Toward Explainable Neural-Symbolic Reasoning

Adrien Bennetot[1,2,3] Jean-Luc Laurent[2] Raja Chatila[3] Natalia Díaz-Rodríguez[1]

[1] U2IS, ENSTA Paris, Institut Polytechnique de Paris, Palaiseau, France - Inria FLOWERS team https://flowers.inria.fr/

[2] Segula Technologies, Parc d'activité de Pissaloup, Trappes, France - [3] Institut des Systèmes Intelligents et de Robotique, Sorbonne Universite, France

Abstract

Truly explainable models should not leave explanation generation to the human user [1]. Therefore, a prediction model must generate itself an explanation of its rationale in natural language. We show how techniques integrating connectionist and symbolic paradigms can achieve performances similar to the state of the art while being more explainable. Our approach is to create a Knowledge Base (KB), without external intervention, that would influence the neural network by modifying its loss functions in order to explain its reasoning.

Contribution

1) Automating bias detection for an image captioning task and creation of a KB directly from the data.

2) Generalization of [2] losses function in order to be automatically applicable to an unlimited number of classes.

3) Obtaining results similar to the state of the art of connectionist methods while using neural-symbolic reasoning

4) Giving an explanation on the prediction of the model by interrogating it on the basis of its KB.

Approach

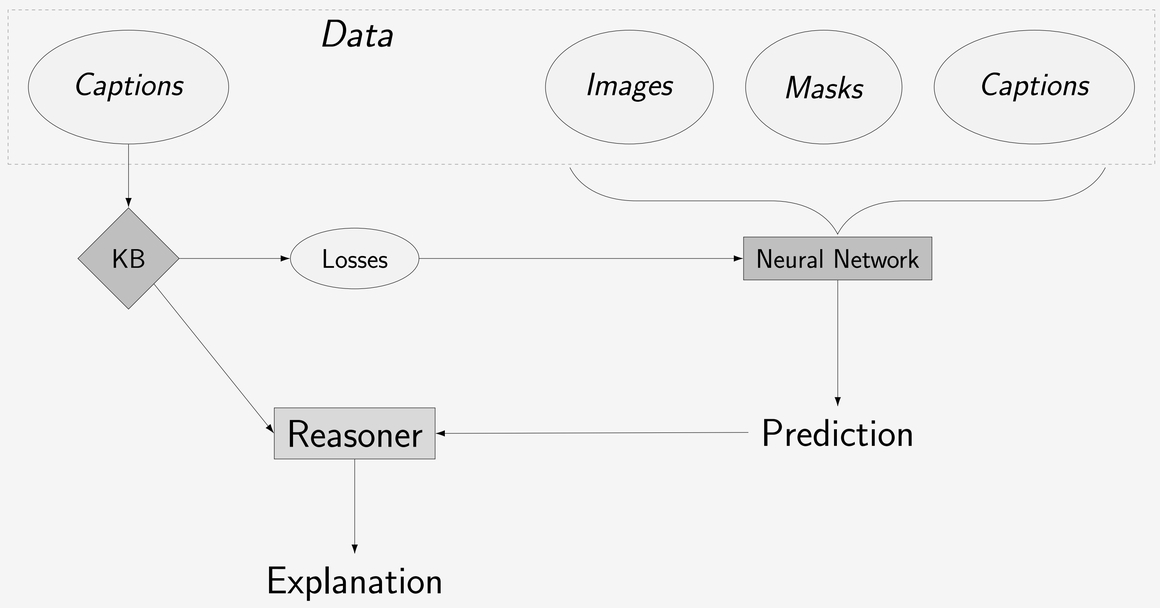

Figure 1: Captions are used to create the KB thanks to a Continuous Bag of Word model. The KB influences our Neural Network by modifying its losses, depending on the relationship between words. Images, segmentation masks, captions and the new losses are used to train the neural network. The prediction is used to retrace the KB in order to give an explanation of the outcome.

Loss Function 1: Confusion: Force our model to be confused when making predictions about a subclass if the input image **does not contain** appropriate evidence for the prediction to be made.

$$C(\tilde{w}_t, I') = |\sum_{b \in B_{word}} (p(\tilde{w}_t = b | w_{0:t-1}, I') - \frac{1}{J})^2| \quad (1)$$

$$\mathcal{L}^{Confusion} = \frac{1}{N} \sum_{n=0}^{N} \sum_{t=0}^{T} \mathbb{1}(w_t \in B_{word}) C(\tilde{w}_t, I'), \quad (2)$$

Loss Function 2: Confidence: Force our model to be confident when making predictions about a subclass if the input image **does contain** appropriate evidence for the prediction to be made.

$$F^j(\tilde{w}_t, I) = \frac{\sum_{b \in B_{word} \setminus b_j} p(\tilde{w}_t = b | w_{0:t-1}, I)}{p(\tilde{w}_t = b_j | w_{0:t-1}, I) + \epsilon} \quad (3)$$

$$\mathcal{L}^{Confidence} = \frac{1}{N} \sum_{n=0}^{N} \sum_{t=0}^{T} \sum_{j=1}^{J} (\mathbb{1}(w_t = b_j) F^j(\tilde{w}_t, I)) \quad (4)$$

A neural network is trained with these losses, in addition to a standard cross entropy for words that are not contained in the KB.

Results

We compare ourselves to Burns et al. (2018) [2] work on gender captioning but adding 2 subclasses to the training, boys and girls. In order to stay fair, we only compare ourselves on the captioning of men and women

Datasets: MSCOCO-Bias and Balanced: images from MSCOCO which are labeled as "man" or "woman" with respectively a 1:3 and 1:1 woman to man ratio

Metrics: Performance for each classes and divergence between classes.

Baselines: Equalizer: linear combination of a confusion loss, a confidence loss and a cross-entropy

Model	Women Correct	Women Incorrect	Women Other	Men Correct	Men Incorrect	Men Other	Divergence
Neural-Symbolic Model	**62.48**	16.46	**21.06**	**69.58**	**4.96**	**25.47**	0.27
Equalizer	59.98	**13.80**	26.22	62.47	5.63	31.90	**0.16**

Figure 2: Mean prediction performance in % and Jensen–Shannon divergence on both MSCOCO-Bias and Balanced for the Neural-Symbolic and the Equalizer model. Woman (resp. men) other when the model predicts a person instead of precisely a woman (resp. man)

Explainability

Figure 3: Example of prediction and explanation where the model is not able to differentiate the subclasses

Figure 4: Example of a prediction and explanation where the model is able to differentiate the subclasses

Conclusion

- We experimented the addition of a KB to a neural network for an image captioning task.
- Preliminary results show a promising research direction where a model can achieve state-of-the-art performances while providing elements of explanation.
- Future work should experiment with more classes, as well as trying to automatically enlarge the KB.

References

[1] D. Doran, S. Schulz, and T. R. Besold, "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives," *arXiv e-prints*, p. arXiv:1710.00794, Oct 2017.

[2] K. Burns, L. A. Hendricks, K. Saenko, T. Darrell, and A. Rohrbach, "Women also Snowboard: Overcoming Bias in Captioning Models," *arXiv e-prints*, p. arXiv:1803.09797, Mar 2018.