RESTRICTING THE FLOW: INFORMATION BOTTLENECKS FOR ATTRIBUTION

Karl Schulz, Leon Sixt, Federico Tombari, Tim Landgraf

Introduction

- What are attribution methods?

  They help us to understand how networks make decisions by assigning some kind of scores to the inputs or the weights of the network.

- Why is it necessary/motivation?

  In several applications it is absolutely necessary to know how networks are making decisions eg: medical diagnosis.

Contributions of the paper

- Propose a information bottleneck framework for attribution.
- Give a information theoretic guarantee for their method i.e areas with zero bit of information are not used by the network.
- Give 2 models: **per-sample bottleneck** and **readout bottleneck** to do the task
- Contribute a novel evaluation method based on bounding boxes
- Amazing and well documented code.

Theory

$$\max I[Y;Z] - \beta I[X,Z]$$

Introduce a new random variable Z, through some operation (here noise), such that the information Z shares with Y is maximized and information Z shared with X is minimized.

What is mutual Information

- Say we have a prediction variable Y. There is of course some uncertainty associated with this which is quantified by entropy (higher the more uncertain)
- Now what mutual information tells us is if we know some X (may be input image or a message or anything that results in Y) how much decrease in entropy will be there or decrease in uncertainty.
- It is Symmetric

Some more equations...

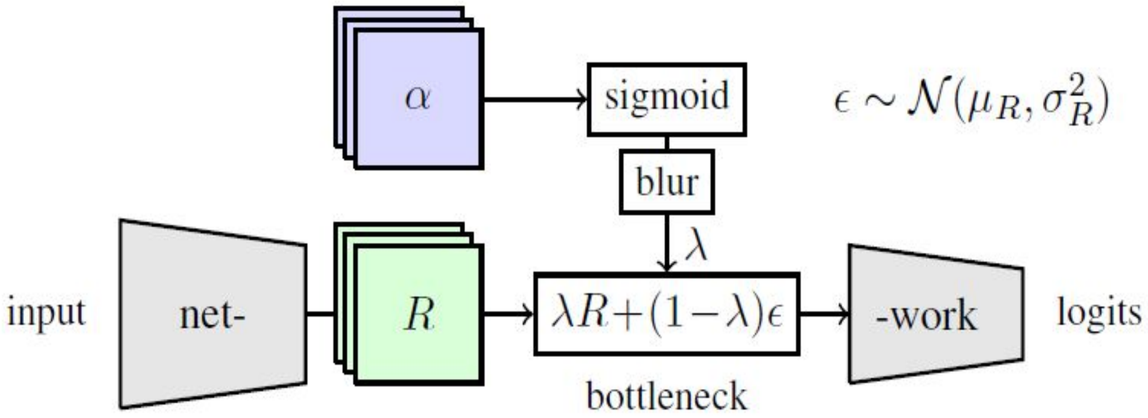
$$\mathrm{I}[R, Z] = \mathbb{E}_R[D_{\mathrm{KL}}[P(Z|R)||P(Z)]] ,$$

$$\mathrm{I}[R, Z] = \mathbb{E}_R[D_{\mathrm{KL}}[P(Z|R)||Q(Z)]] - D_{\mathrm{KL}}[P(Z)||Q(Z)]$$

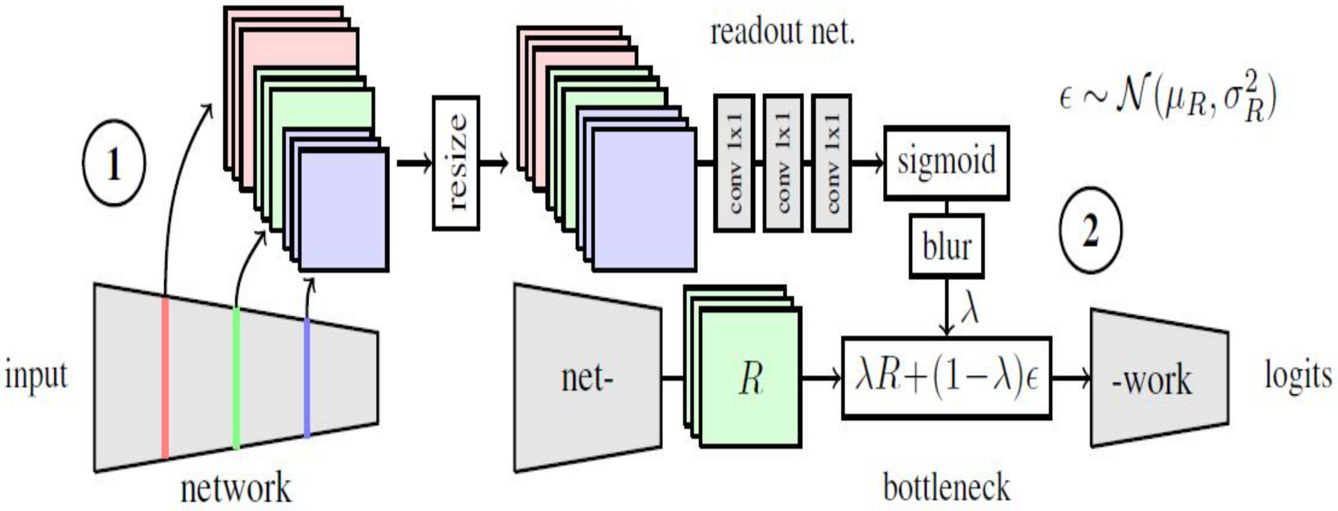
$$\mathcal{L}_I = \mathbb{E}_R[D_{\mathrm{KL}}[P(Z|R)||Q(Z)]]$$

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \mathcal{L}_I$$

Per Sample Bottle neck


α
sigmoid
$\epsilon \sim \mathcal{N}(\mu_R, \sigma_R^2)$
blur
λ
input
net-
R
$\lambda R + (1 - \lambda)\epsilon$
-work
logits
bottleneck

Readout bottle neck



Experiments

Let's look at the paper!

Conclusion

- This paper presents a good approach to understand what the model is not focusing on (also focusing on) during taking decisions.
- There is a strong theoretical background for this approach which can be utilized in future works and its highly flexible.
- You get a score (in bits), which makes it comparable with other models.

THANK YOU