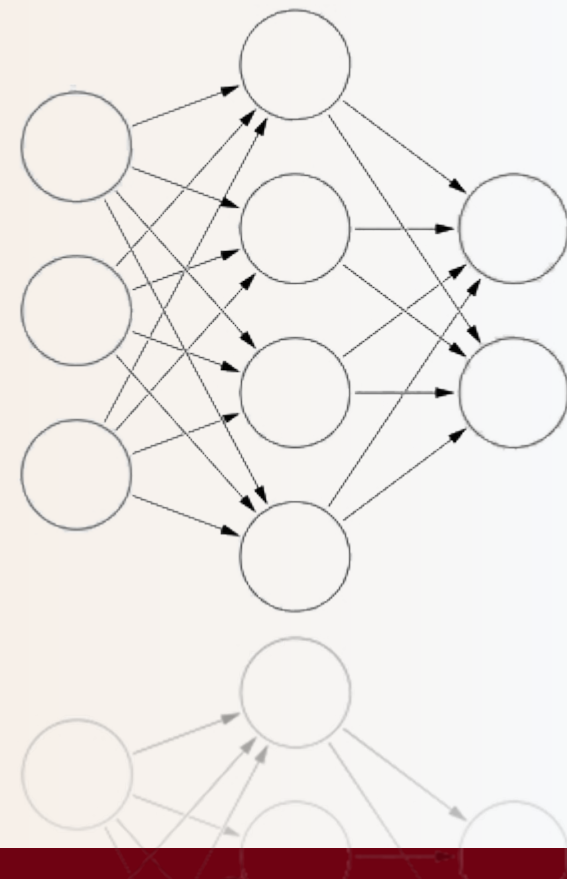


Visualizing classification decision of neural networks (The Way How machine thinks)



Korea University,
2016010662 윤주성
2017020764 김도완
2017010607 김우성

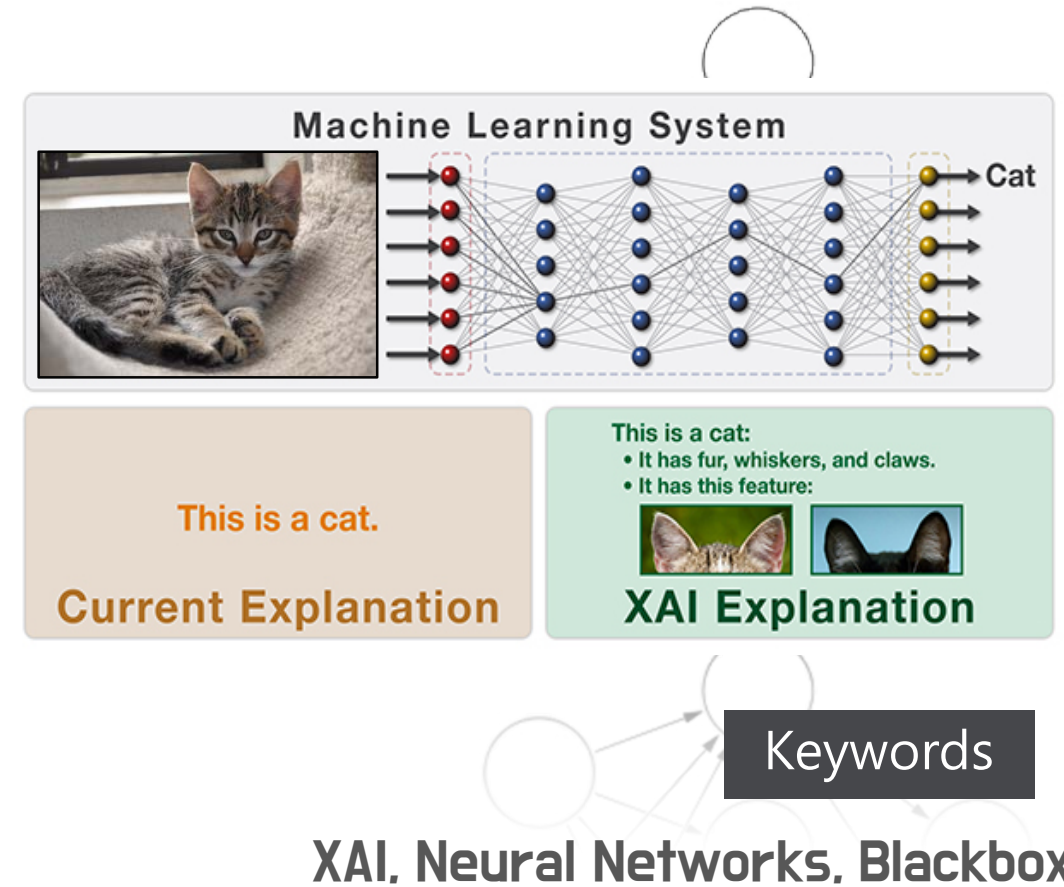
Intro

Deep Learning

High performance / Low explanation

Medicine / Finance / etc

Needs for explainable AI are increasing



Approach - LIME (Local Interpretable Model-agnostic Explanations)

LIME

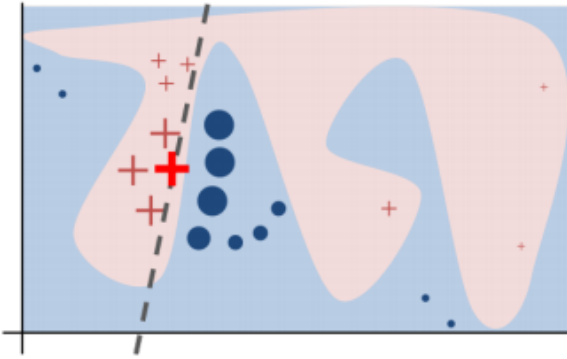
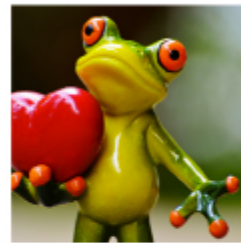
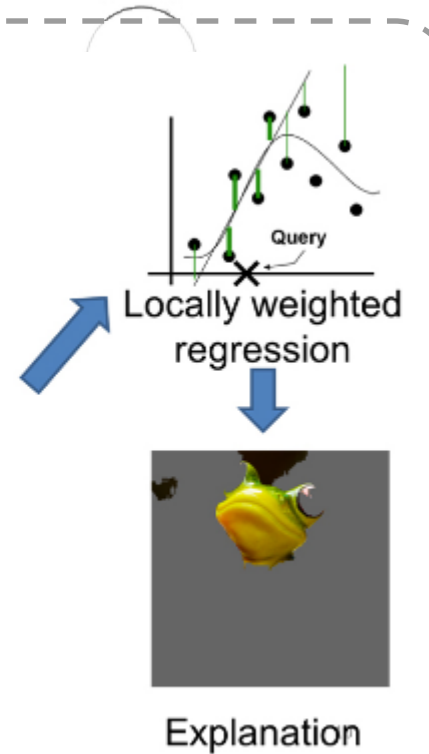


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f , and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

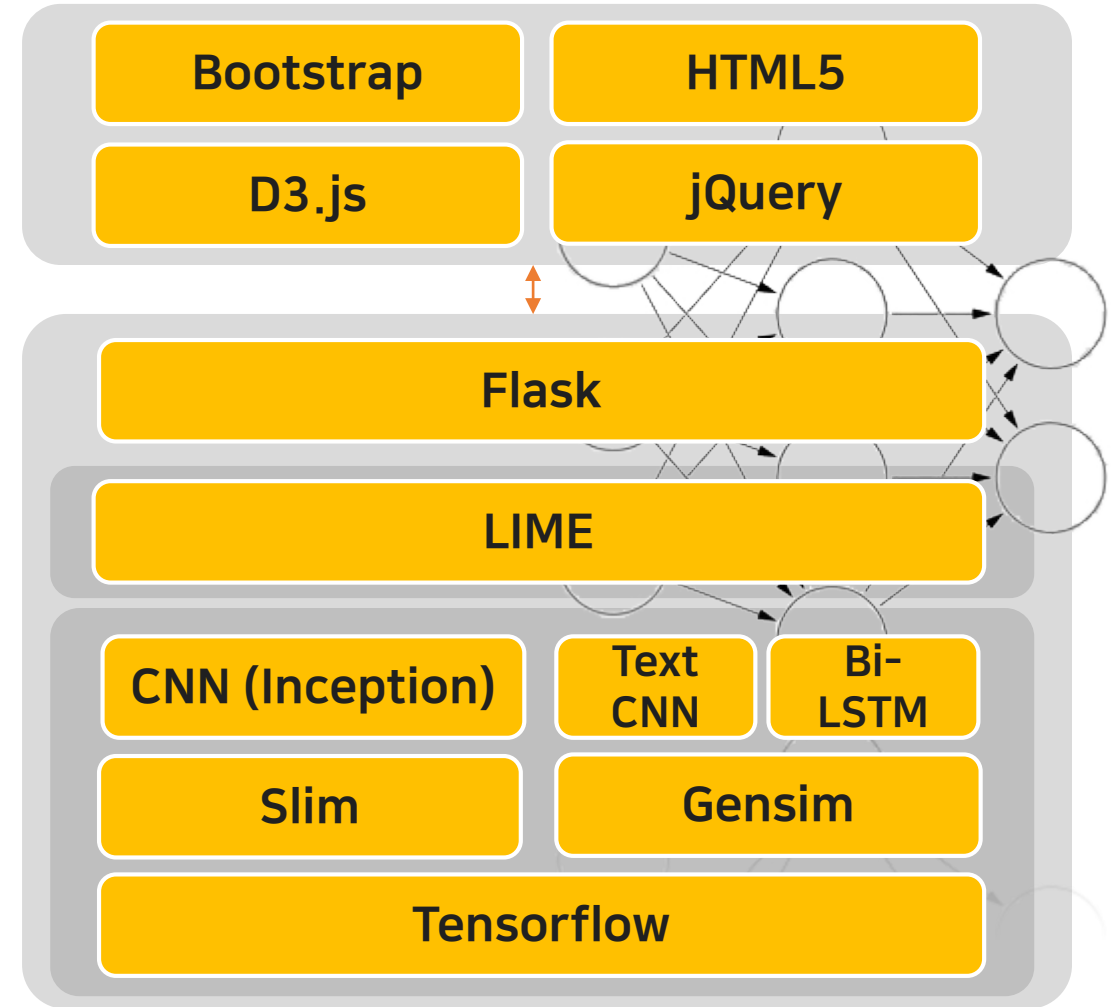
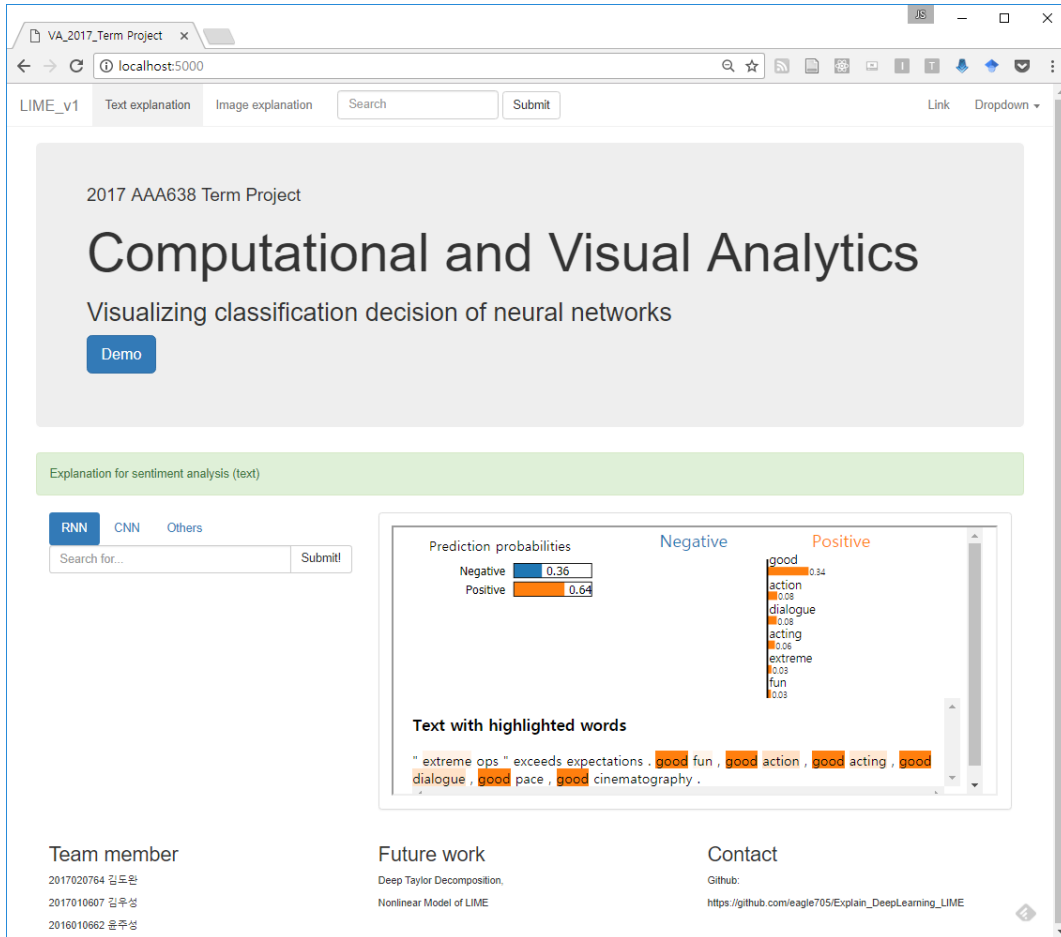


Original Image
 $P(\text{tree frog}) = 0.54$

Perturbed Instances	$P(\text{tree frog})$
	0.85
	0.00001
	0.52



System (Components)



System For Image Explanation

LIME_v1 Text explanation Image explanation Search Submit Link Dropdown

2017 AAA638 Term Project

Computational and Visual Analytics

Visualizing classification decision of neural networks

Demo

Explanation for classification (image)

Guitar Babe Pig

☒ Segmented Image
Features : 1 2 3 4 5 10

[403] (47%) acoustic guitar
[209] (13%) Labrador retriever
[208] (6%) golden retriever
[163] (2%) beagle
[715] (1%) pick, plectrum, plectron
Clear



Team member

2017020764 김도환
2017010607 김우성
2016010662 윤주성

Future work

Deep Taylor Decomposition,
Nonlinear Model of LIME

Contact

Github:
https://github.com/eagle705/Explain_DeepLearning_LIME

Guitar

Babe Pig

☒ Segmented Image

Features : 1 2 3 4 5 10

[403] (47%) acoustic guitar
[209] (13%) Labrador retriever
[208] (6%) golden retriever
[163] (2%) beagle
[715] (1%) pick, plectrum, plectron
Clear

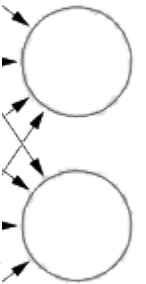
Guitar

Babe Pig

☐ Segmented Image

Features : 1 2 3 4 5 10

[403] (47%) acoustic guitar
[209] (13%) Labrador retriever
[208] (6%) golden retriever
[163] (2%) beagle
[715] (1%) pick, plectrum, plectron
Clear



System For Image Explanation (Segmentation)

LIME_v1 Text explanation Image explanation Search Submit Link Dropdown + Guitar Babe Pig

2017 AAA638 Term Project

Computational
Visualizing classification

Demo

1 : QuickShift : 58 pieces (19.58523416519165 sec)
2 : felzenszwalb : 306 pieces (0.5878908634185791 sec)
3 : SLIC : 90 pieces (3.7114615440368652 sec)

Quickshift felzenszwalb SLIC

Explanation for classification (Image)

Guitar Babe Pig

Segmented Image
Features : 1 2 3 4 5 10

[403] (47%) acoustic guitar
[209] (13%) Labrador retriever
[208] (6%) golden retriever
[163] (2%) beagle
[715] (1%) pick, plectrum, plectron
Clear

Team member
2017020764 김도환
2017010607 김우성
2016010662 윤주성

Future work
Deep Taylor Decomposition,
Nonlinear Model of LIME

Contact
Github
https://github.com/eagle705/Explain_DeepLearning_LIME

[715] (1%) pick, plectrum, plectron
Clear



System For Text Explanation

LIME_v1 Text explanation Image explanation Search Submit Link Dropdown

2017 AAA638 Term Project

Computational and Visual Analytics

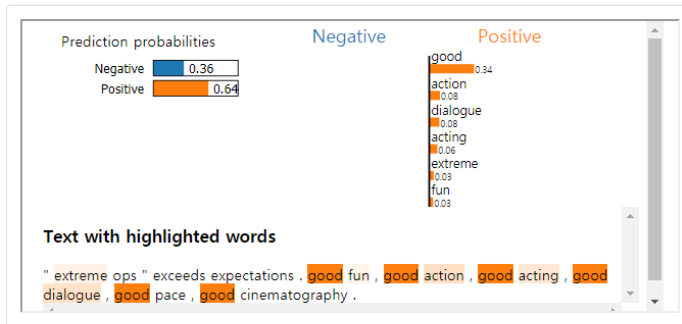
Visualizing classification decision of neural networks

Demo

Explanation for sentiment analysis (text)

RNN CNN Others

Search for... Submit!



Team member

2017020764 김도환
2017010607 김우성
2016010662 윤주성

Future work

Deep Taylor Decomposition,
Nonlinear Model of LIME

Contact

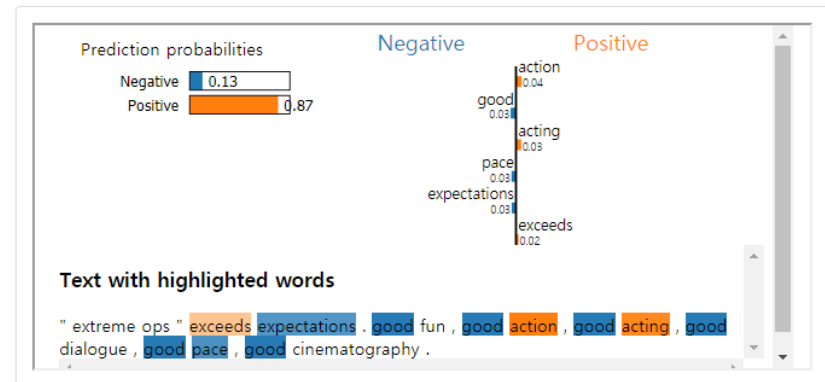
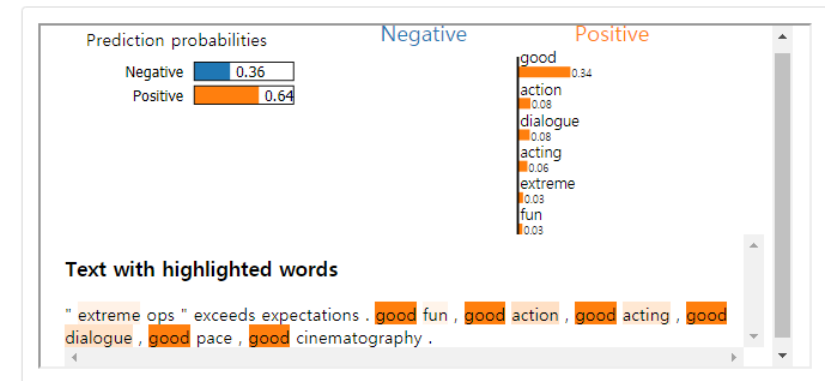
GitHub:
https://github.com/eagle705/Explain_DeepLearning_LIME

RNN CNN Others

" extreme ops " exceeds expectations . go Submit!

RNN CNN Others

" extreme ops " exceeds expectations . go Submit!



System For Text Explanation

LIME_v1 Text explanation Image explanation Search Submit Link Dropdown

2017 AAA638 Term Project

Computational and Visual Analytics

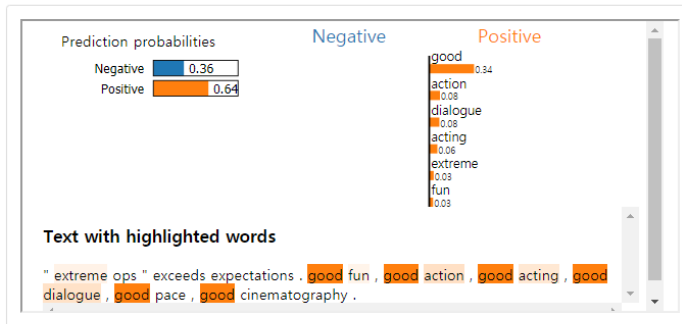
Visualizing classification decision of neural networks

Demo

Explanation for sentiment analysis (text)

RNN CNN Others

Search for... Submit!



Team member

2017020764 김도원
2017010607 김우성
2016010662 윤주성

Future work

Deep Taylor Decomposition,
Nonlinear Model of LIME

Contact

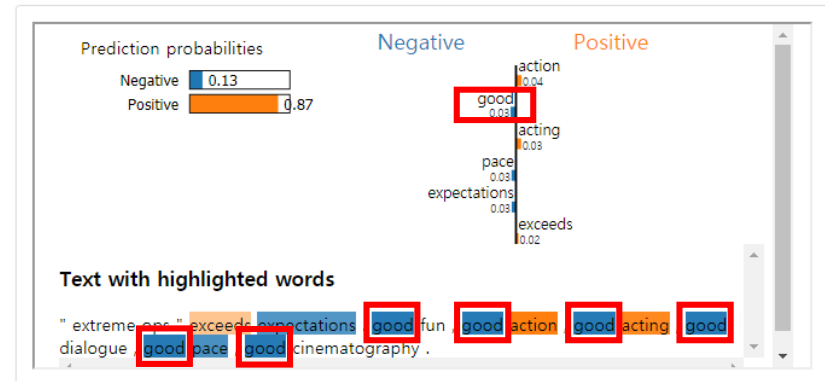
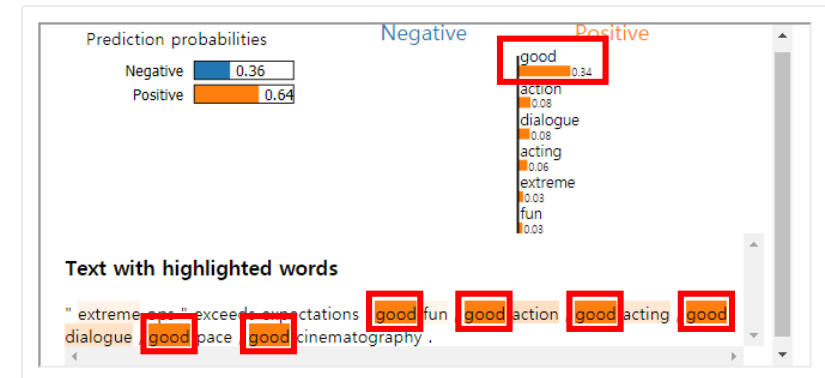
GitHub:
https://github.com/eagle705/Explain_DeepLearning_LIME

RNN CNN Others

"extreme ops" exceeds expectations. go Submit!

RNN CNN Others

"extreme ops" exceeds expectations. go Submit!



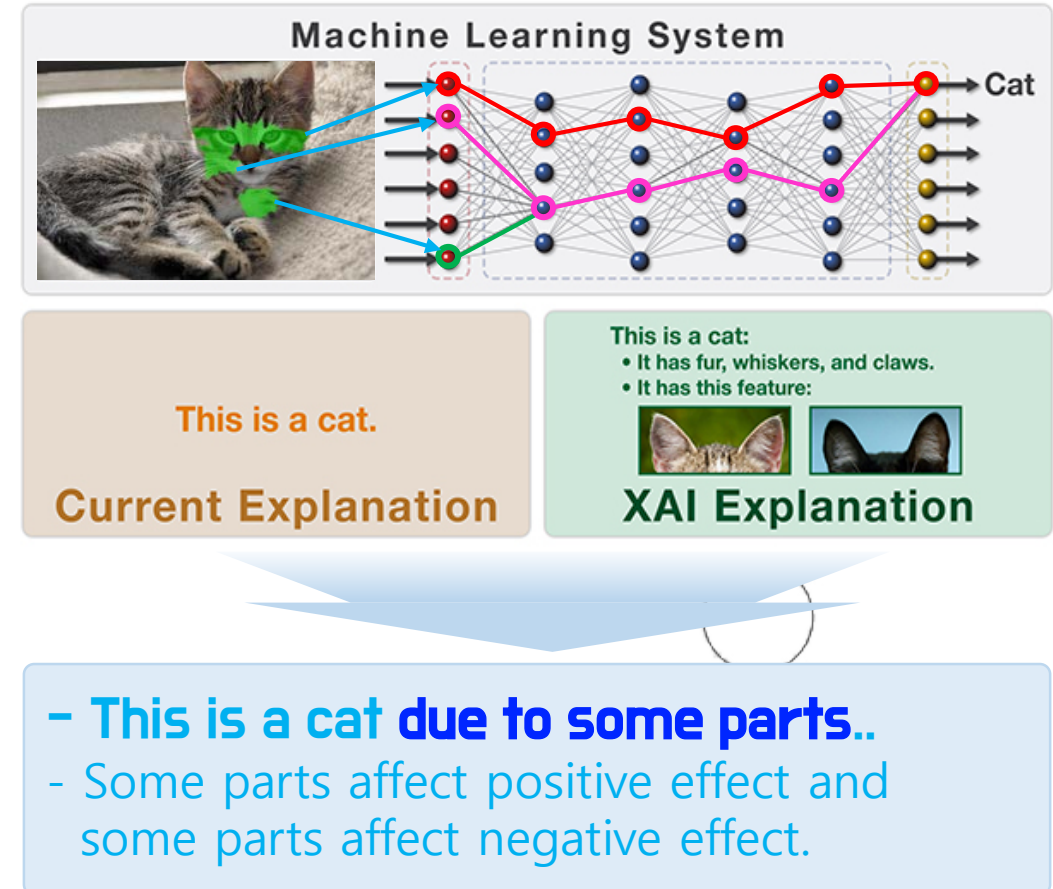
Conclusion & Future Work

□ We can see How Machine thinks

- Machine's thinking is similar with human but **not the same**
- LIME can help human to understand machine's prediction
- LIME for text depends on the model (CNN vs RNN)
(Segmentation can be inappropriate for sequential data)

□ Future Work

- Object Based Interpretation (Segmentation)
- Speed Up (Current 10 mins → a few seconds)
- Upgrade LIME for text (Deal with sequential data)



Demo URL: <https://youtu.be/F3bUDDlaFrc>

Thank you :)

