FLEX: Faithful Linguistic Explanations for Neural Net based Model Decisions

Sandareka Wickramanayake **Wynne Hsu** **Mong Li Lee**

{sandaw, whsu, leeml}@comp.nus.edu.sg

School of Computing, National University of Singapore

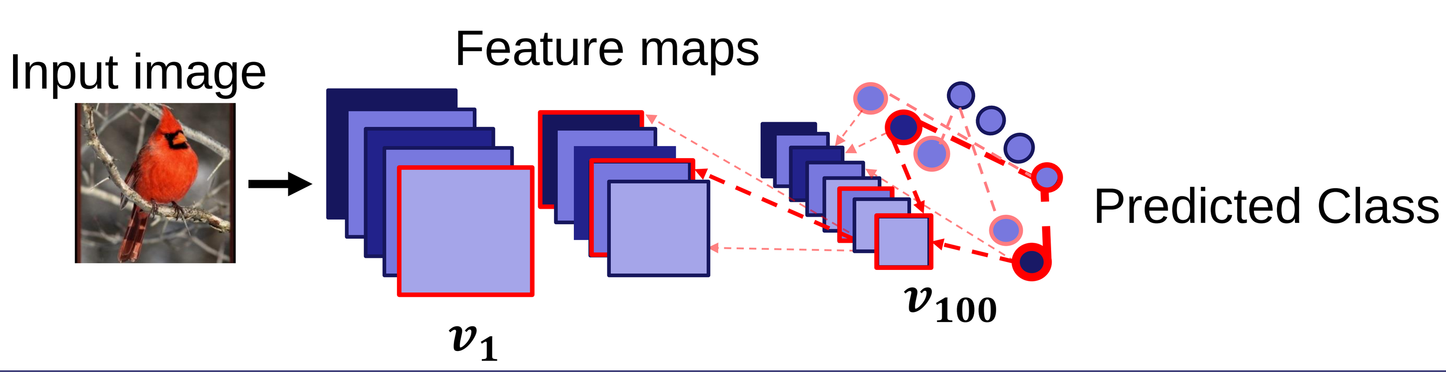
Introduction

- Explanations provide insights into DNN decisions.
- Explanations must be intuitive, descriptive, and faithfully explain why a model makes its decisions.
- FLEX framework generates post-hoc linguistic justifications to rationalize the decision of a convolution neural network in terms of features that are responsible for the decision.

FLEX Framework

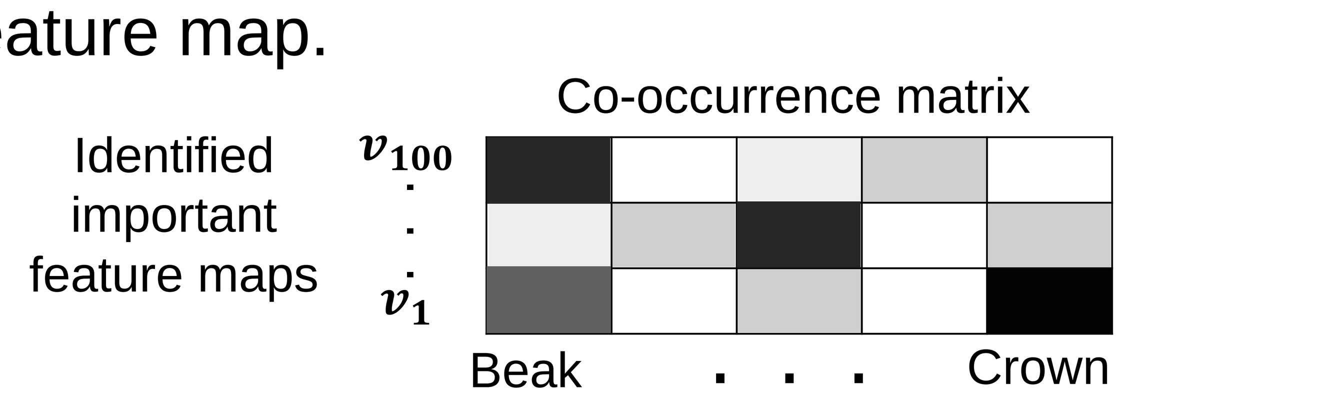
1 Identify Important Features

- Backpropagate gradients of the predicted class score to calculate importance of each feature map.
- Select the minimum set of feature maps such that cumulative importance score is greater than a given threshold.



2 Associate Words to Features

- Calculate co-occurrence score between each important feature map and each noun/adjective in ground truth descriptions.
- The word with the highest score is associated with the feature map.



3 Describe Decision Relevant Features

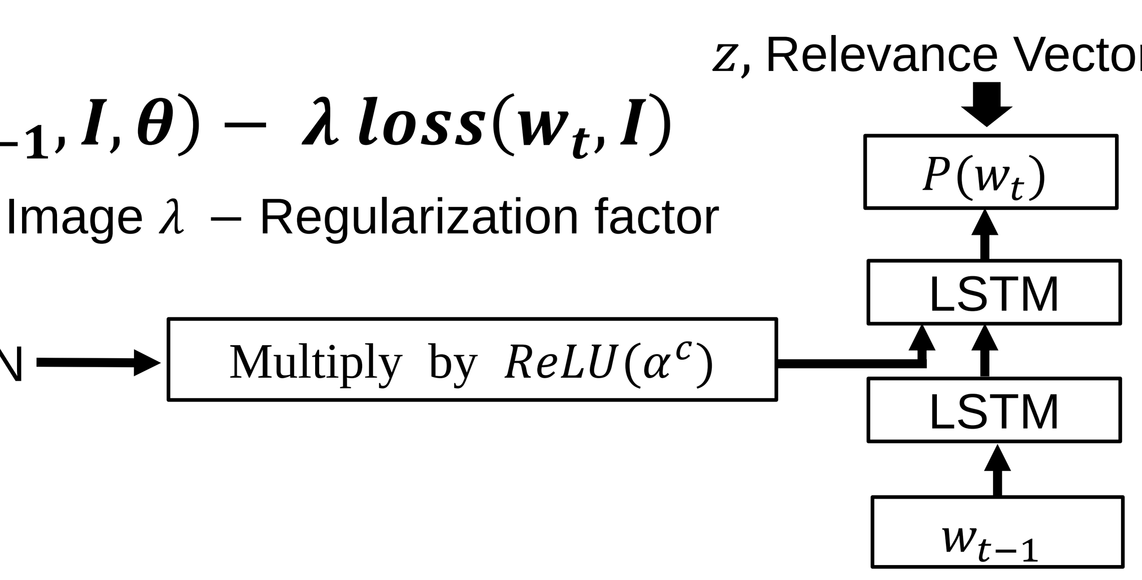
- Get top-k feature maps from the last conv layer based on importance scores, along with associated words.
- For each of these, recursively select top-k feature maps from inner conv layers and their associated words.
- All the words associated with top-k feature maps describe features used by the model for its decision.
- Calculate the relevance vector (z) using these words.

4 Generate Linguistic Justification

- Relevance loss, $loss(w_t, I) = \max(z \odot P(w_t | w_{\le t-1}, I))$
- Train a LSTM network by optimizing the objective function,

$$-\sum_t \log p(w_t | w_{t-1}, I, \theta) - \lambda\, loss(w_t, I)$$

θ – Model parameters, I – Image λ – Regularization factor



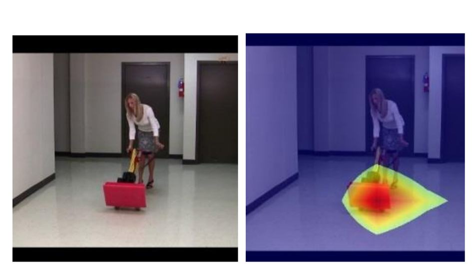
Performance Study

- Compare with **methods**: GVE [1] and MME [2] **Datasets**: CUB [3] and MPII [4]
- **Metrics: BLEU** – indicate fluency of the explanation and how well it matches image content.
  DREL – indicate how well the explanation matches the visual features used by a model in its prediction.

Comparative Study

	DREL		BLEU - 4	
	CUB	MPII	CUB	MPII
FLEX	**17.85**	**16.11**	**30.16**	19.11
GVE	15.67	13.46	28.43	13.71
MME	15.02	13.92	27.94	**19.88**

This is a **Scarlet Tanager** because
FLEX: This bird is red in color with a black beak, and **black eye**.
GVE: This is a red bird with black wings and a small beak.
MME: This bird has a red crown and a red breast.



This is classified to **polishing floors, standing, using electric polishing machine** class because
FLEX: She **is standing** in a room with **a floor polisher** and a rag in her hand.
GVE: She is kneeling on the floor with a carpet and is wearing exercise clothing.
MME: She is holding a mop and is in the middle of moving a mop.

Insights into Incorrect Model Decisions

- Common features between true class and predicted class caused model to misclassify.
- FLEX provides insights for 70.5% of misclassified examples where the generated explanations involve common features.

Fox Sparrow misclassified as Sage Thrasher

Fox Sparrow

Typical images of Sage Thrasher

FLEX: This bird has a **speckled belly and breast** with a **short pointy bill**.

Annotate Decision Relevant Features

- Association between words and feature maps allows automatic annotations of image.
- 88% of the CUB images have at least 1 out of 15 parts correctly annotated.

References

[1] Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell,T. Generating visual explanations. ECCV 2016.
[2] Park, D. H.; Hendricks, L. A.; Akata, Z.; Rohrbach, A.; Schiele, B.; Darrell, T.; and Rohrbach, M. Multimodal explanations: Justifying decisions and pointing to the evidence. CVPR 2018.
[3] Welinder P., Branson S., Mita T., Wah C., Schroff F., Belongie S., Perona, P. "Caltech-UCSD Birds 200". California Institute of Technology. CNS-TR-2010-001. 2010.
[4] Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B.. 2d human pose estimation: New benchmark and state of the art analysis. CVPR 2014.