# Explaining nonlinear classification decisions with deep Taylor decomposition

Pattern recognition
650 citation

Presentor: Jinduk park

# Contents

# Part1. Introduction to XAI

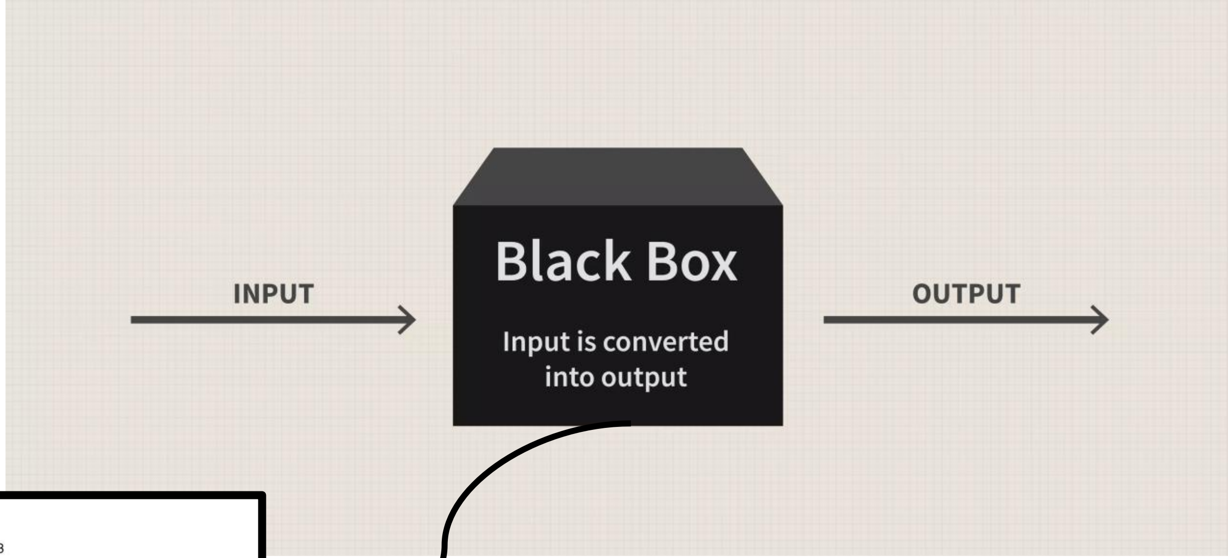
# What is explainable AI (XAI)?

Deep neural network (DNN) is successfully applied to many research area in terms of its performance.
(e.g. natural language processing, image classification, human action recognition..)


source:developer.nvidia.com/deep-learning-course

# What is explainable AI (XAI)?

However, because of its complex relations of nonlinearity,
DNN is regarded as a *black box model*, which means we don't know its internal working.




Source: https://www.investopedia.com/terms/b/blackbox.asp

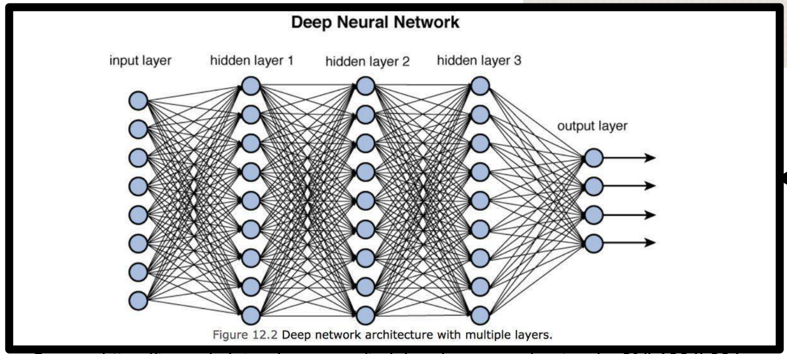


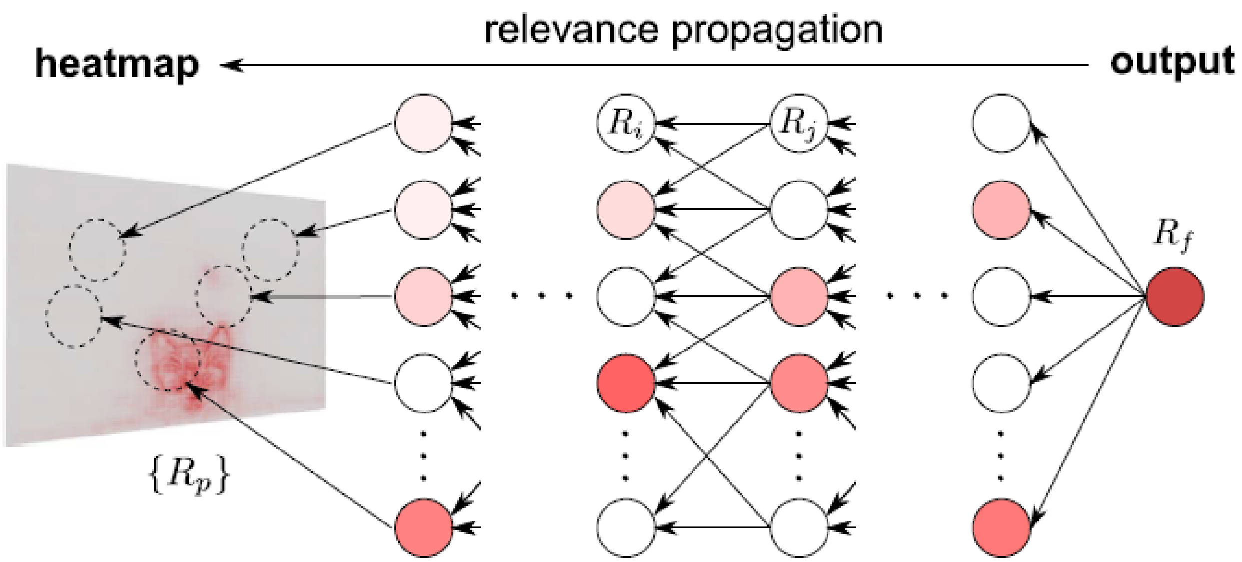

Figure 12.2 Deep network architecture with multiple layers.

Source: https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964

This property makes unfaithfulness of the model prediction.
(which is crucial for medical, bio, law domain)

# What is explainable AI (XAI)?

Explainable AI (XAI) explains decisions of a machine learning model, usually in terms of input variables.

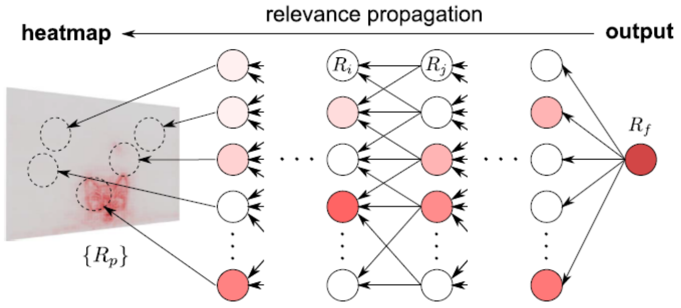



Such a explanation gives users and engineers credibility by providing transparency to the model.

# Taxonomy of XAI

Whether XAI model concerns inside of the model or not

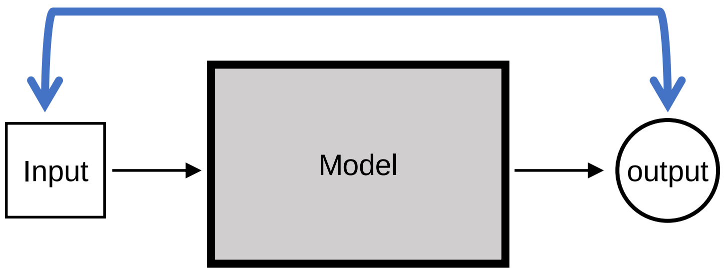
*Model-transparent* method




highlight which particular input features triggered key activations within a model’s weights.

SA [1], GradCAM [2], LRP [3], DTD.

*Model-agnostic* method




treat the model as totally a black-box and attempt to approximate the relationship between the input and the output prediction.

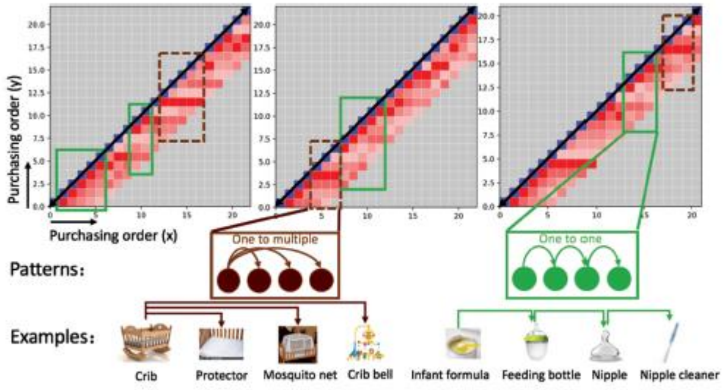
LIME [4], SHAP [5], Anchor [6].

Jeyakumar, Jeya Vikranth, et al. "How can i explain this to you? an empirical study of deep neural network explanation methods." *Advances in Neural Information Processing Systems* (NIPS) (2020).

# Taxonomy of XAI

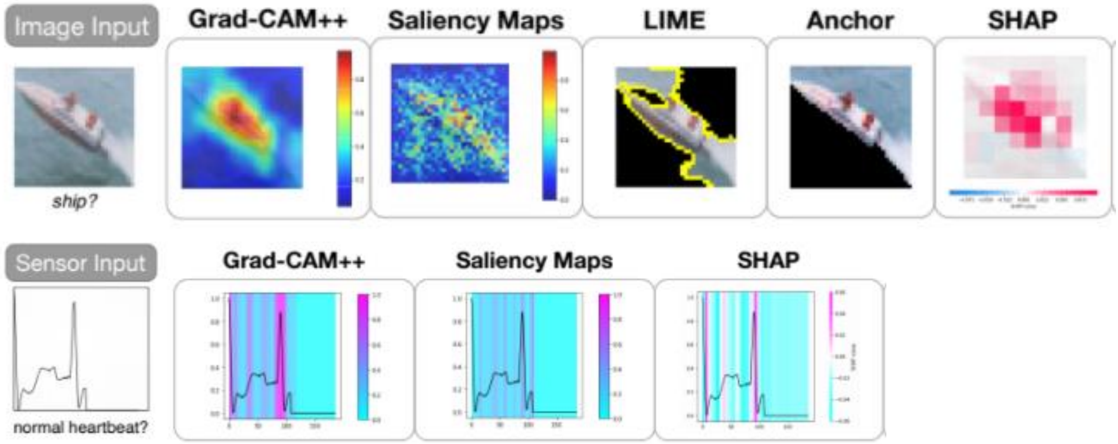
How to provide model explanation to user
varies depending on specific domain and XAI method.

[Behavior sequence]




Source: Zhang, Yongfeng, and Xu Chen. "Explainable recommendation: A survey and new perspectives.“ *arXiv preprint arXiv:1804.11192* (2018).

[Heatmap visualization]




Source: Jeyakumar, Jeya Vikranth, et al. "How can i explain this to you? an empirical study of deep neural network explanation methods.“ *Advances in Neural Information Processing Systems* (NIPS) (2020).

[Word cloud / sentence generation]


Source: Zhang, Yongfeng, and Xu Chen. "Explainable recommendation: A survey and new perspectives.“ *arXiv preprint arXiv:1804.11192* (2018).

## Part2. Deep Taylor decomposition (DTD)

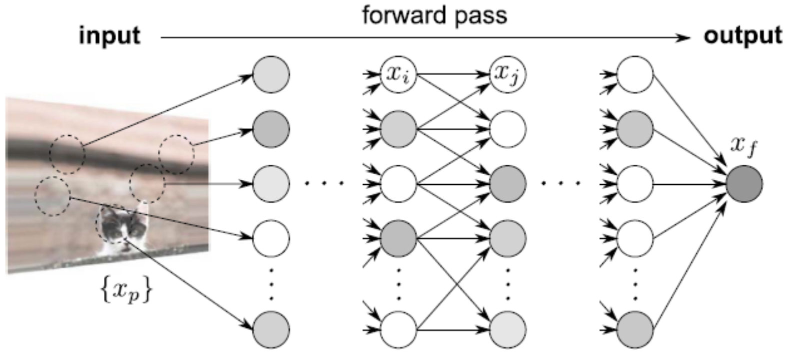
One of the *Model-transparent* XAI method that provides *heatmap* when explain model prediction.
(specifically assume image classification task)

# Definition of relevance score

- Basic concept & notations

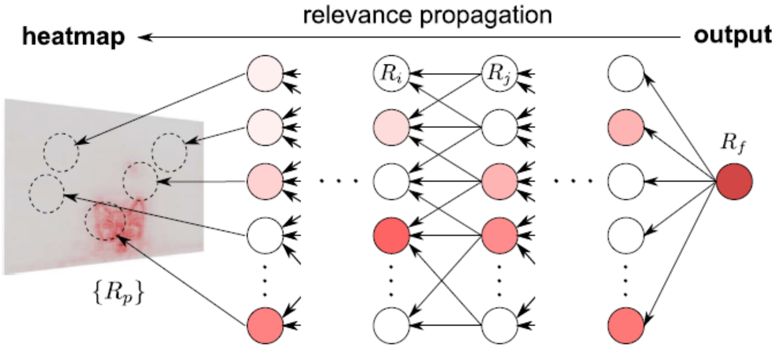
1. Model $f(x)$

   $f: \mathbb{R}^d \rightarrow \mathbb{R}^+$ : Quantifies the presence of certain object.




2. Relevance score $R_p(x)$

   To what extent the pixel $p$ contributes to explaining the classification decision in f(x).




3. Heat mapping $\boldsymbol{R}(x)$

   Contains set of relevance scores designated to each pixels.

# Definition of relevance score

- Required properties for the relevance score

* These are not sufficient conditions nor strict definition of relevance score though.

Definition 1. conservative

$$\forall\, \boldsymbol{x}: f(\boldsymbol{x}) = \sum_p R_p(\boldsymbol{x}).$$

Definition 2. positive

$$\forall\, \boldsymbol{x},\, p: R_p(\boldsymbol{x}) \geq 0$$

Definition 3. consistent

- Conservative + positive = consistent
- If the relevance score is consistent, than it is naturally forced to follow:

$$(f(\boldsymbol{x}) = 0) \Rightarrow (\boldsymbol{R}(\boldsymbol{x}) = 0)$$

# Motivation to use decomposition-based method

- Natural decomposition vs SA [1]

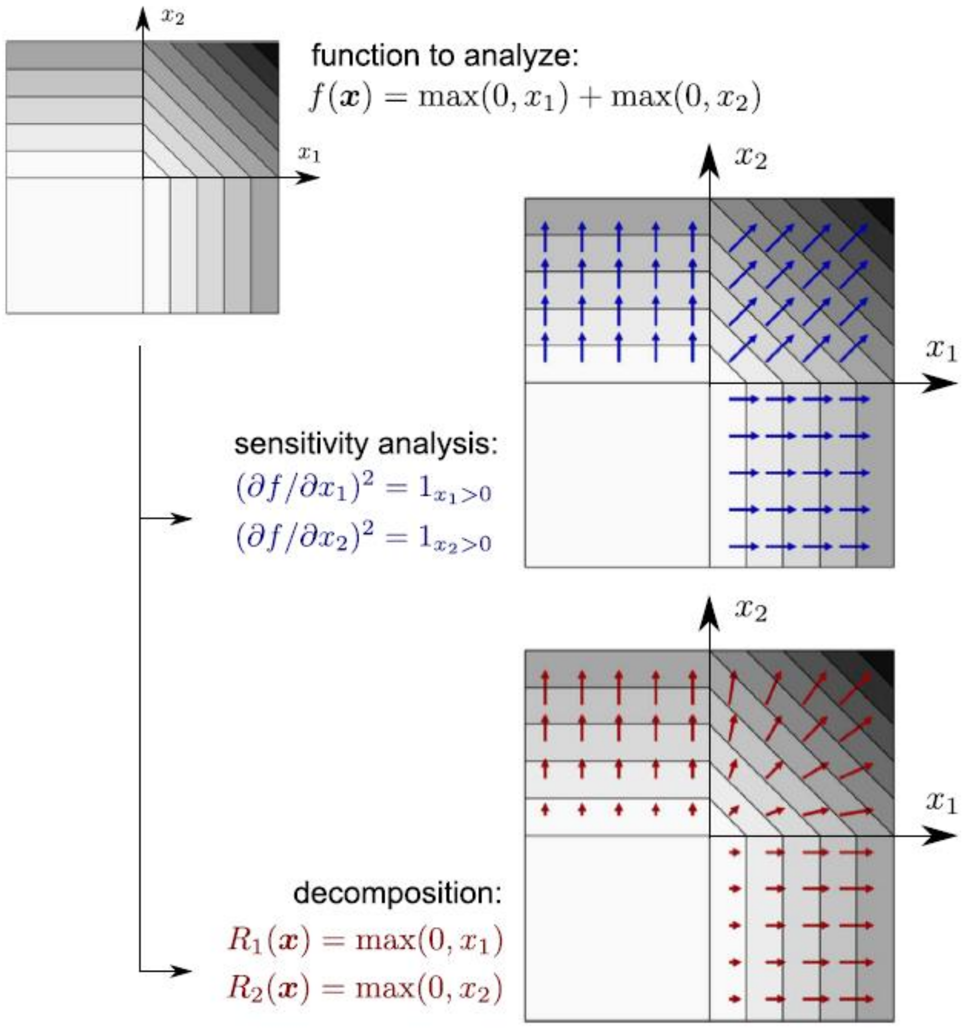
Natural decomposition method that follows definition 1~3.

$$f(\boldsymbol{x}) = \sum_p \sigma_p(x_p)$$

$$R_p(\boldsymbol{x}) = \sigma_p(x_p)$$

Even naïve decomposition method has more expressive power than SA method.

However, It is still not enough.



# Taylor decomposition

A decomposition method based on the Taylor expansion of the function at some well-chosen root point $\widetilde{\boldsymbol{x}}$

root point $\widetilde{\boldsymbol{x}}$ where $f(\widetilde{\boldsymbol{x}}) = 0$

The Taylor expansion gives:

$$f(\boldsymbol{x}) = \boxed{f(\widetilde{\boldsymbol{x}})} + \left( \frac{\partial f}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}} \right)^{\top} \cdot (\boldsymbol{x} - \widetilde{\boldsymbol{x}}) + \varepsilon$$

$$= \boxed{0} + \sum_p \underbrace{\frac{\partial f}{\partial x_p}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}} \cdot (x_p - \widetilde{x}_p)}_{\boxed{R_p(\boldsymbol{x})}} + \boxed{\varepsilon}$$

★★★

Higher order terms are complex and hard to redistribute

We takes 1-st order term as a relevance score $R_p(x)$

# Taylor decomposition

" What is the philosophy behind the formulation? "

$$f(\boldsymbol{x}) = \boxed{f(\widetilde{\boldsymbol{x}})} + \overbrace{\left(\frac{\partial f}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}}\right)^{\top} \cdot (\boldsymbol{x} - \widetilde{\boldsymbol{x}})}^{\boldsymbol{R}(\boldsymbol{x})} + \varepsilon$$

Transposing the term gives:

$$f(\boldsymbol{x}) - \overset{0}{f(\widetilde{\boldsymbol{x}})} = \boldsymbol{R}(\boldsymbol{x})$$

" From the absent of the object ($f(\widetilde{\boldsymbol{x}}) = 0$),
how much $\boldsymbol{x}$ contributes to the model classification f(x)."

# Taylor decomposition

" Then, how to find such a root point $\tilde{\boldsymbol{x}}$ that satisfying ($f(\tilde{\boldsymbol{x}}) = 0$)? "

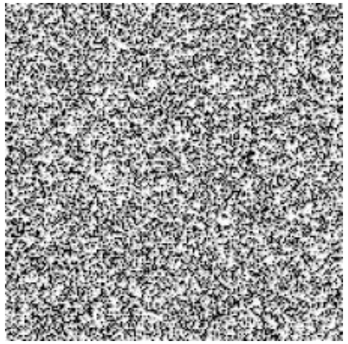
+ The root point should be admissible:
nearest in the Euclidean sense to the actual data point $\boldsymbol{x}$

$f(.)$ : white car classifier

$\boldsymbol{x}$

$\tilde{\boldsymbol{x}}$ (good !)

It will not exactly look like this though.. ☺



$\tilde{\boldsymbol{x}}$ (bad !)

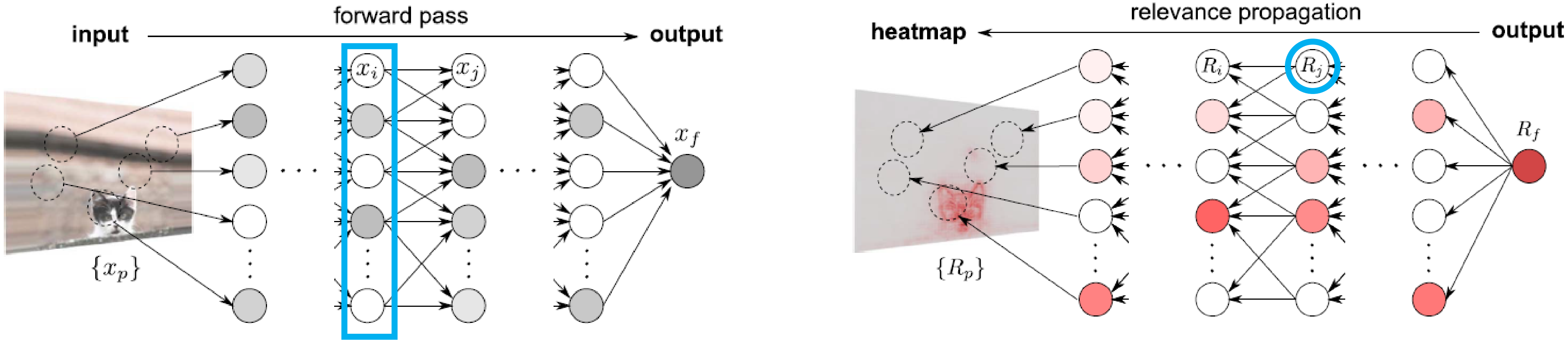
One possible way is to solve optimization problem with the minimization objective:

$$\min_{\boldsymbol{\xi}} \|\boldsymbol{\xi} - \boldsymbol{x}\|^2 \quad \text{subject to} \quad f(\boldsymbol{\xi}) = 0 \quad \text{and} \quad \boldsymbol{\xi} \in \mathcal{X},$$

But It is time-consuming and thus undesirable.

# Deep Taylor decomposition (DTD)

Deep Taylor decomposition, specifically designed Taylor-decomposition-based redistribution method to apply to deep neural network (DNN).




because of its top-down dependency property of DNN, we can always decompose $R_j$ in terms of previous layer's units $\{x_i\}$.

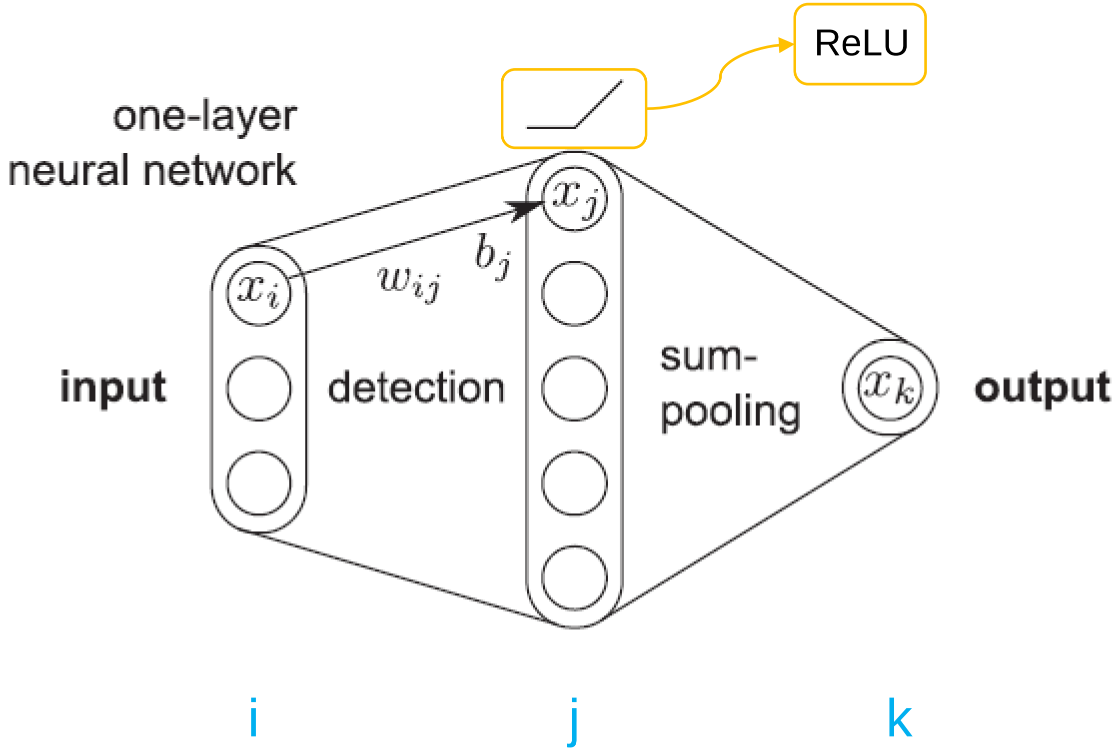
By Taylor expansion,

Taylor residual

$$R_j = \left( \frac{\partial R_j}{\partial \{x_i\}} \Big|_{\{\tilde{x}_i\}^{(j)}} \right)^{\top} \cdot \left( \{x_i\} - \{\tilde{x}_i\}^{(j)} \right) + \varepsilon_j = \sum_i \underbrace{\frac{\partial R_j}{\partial x_i} \Big|_{\{\tilde{x}_i\}^{(j)}} \cdot \left( x_i - \tilde{x}_i^{(j)} \right)}_{R_{ij}} + \boxed{\varepsilon_j}$$

# DTD to one-layer networks

Application DTD on one-layer network with ReLU non-linear activation.

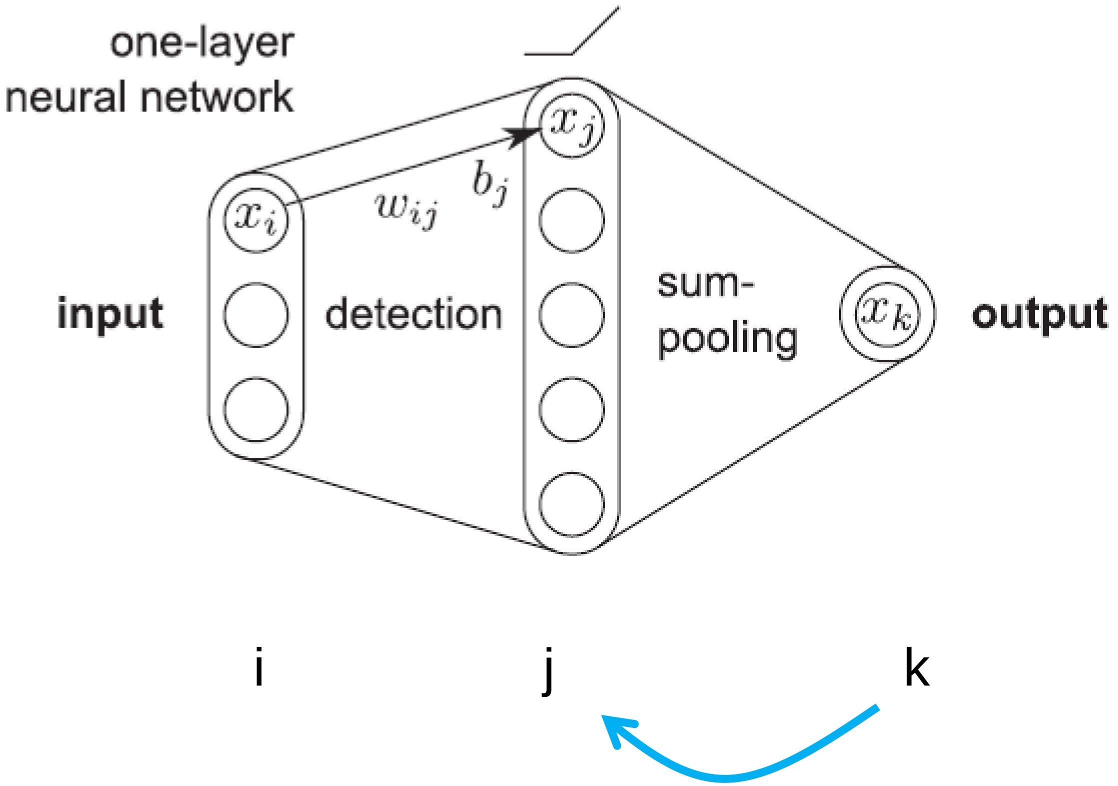



$$x_j = \max\left(0, \sum_i x_i w_{ij} + b_j\right) \quad \text{and} \quad x_k = \sum_j x_j$$

* Logic flow for deriving DTD method.
1. Find root point $\tilde{\boldsymbol{x}}$.
2. Based on the found root point, conduct Taylor decomposition w.r.t previous layer.
3. Derive relevance redistribution rule.

* Such root point($f(\tilde{\boldsymbol{x}}) = 0$) may not exist in some DNN:
Give constraint: $b_j \leq 0$ to ensure *existence* of the root point

# DTD to one-layer networks




- First rule for relevance redistribution.

$$R_k = \sum_j x_j \qquad (1)$$

$$R_j = \frac{\partial R_k}{\partial x_j}\Big|_{\{\tilde{x}_j\}} \cdot (x_j - \tilde{x}_j) \qquad (2)$$

Let's find root point $\{\tilde{\boldsymbol{x}}_j\}$.

1. From (1), $f(\tilde{x}_j) = \sum_j \tilde{x}_j = 0$
2. Admissible ($\forall j : \tilde{x}_j \geq 0$ ) : ReLU

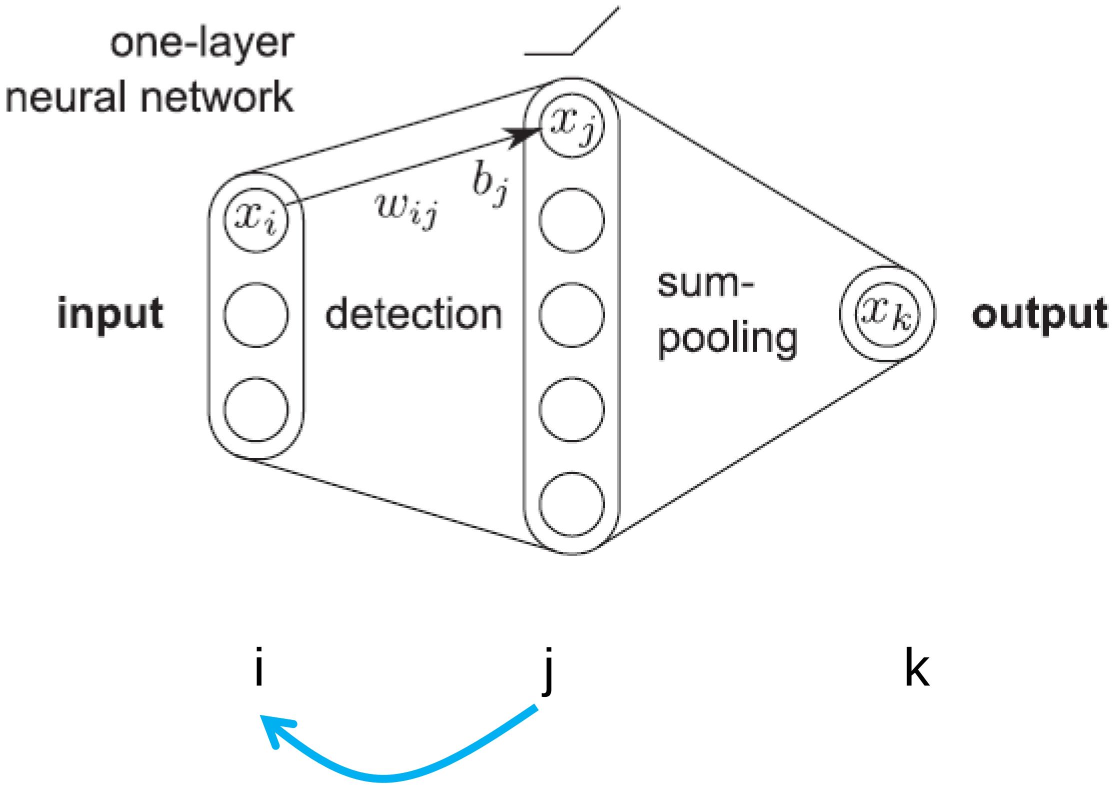
$\tilde{\boldsymbol{x}}_j = \mathbf{0}$

From (2), $\frac{\partial R_k}{\partial x_j} = 1$ , $\tilde{\boldsymbol{x}}_j = \mathbf{0}$ gives

first rule for relevance redistribution.

$$\boxed{R_j = x_j}$$

# DTD to one-layer networks




- Second rule for relevance redistribution.

$$R_j = \max\left(0, \sum_i x_i w_{ij} + b_j\right),$$ (from first rule)

Establish mapping between $\{x_i\}$ to $R_j$ by Taylor expansion

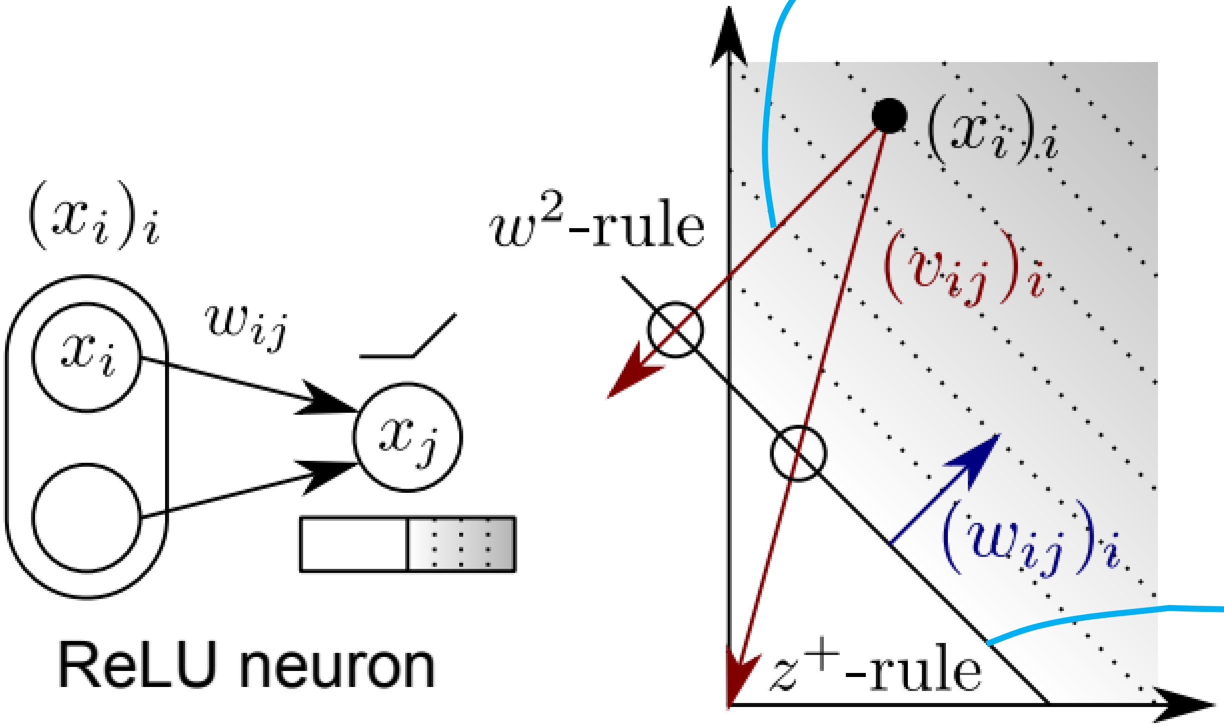
$$R_i = \sum_j \frac{\partial R_j}{\partial x_i}\Big|_{\{\widetilde{x}_i\}^{(j)}} \cdot \left(x_i - \widetilde{x}_i^{(j)}\right).$$

Let's find root point $\{\widetilde{x}_i\}^j$ again.

Here, each choice of input domain will lead to different rule for propagating relevance.

# DTD to one-layer networks

2. The root point should be admissible.




Bach, Sebastian, et al. "Deep Taylor Decomposition of Neural Networks."

1. root point condition: $f(\tilde{x}) = 0$

Illustration of a root point search in the two-dimensional input space of a ReLU neuron.

We solve the intersection of the two condition to find admissible root point.

# DTD to one-layer networks

## 1. $w^2$-rule

Unconstrained input search space

The admissible root point is the intersection of the plane equation $\sum_i \tilde{x}_i{}^j w_{ij} + b_j = 0$ and the line of maximum descent $\{\tilde{x}_i\}^{(j)} = \{x_j\} + t\boldsymbol{w}_j$

root point condition

Admissible : be in vicinity

$$\{\tilde{x}_i\}^{(j)} = \{x_j - \frac{w_{ij}}{\sum_i w_{ij}^2}(\sum_i x_i w_{ij} + b_j)\}$$

Inject this into the equation $R_i = \sum_j \frac{\partial R_j}{\partial x_i}\Big|_{\{\tilde{x}_i\}^{(j)}} \cdot (x_i - \tilde{x}_i^{(j)}).$

$$\boxed{R_i = \sum_j \frac{w_{ij}^2}{\sum_{i'} w_{i'j}^2} R_j}$$

# DTD to one-layer networks

## 2. $z$-rules

We can give some constraints, which leads to different propagation rules

**Positive input search space**

$$\{\{x_i\} = 0 \leq x_i\}$$

⬇

**$z^+$-rules**

$$R_i = \sum_j \frac{z_{ij}^+}{\sum_{i'} z_{i'j}^+} R_j$$

$$z_{ij}^+ = x_i w_{ij}^+$$

**Bounded input search space**

$$\{\{x_i\} = l_i \leq x_i \leq h_i\}$$

$l$: lower bound

$h$: upper bound

⬇

**$z^\beta$-rules**

$$R_i = \sum_j \frac{z_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_{i'} z_{i'j} - l_i w_{i'j}^+ - h_i w_{i'j}^-} R_j$$

$$z_{ij} = x_i w_{ij}$$

# DTD to deep networks

DNN is basically constructed with stacking such a simple layer by layer,
so we can serially apply redistribution rule for different layer considering it activation and constraint.

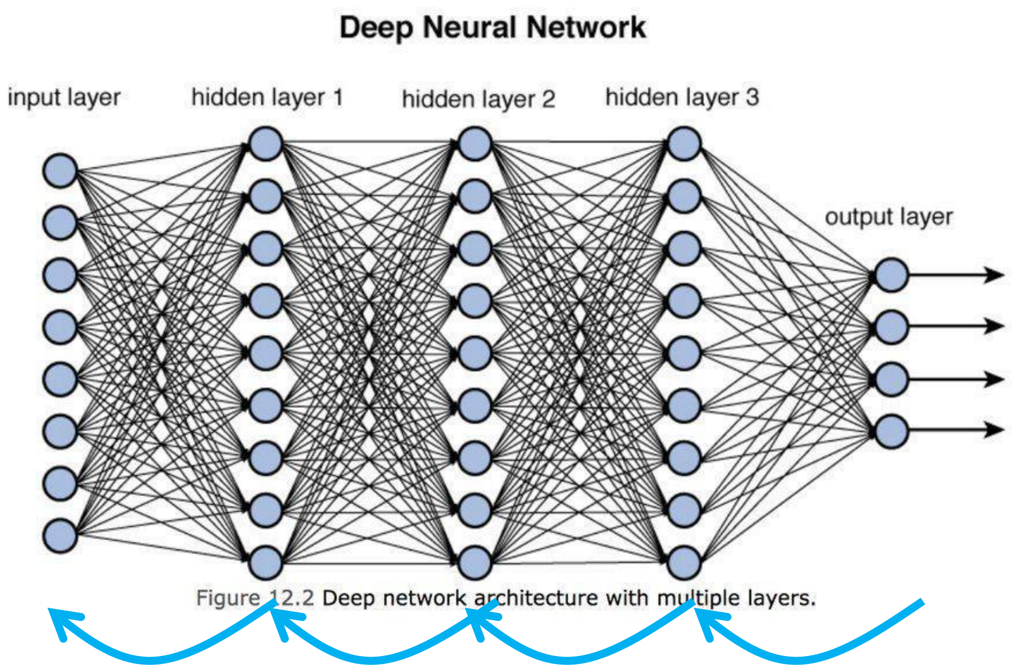


Figure 12.2 Deep network architecture with multiple layers.

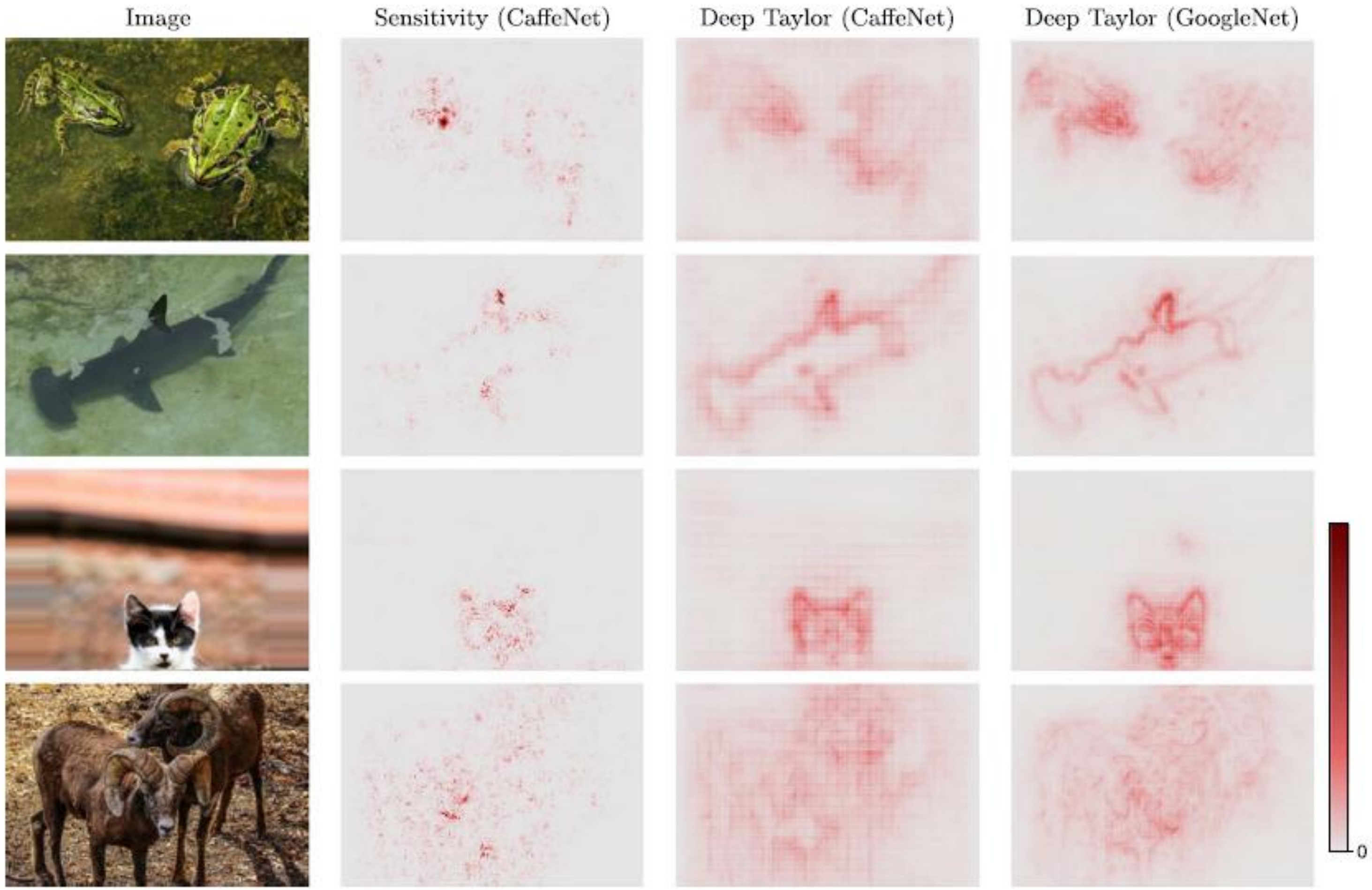
$z^{\beta}$-rule $w^{2}$-rule $z^{+}$-rule

Source: https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964

Stacking redistribution rules makes *training-free* decomposition-based XAI method

# DTD to deep networks

The results (heatmap visualization)




Qualitative evaluation of the method

* There is no ground-truth for the explanation.

# Wrap up

# Consideration of limitations

1. Deriving every relevance redistribution rules for different conditions might be expensive.

2. Plus, It seems to be difficult to apply DTD to the more complex network architecture like LSTM cell.

3. We could infer that explanation error occur because of some assumptions and Taylor residuals in the process of rule derivation.

# Summary and conclusion

- DTD is theoretically well-established model-transparent XAI method that can be applicable for DNN architecture.

- The redistribution rules in DNN are vary depending on the model and data constraints

(Next talk)

- Introduction to some model-agnostic XAI methods

- How to quantitively evaluate XAI model instead of visual one.

# References

[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 839–847. IEEE, 2018.

[3] Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." *PloS one* 10.7 (2015): e0130140.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Advances in neural information processing systems, pages 4765–4774, 2017.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In AAAI Conference on Artificial Intelligence (AAAI), 2018.

Thanks for your listening.