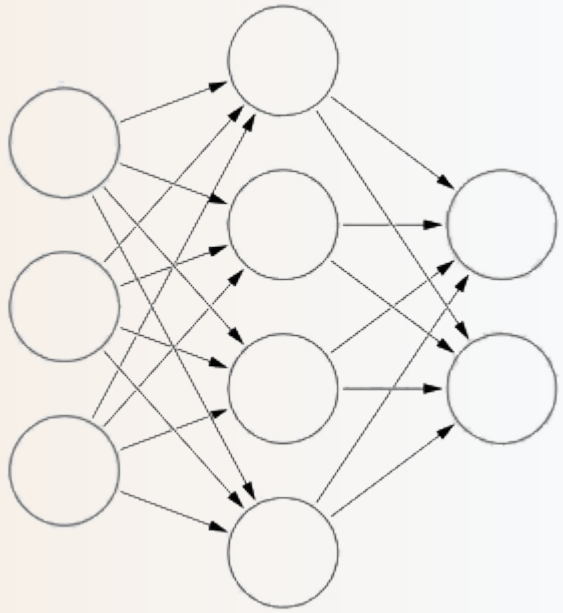
Visualizing classification decision of neural networks
(The Way How machine thinks)



Korea University,
2016010662 윤주성
2017020764 김도완
2017010607 김우성

AAA638 Visual Analytics Spring 2017

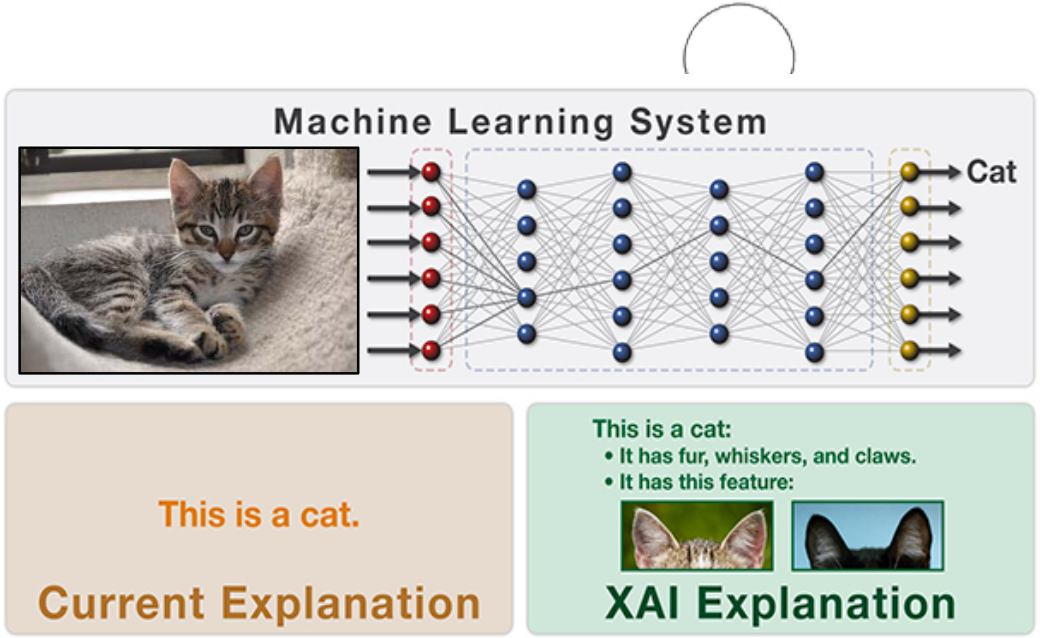
Intro

Deep Learning

High performance / Low explanation

Medicine / Finance / etc

Needs for explainable AI are increasing



Keywords

XAI, Neural Networks, Blackbox

Approach - LIME (Local Interpretable Model-agnostic Explanations)

LIME

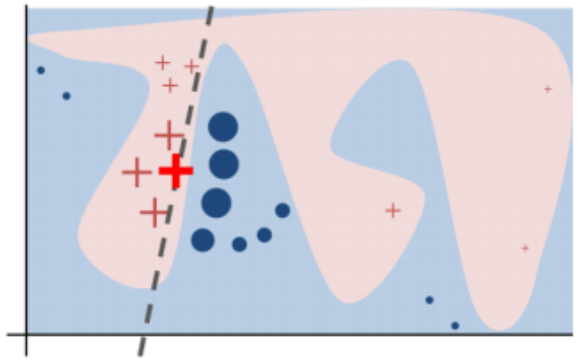
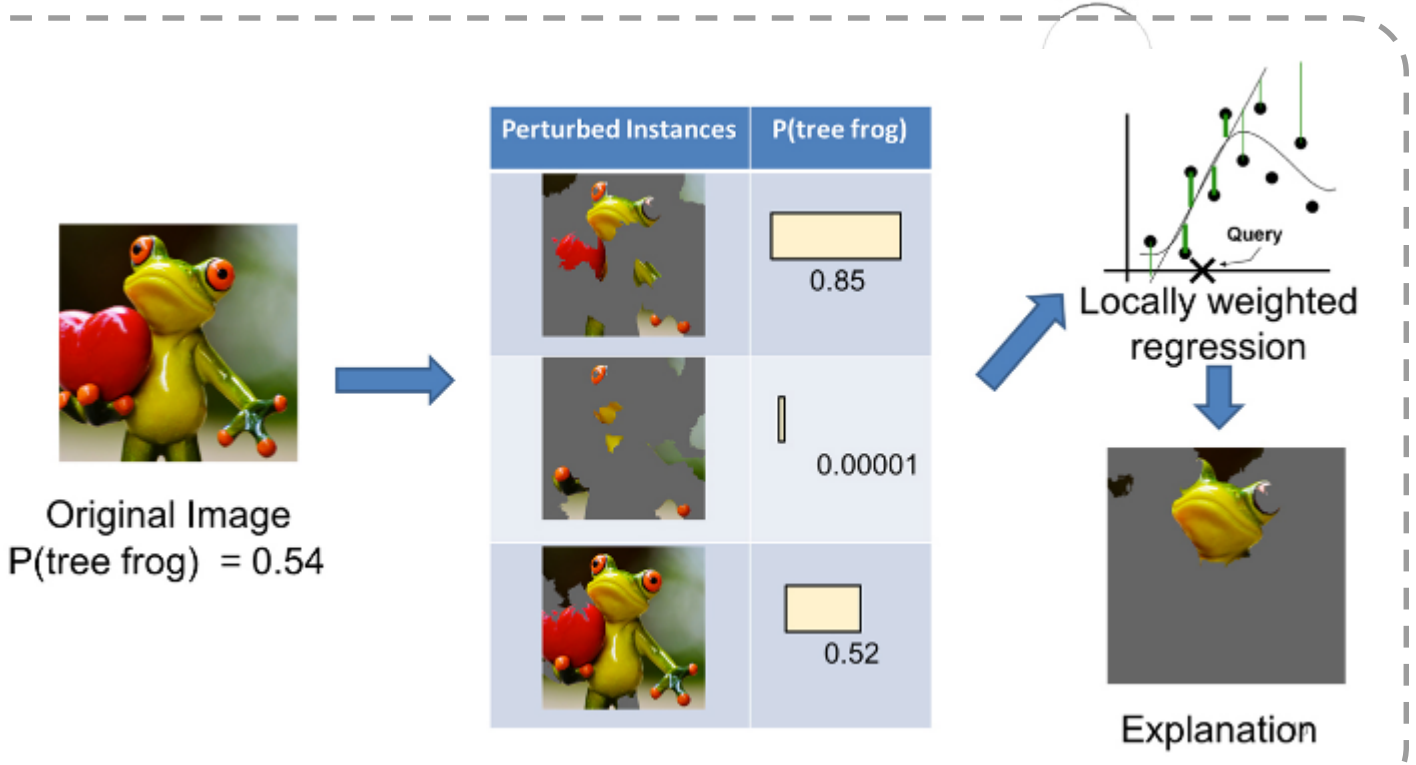
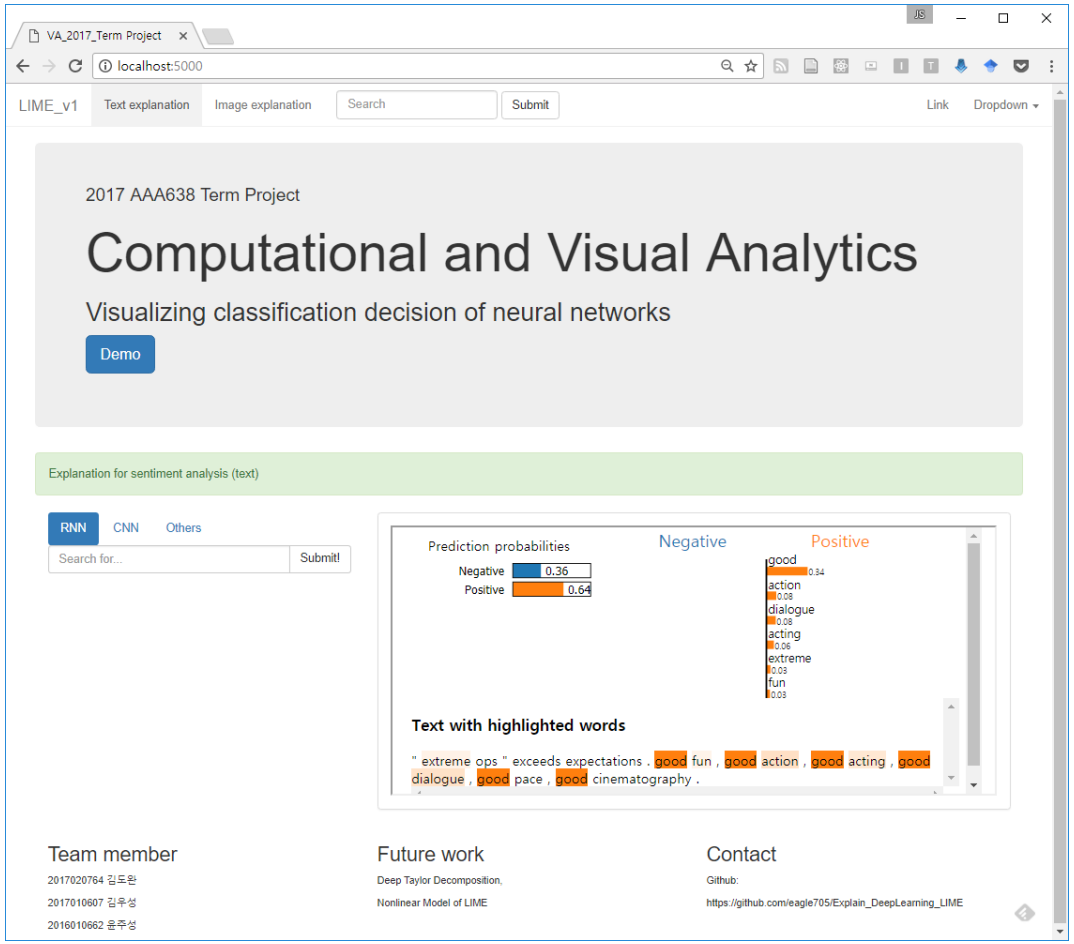
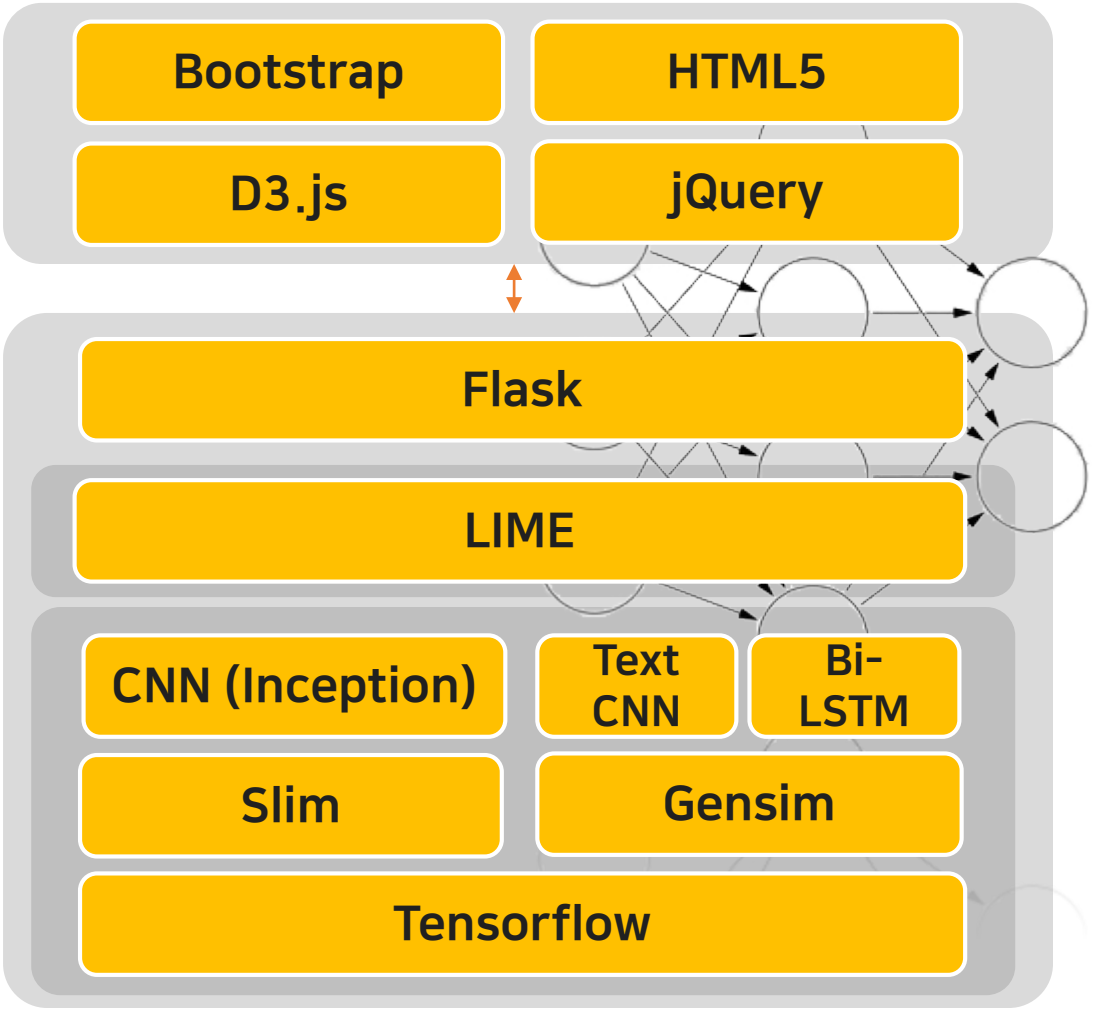


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions 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.

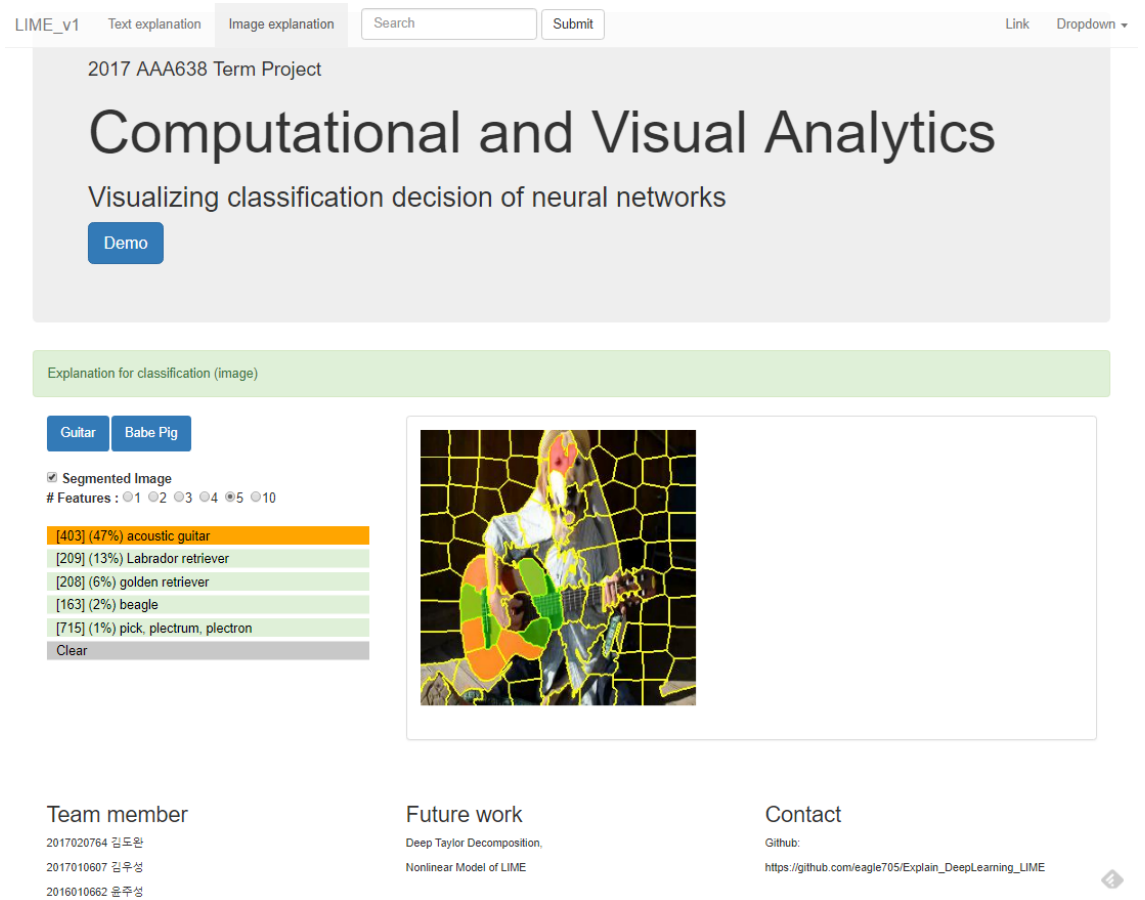


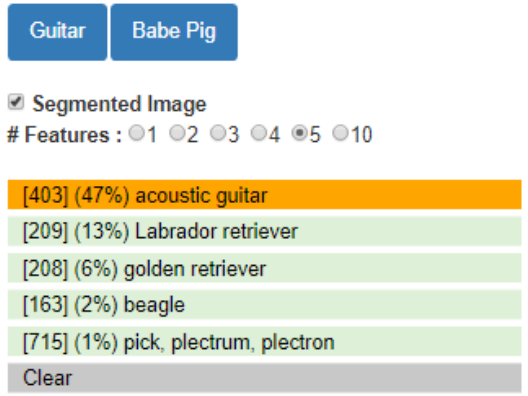
System (Components)

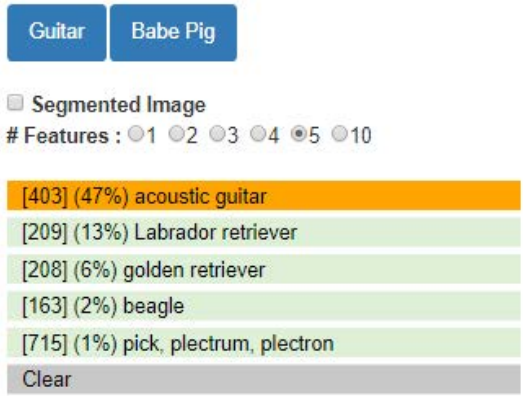


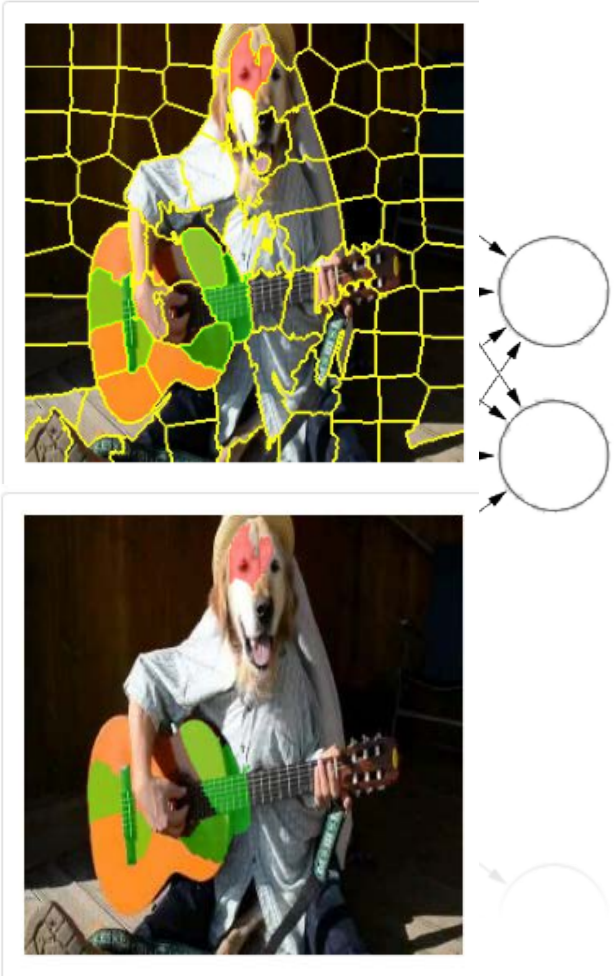


System For Image Explanation









System For Image Explanation (Segmentation)

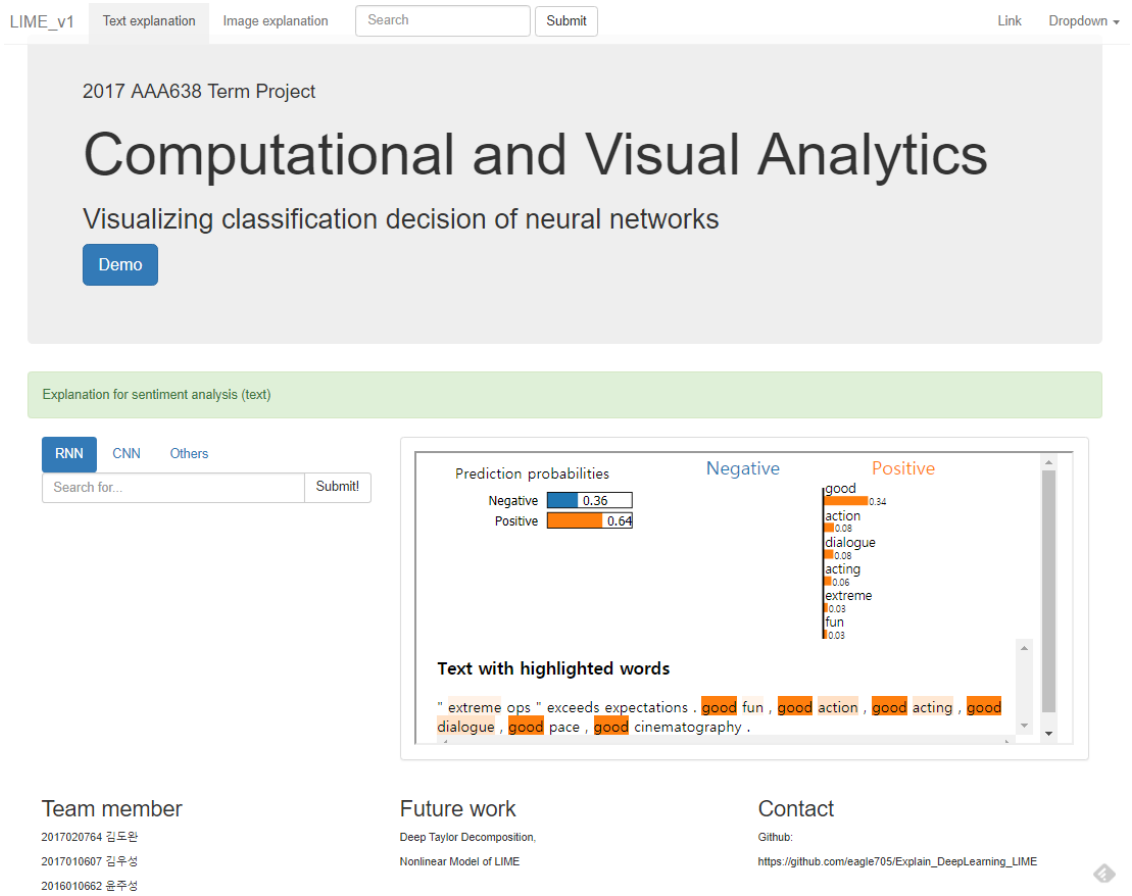
1 : QuickShift : 58 pieces (19.585234165191650 sec)
2 : felzenszwalb : 306 pieces (0.5878908634185791 sec)
3 : SLIC : 90 pieces (3.7114615440368652 sec)

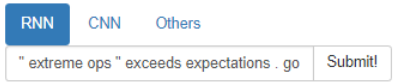
Quickshift

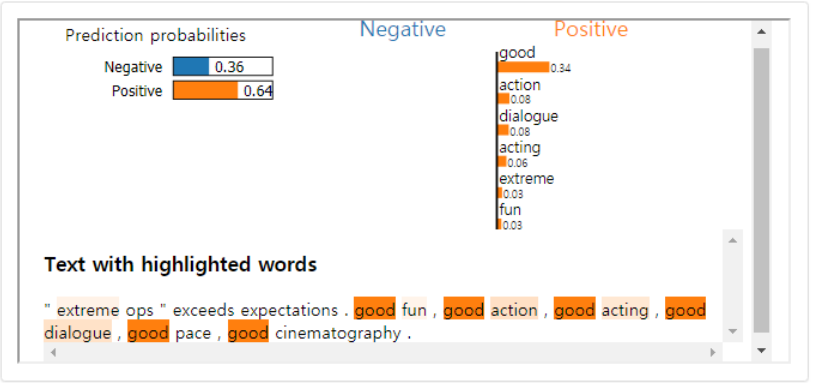
felzenszwalb

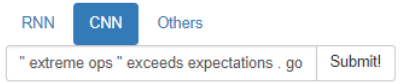
SLIC

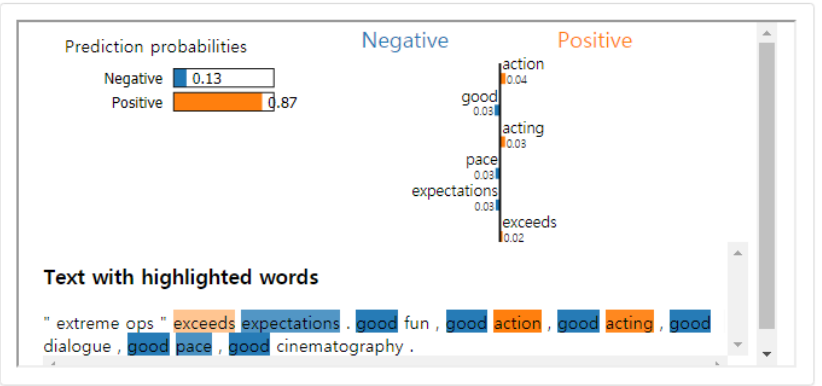
System For Text Explanation



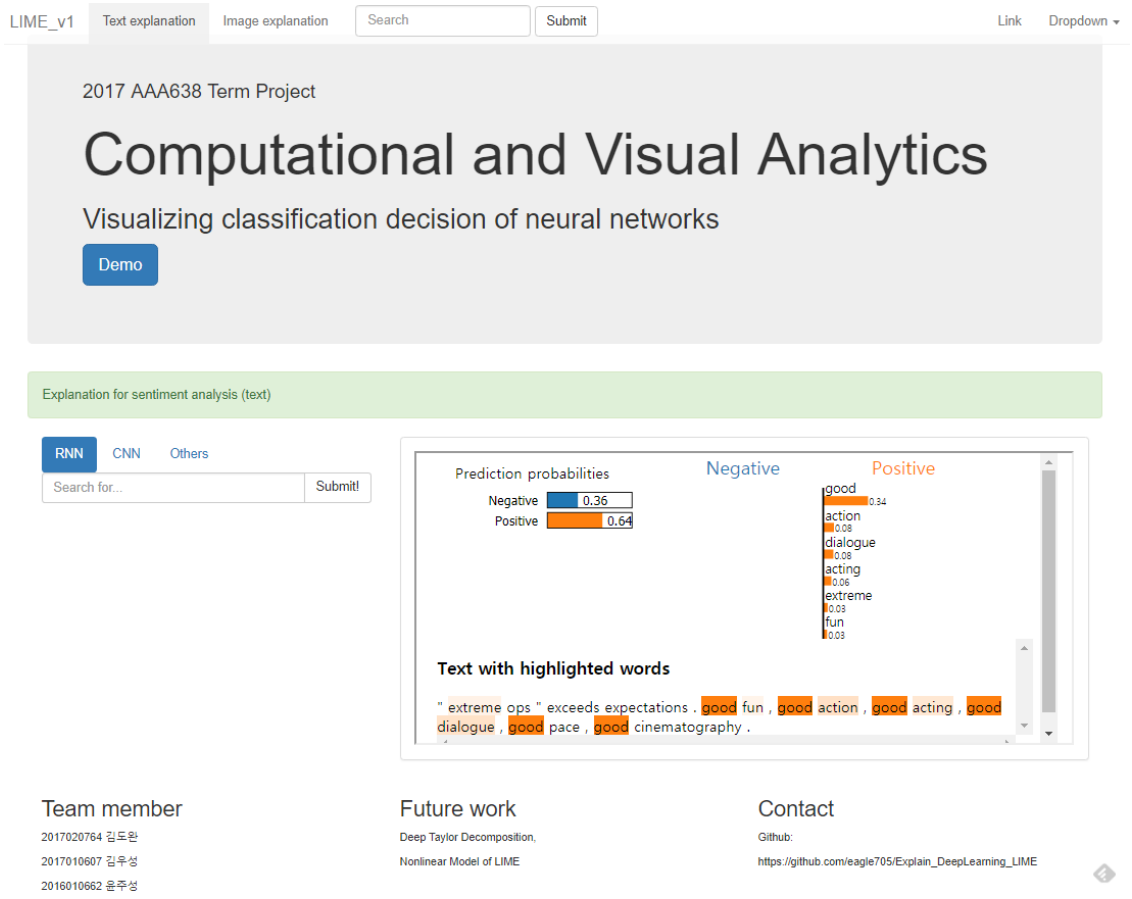


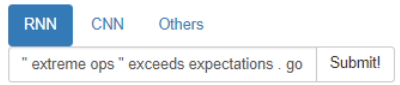


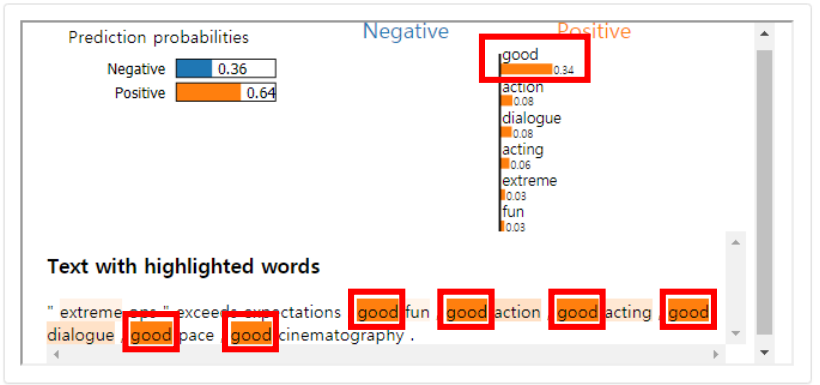


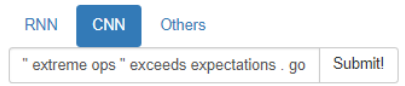


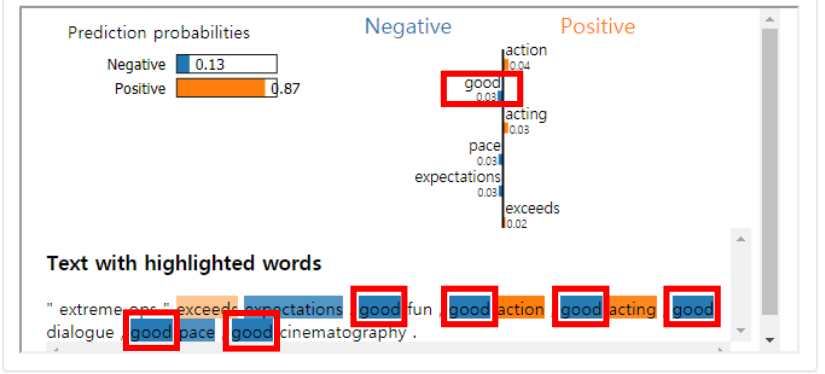
System For Text Explanation











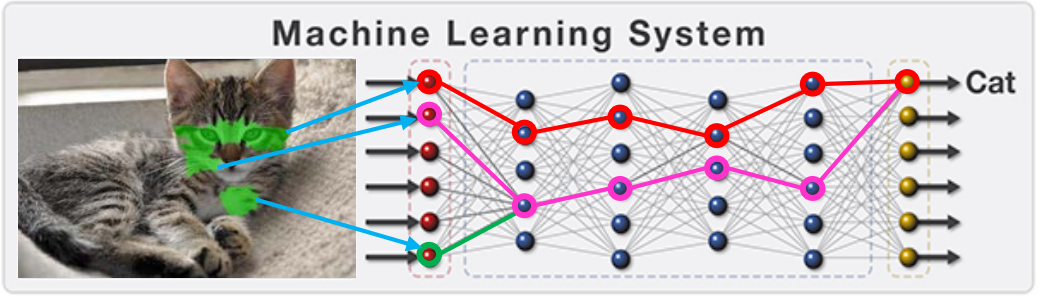
Conclusion & Future Work

☐ 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)



This is a cat.

Current Explanation

This is a cat:
- It has fur, whiskers, and claws.
- It has this feature:

XAI Explanation

- This is a cat due to some parts..

- Some parts affect positive effect and some parts affect negative effect.

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

Q&A

Thank you :)