n words: $\boldsymbol{x} = (x_1, \cdots, x_n)$
  - A text span: $\boldsymbol{x}_{(s_i, s_{i+1}]} = \left(x_{s_i+1}, \cdots, x_{s_{i+1}}\right)$
  - A partition: $\mathcal{P} = \left\{\boldsymbol{x}_{(0,s_1]}, \boldsymbol{x}_{(s_1, s_2]}, \cdots, \boldsymbol{x}_{(s_{P-1},n]}\right\}$
  - Interaction score: $\phi(\cdot,\cdot)$
  - Importance score: $\psi(\cdot)$

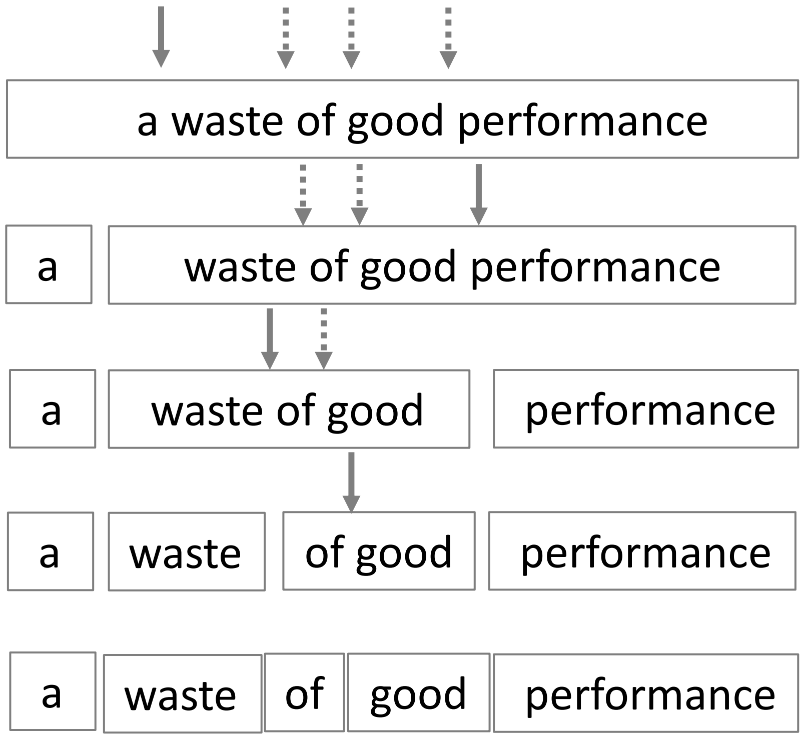


HEDGE

- Where is the dividing point?

The *local weakest* interaction point:

$$\min_{j \in (s_i, s_{i+1})} \phi(\boldsymbol{x}_{(s_i, j]}, \boldsymbol{x}_{(j, s_{i+1}]} | \mathcal{P})$$



HEDGE

- Where is the dividing point?

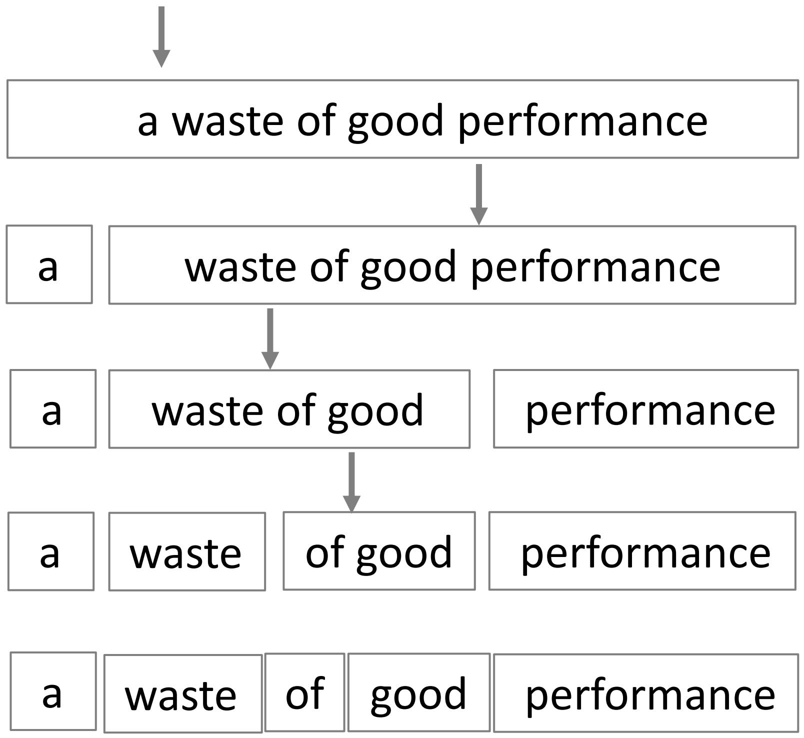
  The *local weakest* interaction point:

$$\min_{j \in (s_i,\, s_{i+1})} \phi\left(\boldsymbol{x}_{(s_i, j]}, \boldsymbol{x}_{(j,\, s_{i+1}]} | \mathcal{P}\right)$$

- Which text segment should be split?

  The *global weakest* interaction point:

$$\min_{\boldsymbol{x}_{(s_i,\, s_{i+1}]} \in \mathcal{P}} \min_{j \in (s_i,\, s_{i+1})} \phi\left(\boldsymbol{x}_{(s_i, j]}, \boldsymbol{x}_{(j,\, s_{i+1}]} | \mathcal{P}\right)$$



HEDGE

- Where is the dividing point?

The *local weakest* interaction point:

$$\min_{j \in (s_i, s_{i+1})} \phi\left(\boldsymbol{x}_{(s_i, j]}, \boldsymbol{x}_{(j, s_{i+1}]} | \mathcal{P}\right)$$

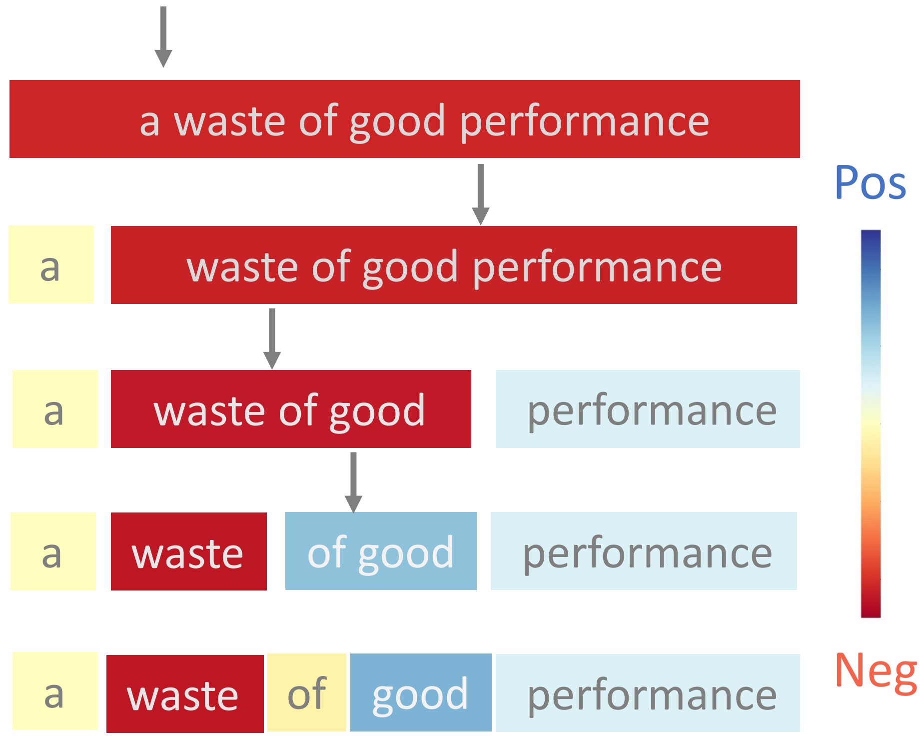
- Which text segment should be split?

The *global weakest* interaction point:

$$\min_{\boldsymbol{x}_{(s_i, s_{i+1}]} \in \mathcal{P}} \min_{j \in (s_i, s_{i+1})} \phi\left(\boldsymbol{x}_{(s_i, j]}, \boldsymbol{x}_{(j, s_{i+1}]} | \mathcal{P}\right)$$

- How to quantify feature importance?

Feature importance score: $\psi(\cdot)$



HEDGE

- Where is the dividing point?

The *local weakest* interaction point:

$$\min_{j\in(s_i,\,s_{i+1})} \phi\left(\boldsymbol{x}_{(s_i,j]}, \boldsymbol{x}_{(j,\,s_{i+1}]} | \mathcal{P}\right)$$

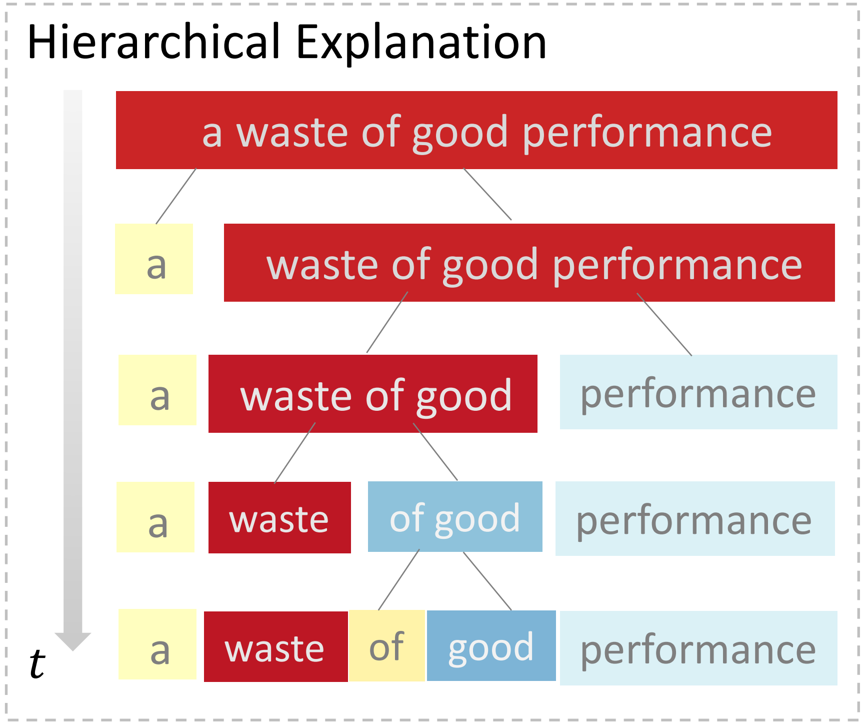
- Which text segment should be split?

The *global weakest* interaction point:

$$\min_{\boldsymbol{x}_{(s_i,\,s_{i+1}]}\in\mathcal{P}} \min_{j\in(s_i,\,s_{i+1})} \phi\left(\boldsymbol{x}_{(s_i,j]}, \boldsymbol{x}_{(j,\,s_{i+1}]} | \mathcal{P}\right)$$

- How to quantify feature importance?

Feature importance score: $\psi(\cdot)$



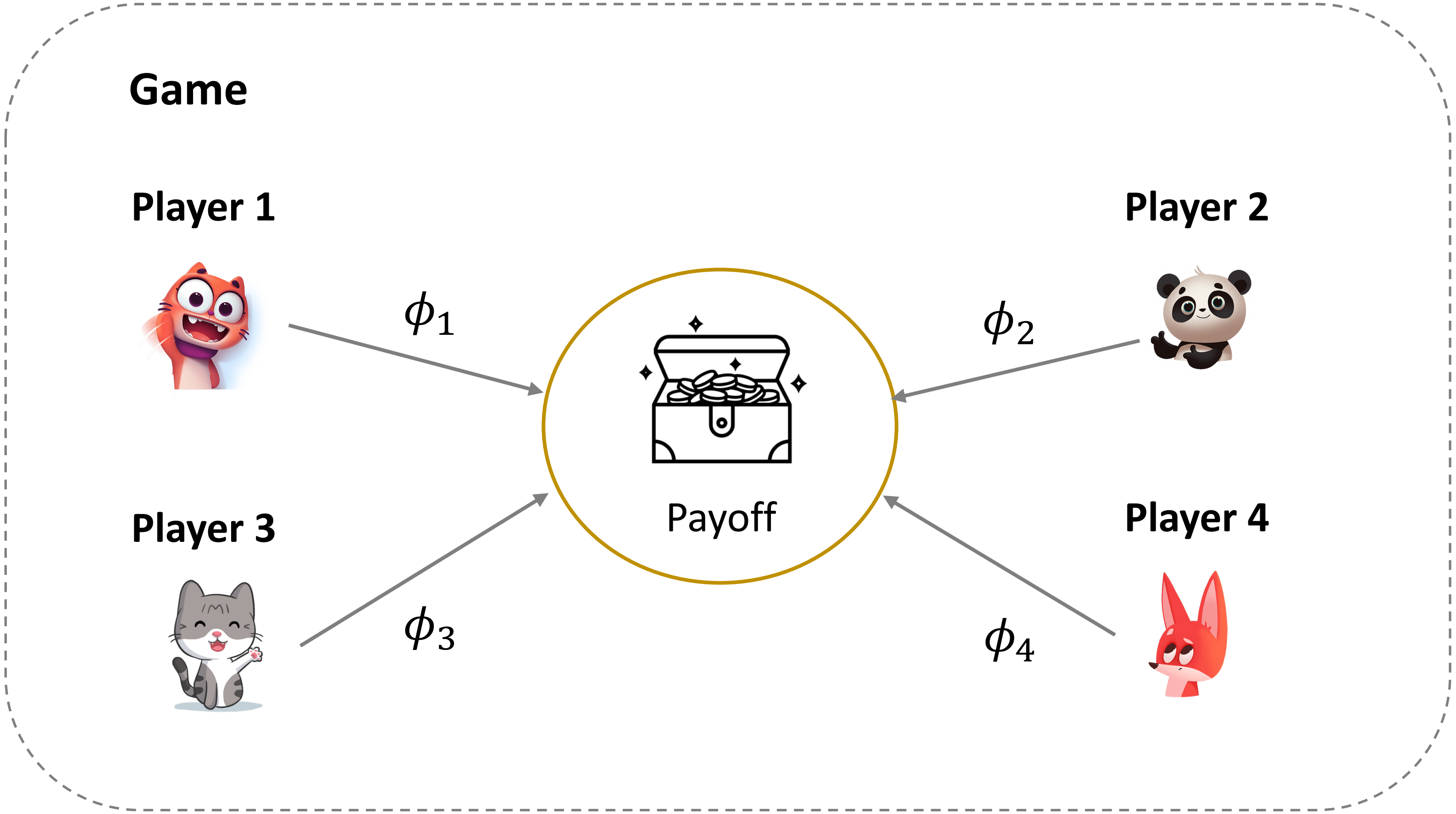
HEDGE

- Feature interaction score $\phi(\cdot,\cdot)$

Calculate the interaction between $\boldsymbol{j_1}$ and $\boldsymbol{j_2}$ via Shapley interaction index
[Fujimoto et al., 2006, Lundberg et al., 2018]

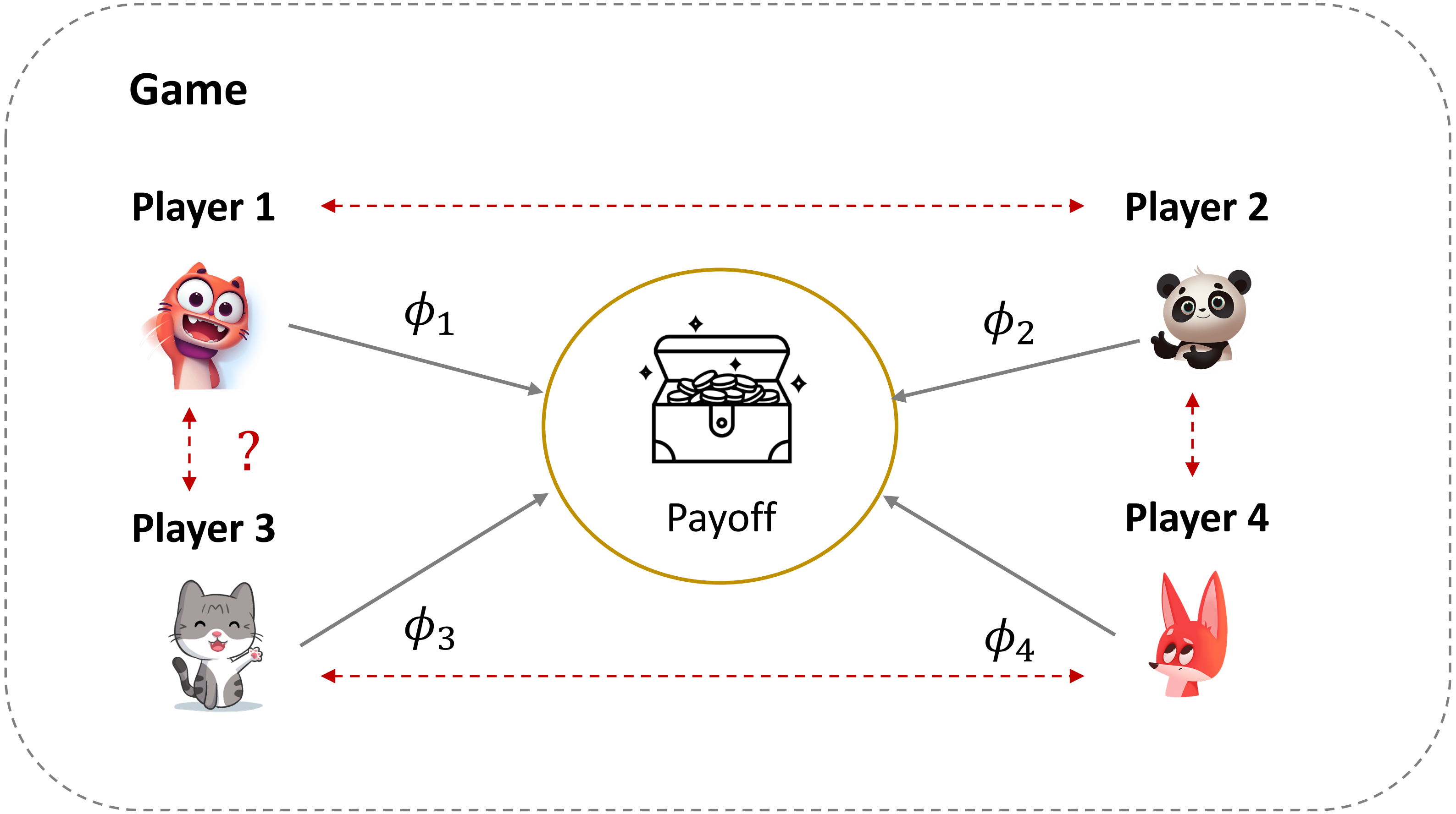
Shapley

Quantifying the contribution of each player



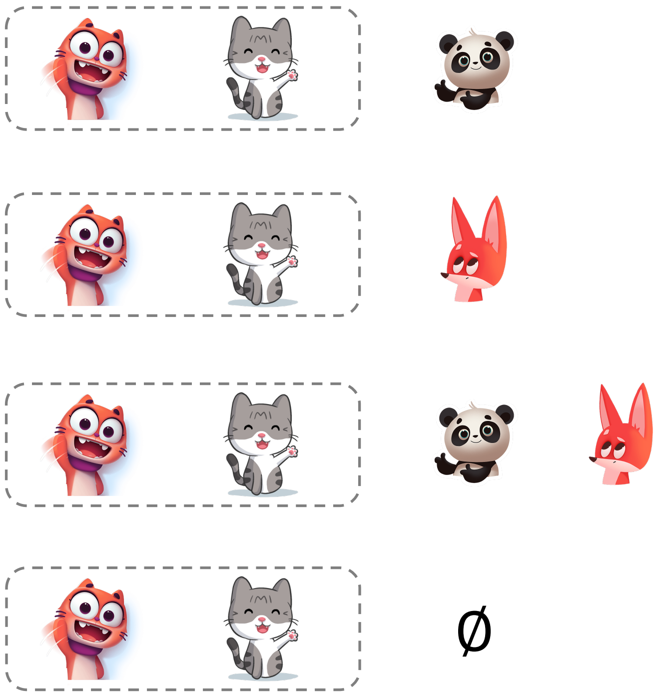
Shapley Interaction Index

Quantifying the interaction between players



Shapley Interaction Index

Coalitions **Payoff**



P_1

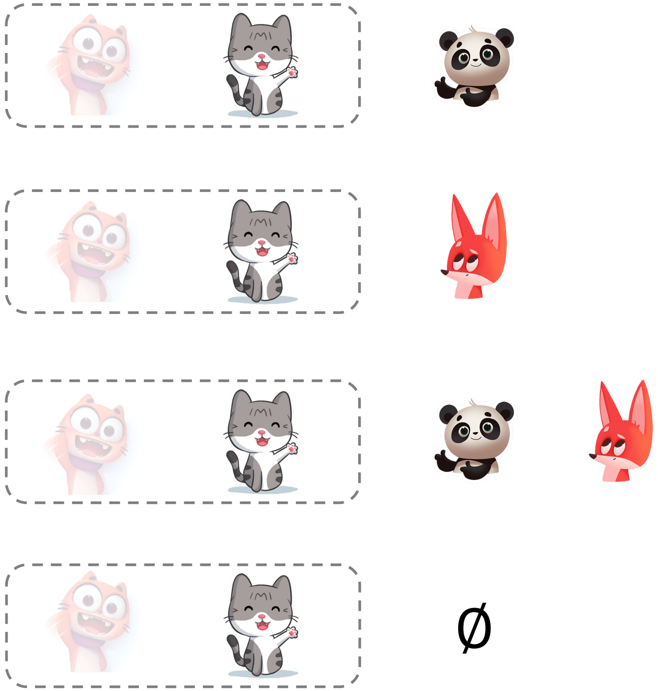
P_2

P_3

P_4

Shapley Interaction Index

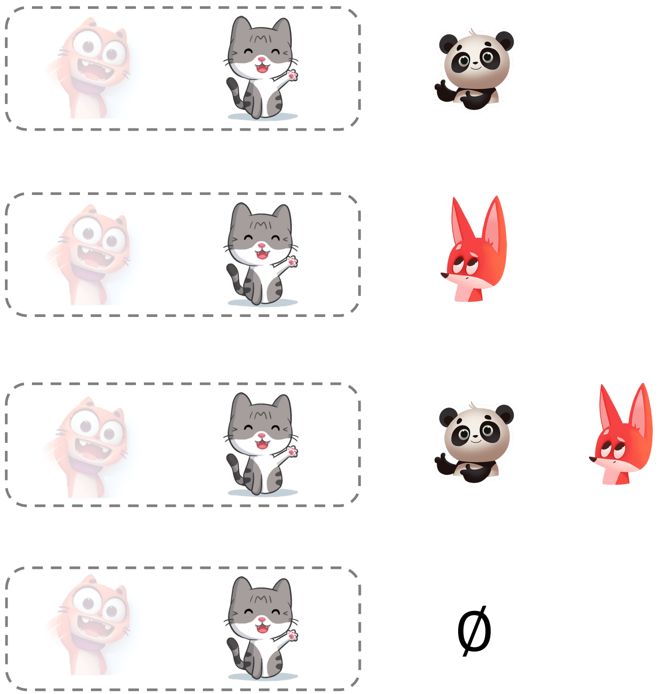
Coalitions **Payoff**



P_1	P_1'
P_2	P_2'
P_3	P_3'
P_4	P_4'

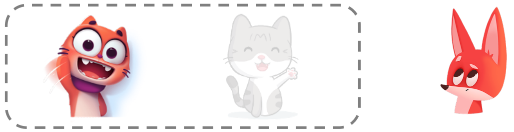
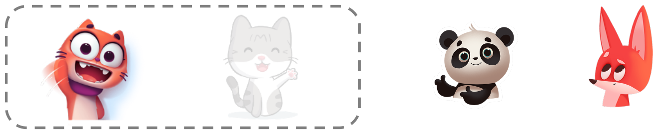
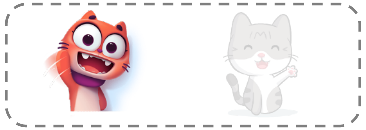
Shapley Interaction Index

Coalitions

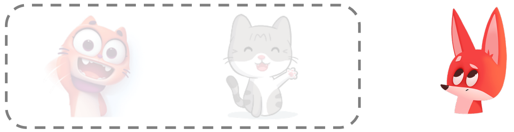
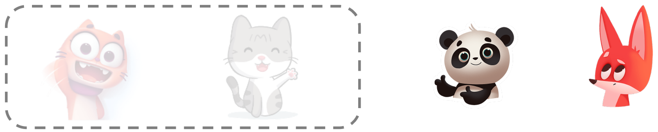
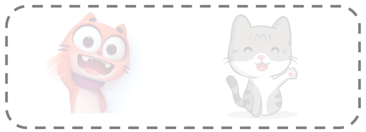


Payoff	Marginal contribution with player 3
$P_1 - P_1'$	ΔP_1
$P_2 - P_2'$	ΔP_2
$P_3 - P_3'$	ΔP_3
$P_4 - P_4'$	ΔP_4

Shapley Interaction Index

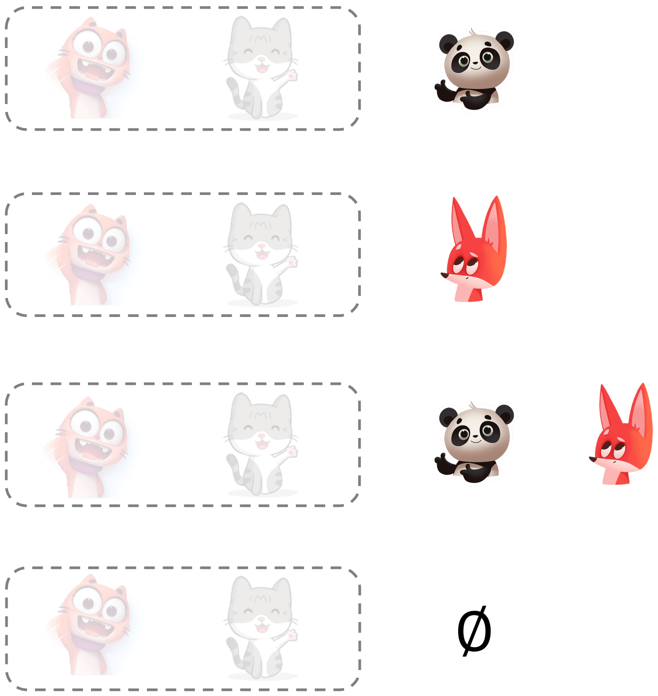
Coalitions	Payoff			Marginal contribution with player 3	Payoff
	P_1	$-$	$P_1{}'$	ΔP_1	Q_1
	P_2	$-$	$P_2{}'$	ΔP_2	Q_2
	P_3	$-$	$P_3{}'$	ΔP_3	Q_3
 $\emptyset$	P_4	$-$	$P_4{}'$	ΔP_4	Q_4

Shapley Interaction Index

Coalitions	Payoff	Marginal contribution with player 3	Payoff	Marginal contribution without player 3
	$P_1 - P_1'$	ΔP_1	$Q_1 - Q_1'$	ΔQ_1
	$P_2 - P_2'$	ΔP_2	$Q_2 - Q_2'$	ΔQ_2
	$P_3 - P_3'$	ΔP_3	$Q_3 - Q_3'$	ΔQ_3
 $\emptyset$	$P_4 - P_4'$	ΔP_4	$Q_4 - Q_4'$	ΔQ_4

Shapley Interaction Index

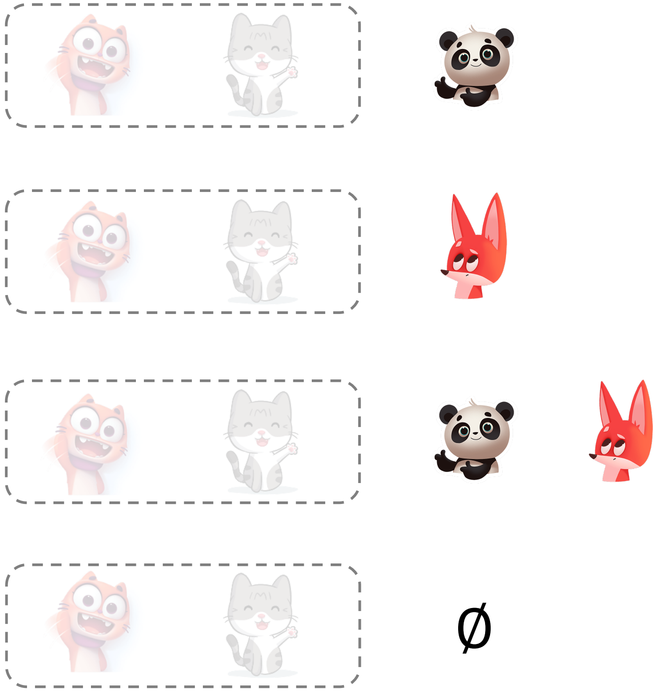
Coalitions



Payoff			Marginal contribution with player 3	Payoff			Marginal contribution **without** player 3
P_1	$-$	$P_1{}'$	ΔP_1	Q_1	$-$	$Q_1{}'$	ΔQ_1
P_2	$-$	$P_2{}'$	ΔP_2	Q_2	$-$	$Q_2{}'$	ΔQ_2
P_3	$-$	$P_3{}'$	ΔP_3	Q_3	$-$	$Q_3{}'$	ΔQ_3
P_4	$-$	$P_4{}'$	ΔP_4	Q_4	$-$	$Q_4{}'$	ΔQ_4

$$\phi_{1,3} = \sum \Delta P_i - \Delta Q_i$$

Shapley Interaction Index



Coalitions	Payoff	Marginal contribution with player 3	Payoff	Marginal contribution without player 3
	$P_1 - P_1'$	ΔP_1	$Q_1 - Q_1'$	ΔQ_1
	$P_2 - P_2'$	ΔP_2	$Q_2 - Q_2'$	ΔQ_2
	$P_3 - P_3'$	ΔP_3	$Q_3 - Q_3'$	ΔQ_3
$\emptyset$	$P_4 - P_4'$	ΔP_4	$Q_4 - Q_4'$	ΔQ_4

$$\phi_{1,3} = \sum \Delta P_i - \Delta Q_i$$

$$\phi_{3,1} = \phi_{1,3}$$

interaction

$$\phi_{3,1} + \phi_{1,3}$$

HEDGE

- Feature interaction score $\phi(\cdot,\cdot)$

Calculate the interaction between $\boldsymbol{j_1}$ and $\boldsymbol{j_2}$ via Shapley interaction index
[Fujimoto et al., 2006, Lundberg et al., 2018]

$$\phi(j_1, j_2|\mathcal{P}) = \sum_{S \subseteq \mathcal{N} \setminus \{j_1, j_2\}} \frac{|S|!\,(P - |S| - 1)!}{P!} \gamma(j_1, j_2, S)$$

$$\gamma(j_1, j_2, S) = \underbrace{\mathbb{E}[f(\boldsymbol{x}')|S \cup \{j_1, j_2\}] - \mathbb{E}[f(\boldsymbol{x}')|S \cup \{j_2\}]}_{\text{The influence of } j_1 \text{ on the model output with } j_2 \text{ considered}} - (\underbrace{\mathbb{E}[f(\boldsymbol{x}')|S \cup \{j_1\}] - \mathbb{E}[f(\boldsymbol{x}')|S]}_{\text{without } j_2 \text{ considered}})$$

HEDGE

- Feature interaction score $\phi(\cdot,\cdot)$

Calculate the interaction between $\boldsymbol{j_1}$ and $\boldsymbol{j_2}$ via Shapley interaction index
[Fujimoto et al., 2006, Lundberg et al., 2018]

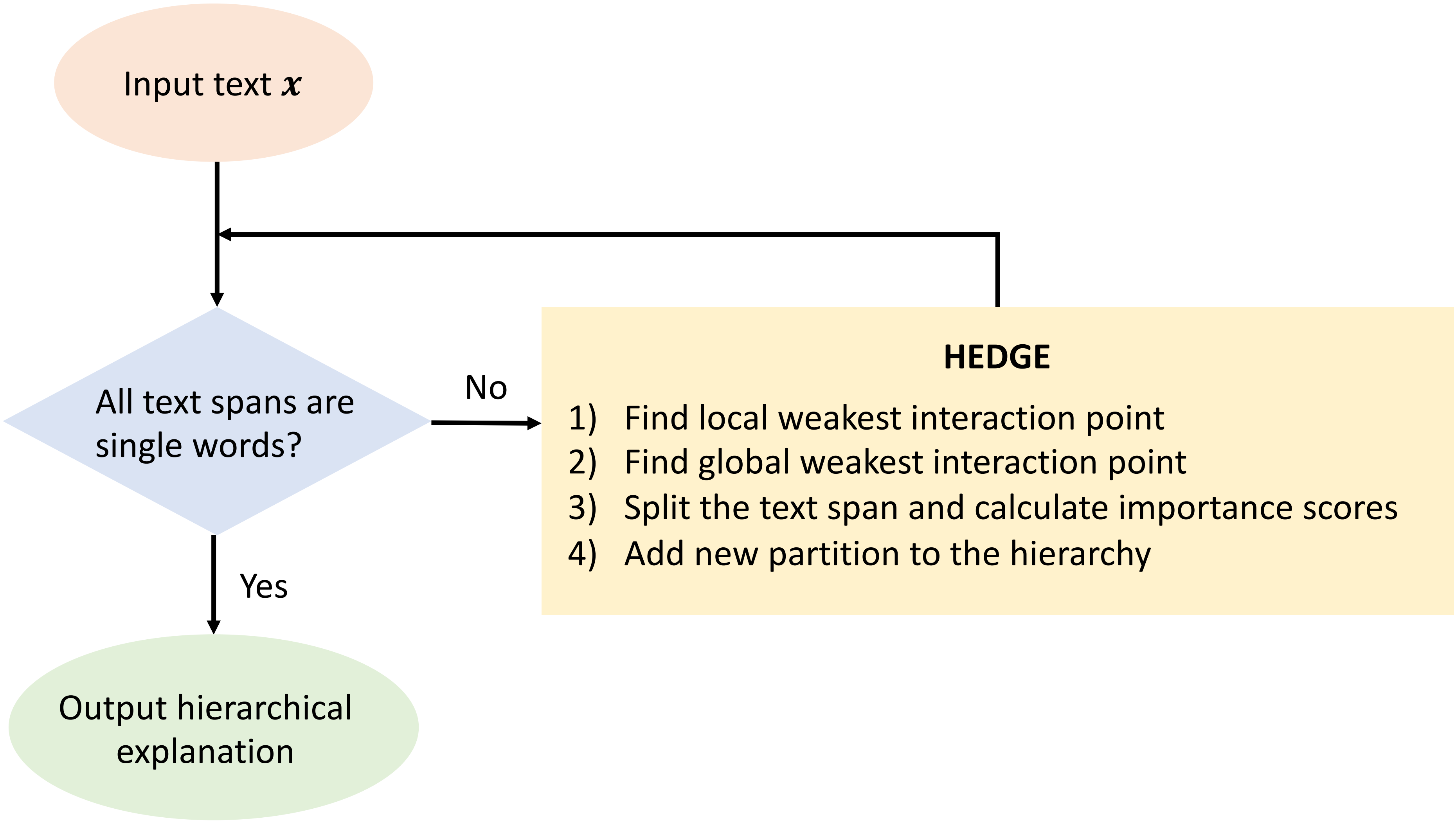
$$\phi(j_1, j_2|\mathcal{P}) = \sum_{S \subseteq \mathcal{N} \setminus \{j_1, j_2\}} \frac{|S|!\,(P - |S| - 1)!}{P!} \gamma(j_1, j_2, S)$$

$$\gamma(j_1, j_2, S) = \underbrace{\mathbb{E}[f(\boldsymbol{x}')|S \cup \{j_1, j_2\}] - \mathbb{E}[f(\boldsymbol{x}')|S \cup \{j_2\}]}_{\text{The influence of } j_1 \text{ on the model output with } j_2 \text{ considered}} - (\underbrace{\mathbb{E}[f(\boldsymbol{x}')|S \cup \{j_1\}] - \mathbb{E}[f(\boldsymbol{x}')|S]}_{\text{without } j_2 \text{ considered}})$$

- Feature importance score $\psi(\cdot,\cdot)$

$$\psi(\boldsymbol{x}_{(s_i,\, s_{i+1}]}) = f_{\underbrace{\hat{y}}_{\text{Predicted label on } \boldsymbol{x}}}(\boldsymbol{x}_{(s_i,\, s_{i+1}]}) - \max_{y' \neq \hat{y}, y' \in \mathcal{Y}} f_{y'}(\boldsymbol{x}_{(s_i,\, s_{i+1}]})$$

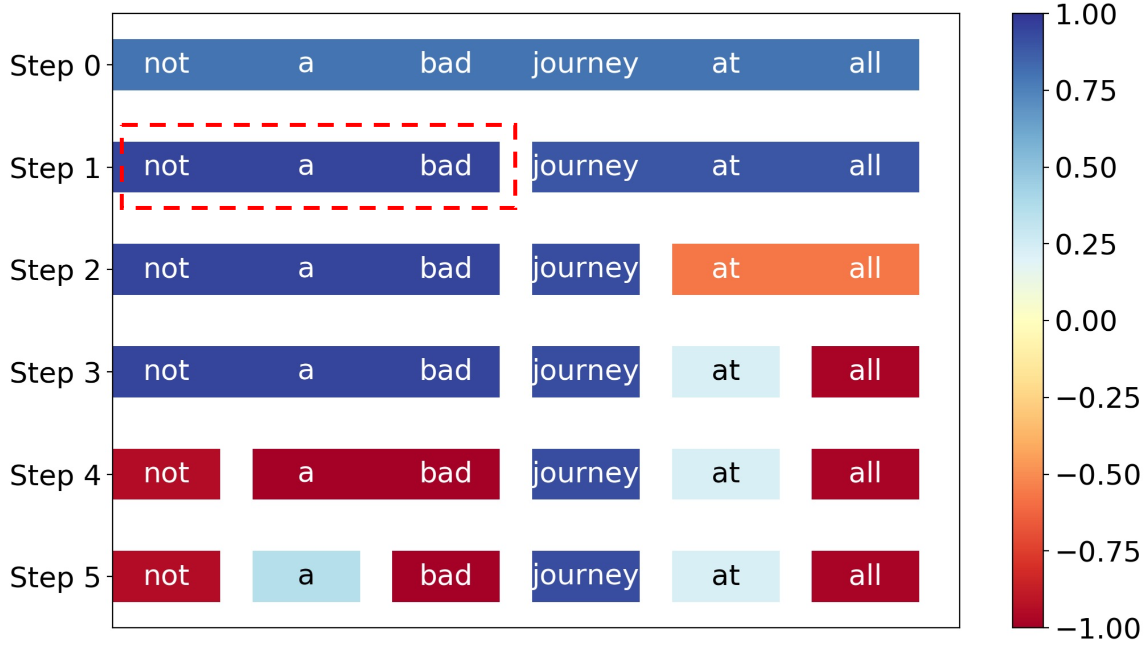
Pipeline


Input text $\boldsymbol{x}$
All text spans are single words?
No
Yes
HEDGE
1) Find local weakest interaction point
2) Find global weakest interaction point
3) Split the text span and calculate importance scores
4) Add new partition to the hierarchy
Output hierarchical explanation

Question?

Qualitative Analysis

- Compare HEDGE in interpreting the LSTM and BERT model
  - BERT gives the correct prediction “positive”, while LSTM makes a wrong prediction “negative”
  - HEDGE can explain different model prediction behaviors



(BERT)

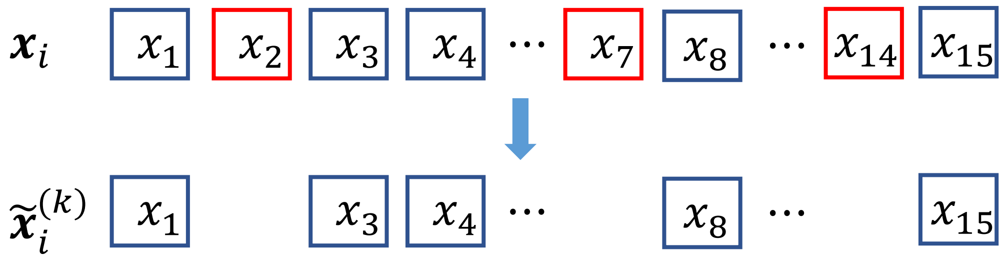
(LSTM)

Quantitative Evaluation

- The area over the perturbation curve (AOPC) [Nguyen, 2018, Samek et al., 2016]

$$AOPC(k) = \frac{1}{N}\sum_{i=1}^{N}\left\{p(\hat{y}|\boldsymbol{x}_i) - p\left(\hat{y}\,\middle|\,\widetilde{\boldsymbol{x}}_i^{(k)}\right)\right\}$$

✓ Higher AOPCs are better

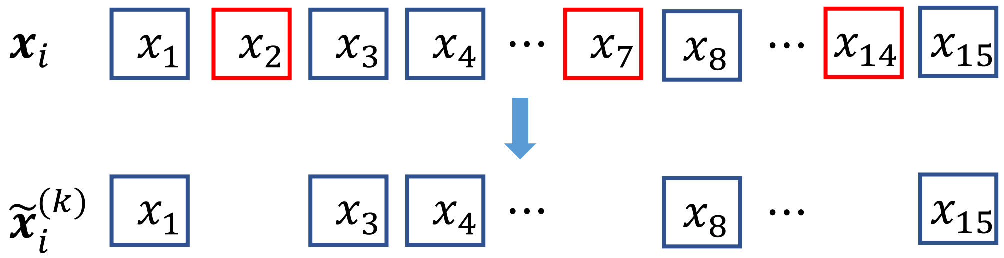


Quantitative Evaluation

- The area over the perturbation curve (AOPC) [Nguyen, 2018, Samek et al., 2016]

$$AOPC(k) = \frac{1}{N}\sum_{i=1}^{N}\left\{p(\hat{y}|\boldsymbol{x}_i) - p\left(\hat{y}\,\middle|\,\tilde{\boldsymbol{x}}_i^{(k)}\right)\right\}$$

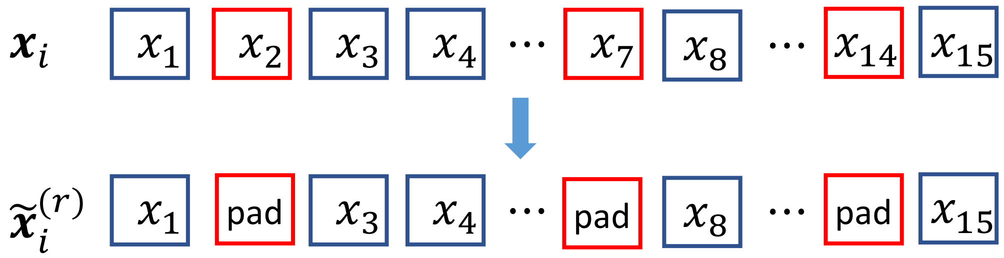
✓ Higher AOPCs are better



- Log-odds [Shrikumar et al., 2017, Chen et al., 2018]

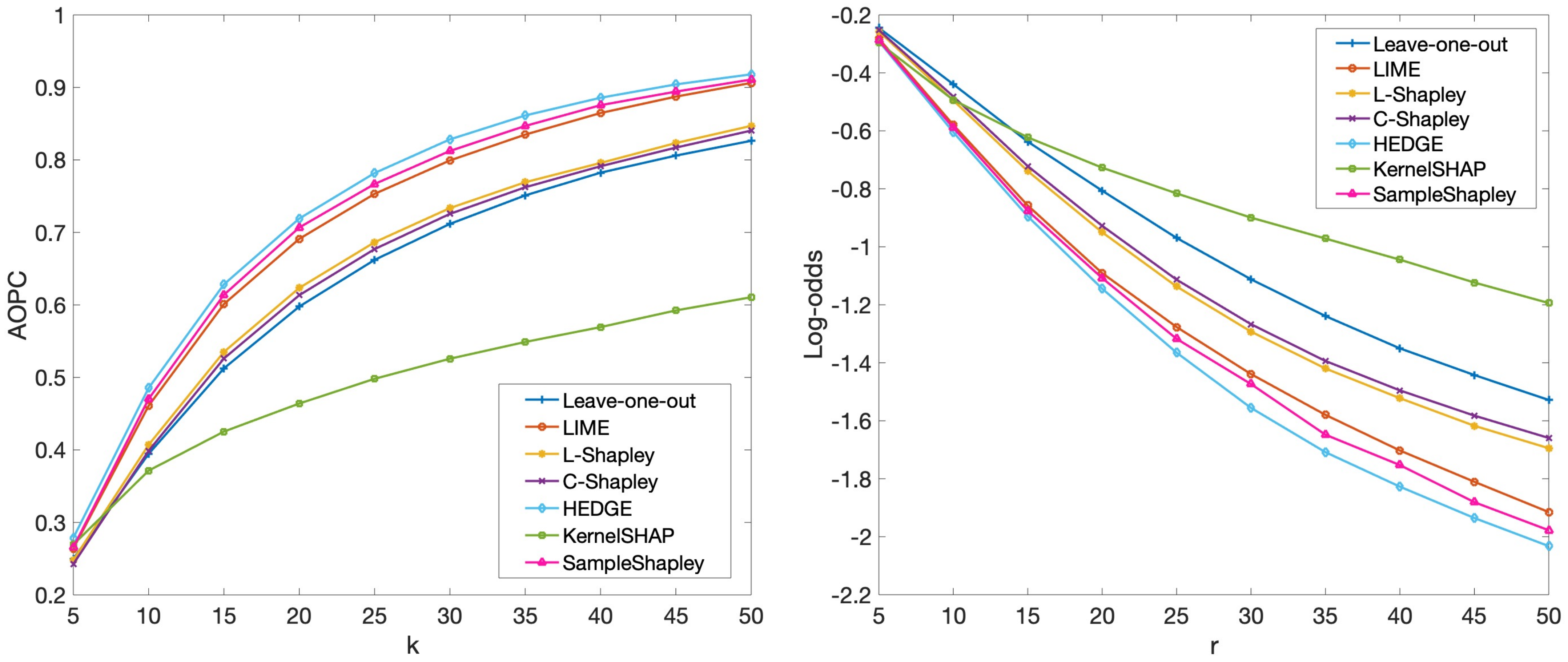
$$Log - odds(r) = \frac{1}{N}\sum_{i=1}^{N}\log\frac{p\left(\hat{y}\,\middle|\,\tilde{\boldsymbol{x}}_i^{(r)}\right)}{p(\hat{y}|\boldsymbol{x}_i)}$$

✓ Lower log-odds scores are better



Quantitative Evaluation

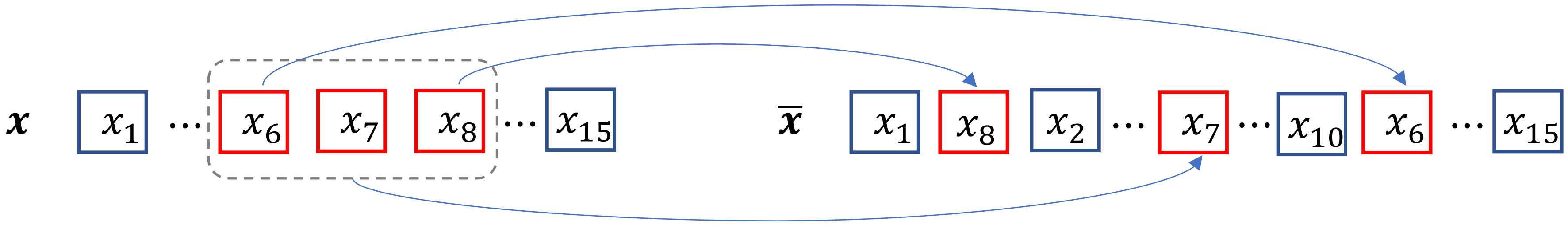
- AOPC and log-odds scores of the CNN model on the IMDB dataset
  - HEDGE achieves the best performance under both evaluation metrics



Quantitative Evaluation

- Cohesion-score
  - $Cohesion-score = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}\sum_{q=1}^{Q}\left\{p\left(\hat{y}\middle|\boldsymbol{x}_i\right) - p\left(\hat{y}\middle|\bar{\boldsymbol{x}}_i^{(q)}\right)\right\}$

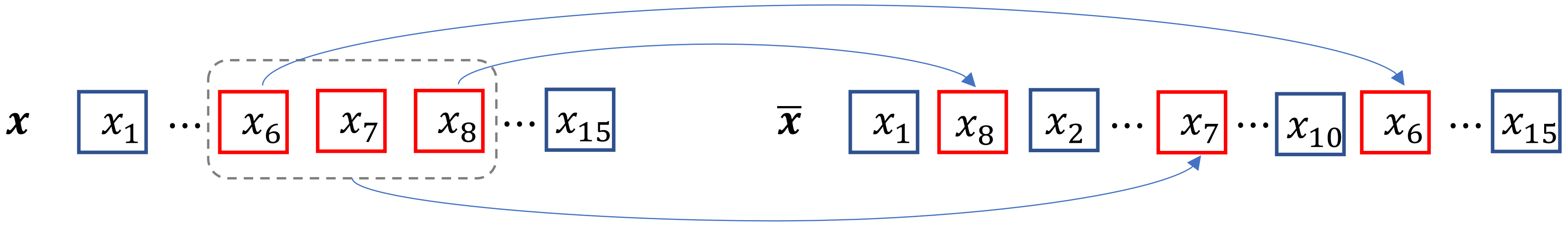
✓ Higher cohesion-scores are better



Quantitative Evaluation

- Cohesion-score
  - $Cohesion - score = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}\sum_{q=1}^{Q}\left\{p(\hat{y}|\boldsymbol{x}_i) - p\left(\hat{y}\,\middle|\,\overline{\boldsymbol{x}}_i^{(q)}\right)\right\}$

✓ Higher cohesion-scores are better



- Results

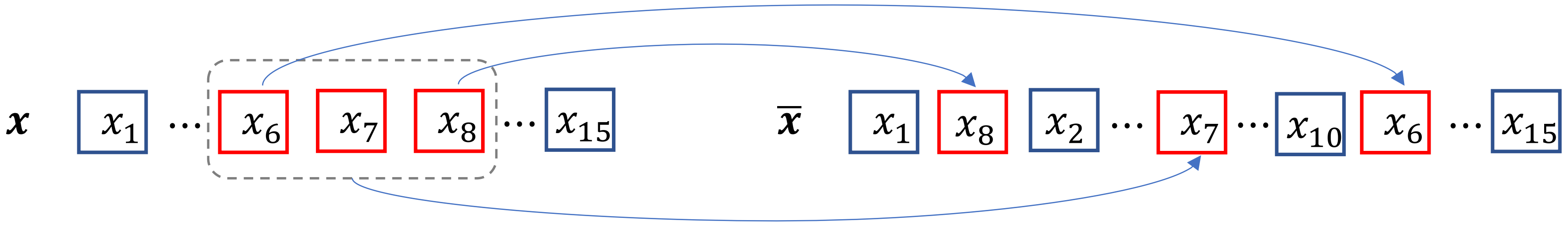
Methods	Models	Cohesion-score SST	Cohesion-score IMDB
HEDGE	CNN	0.016	0.012
	BERT	0.124	0.103
	LSTM	0.020	0.050
ACD	LSTM	0.015	0.038

✓ HEDGE is better at capturing feature interactions

Quantitative Evaluation

- Cohesion-score
  - $Cohesion-score = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}\sum_{q=1}^{Q}\left\{p(\hat{y}|\boldsymbol{x}_i) - p\left(\hat{y}\,\middle|\,\overline{\boldsymbol{x}}_i^{(q)}\right)\right\}$

✓ Higher cohesion-scores are better



- Results

Methods	Models	Cohesion-score	
		SST	IMDB
HEDGE	CNN	0.016	0.012
	BERT	0.124	0.103
	LSTM	0.020	0.050
ACD	LSTM	0.015	0.038

✓ BERT is more sensitive to perturbations on important phrases

Human Evaluation

- Compare human annotations and model predictions

Movie review

Not a bad movie at all

✓ Higher coherence scores are better

Explanation A

bad movie

Explanation B

not a bad

Guess predictions

A: negative

B: positive

Human Evaluation

- Coherence scores of different explanation methods with LSTM model on the IMDB dataset

Methods	Coherence Score
Leave-one-out	0.82
ACD	0.68
LIME	0.85
L-Shapley	0.75
C-Shapley	0.73
KernelSHAP	0.56
SampleShapley	0.78
HEDGE	**0.89**

Question?

Reference

- Murdoch, W. James, Peter J. Liu, and Bin Yu. "Beyond word importance: Contextual decomposition to extract interactions from lstms." *arXiv preprint arXiv:1801.05453* (2018).
- Chen, Hanjie, Guangtao Zheng, and Yangfeng Ji. "Generating hierarchical explanations on text classification via feature interaction detection." *arXiv preprint arXiv:2004.02015* (2020).
- Yeh, Chih-Kuan, et al. "On completeness-aware concept-based explanations in deep neural networks." *Advances in Neural Information Processing Systems* 33 (2020): 20554-20565.
- Sepp Hochreiter and J¨urgen Schmidhuber. Long short-term memory. Neural computation, 9(8): 1735–1780, 1997.
- Scott M Lundberg, Gabriel G Erion, and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.