# Towards Explainable Neural-Symbolic Visual Reasoning

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

adrien.bennetot@ensta-paristech.fr

[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

## Neural-Symbolic computation for Explainable AI

High-performance models suffer from a lack of interpretability. We show why techniques integrating connectionist and symbolic paradigms are the most efficient solutions to produce explanations and we propose a reasoning model to explain a neural network's decision. We use this explanation in order to correct bias in the network's decision rationale. We accompany this model with an example of its potential use for image captioning.

## Goal : explainable neural network

Truly explainable models should directly integrate reasoning in order to not leave explanation generation to the human user. The neural network, considered as a black box, must generate itself an explanation in natural language. We propose to create a Knowledge Base (KB), directly from the data, that would influence the neural network by modifying its loss function. We hypothesize that the use of loss functions that have a concrete and more graspable perceptible meaning could make it easier to provide an explanation than a classic non intuitive cross-entropy.

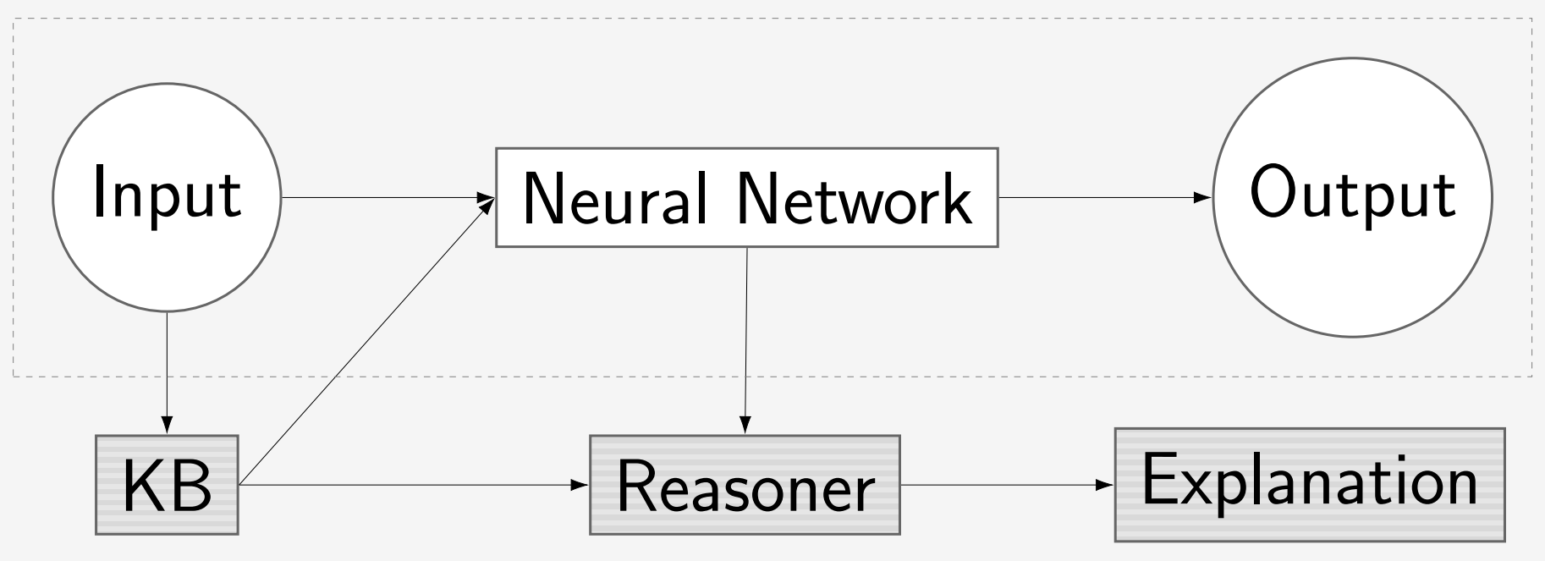

Figure 1: *Explainable Model*: in grey what is added compared to an usual network

Inspired by the work of [1], we apply this model to an image captioning task where we use images I, captions S and segmentation masks M to learn to a neural network how to fight bias in a dataset. We introduce two new losses representing the notions of confusion and confidence.

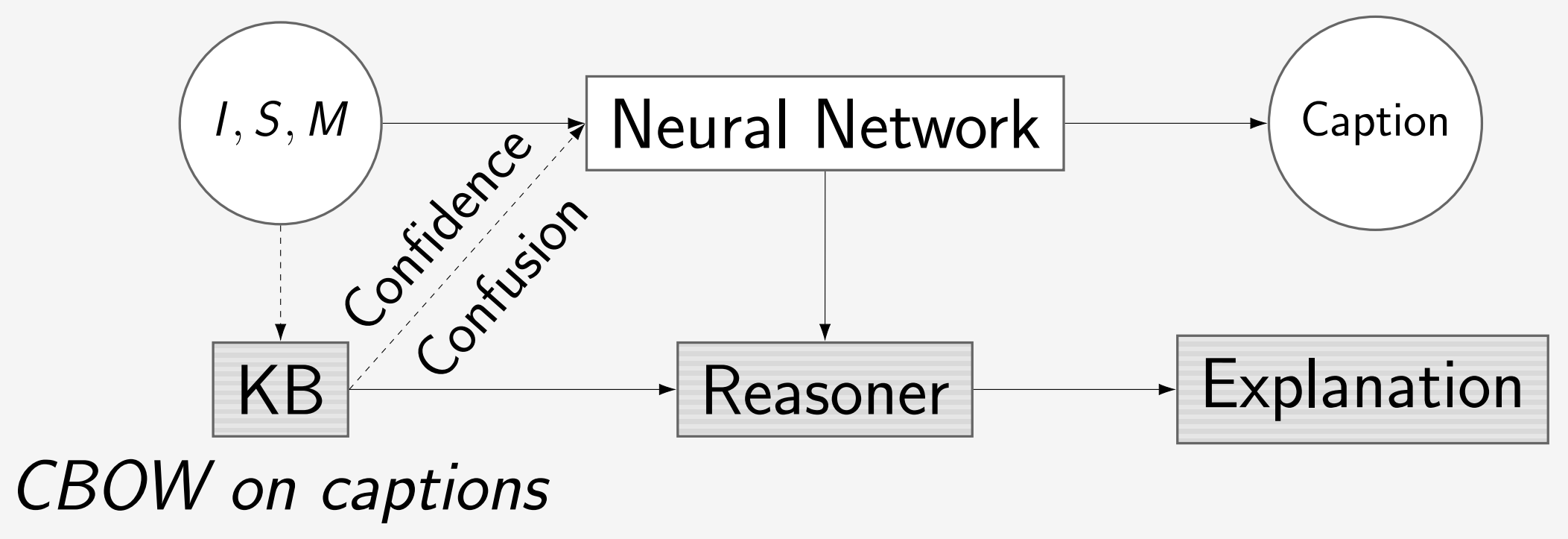

Figure 2: The KB is created thanks to a Continuous Bag Of Word (CBOW) model. Dashed arrows when only captions S are used

## Knowledge Base

We propose creating the reasoning-facilitating KB by performing word-embedding on the black box model labels in order to determine which words are particularly exposed to a risk of errors due to learning priors or biased data collection. Interchangeable words are more likely to be victims of overuse of context since they are involved in the same situations.

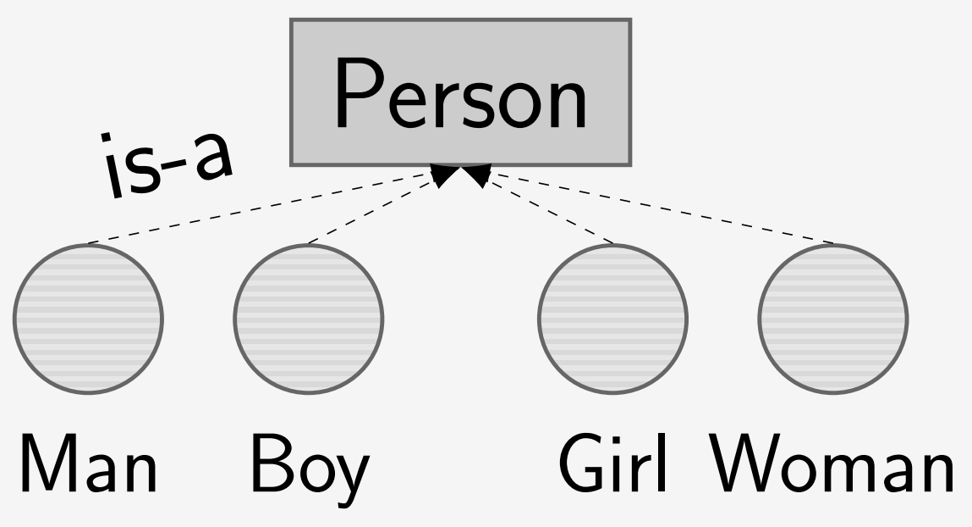

Figure 3: Example of knowledge obtained from the captions, representing relations between words. Dashed arrow for an *is-a* relation.

It allows us to create $B_{word}$ lists grouping words (sub-classes) having the same relationship with *word*, meaning that the model must be particularly careful when prediction one of them. In this case, we show an exemple of the list $B_{person} = [man, boy, girl, woman]$

## Confidence and confusion losses

We want to force our model to be confused when making predictions if the input image does not contain appropriate evidence for the prediction to be made. We use masked images $I' = I \cdot M$, where the information relevant to making a decision is removed. We note confusion function $C$ which operates over the predicted distribution of words $p(\tilde{w}_t)$:

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

where $J$ is the length of $B_{word}$. As we try to minimize $C(\tilde{w}_t, I')$, we have a sum of squares that tends toward zero, meaning that each probability tends to be equal. We define the confusion loss $\mathcal{L}^{Confusion}$ as:

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

with $\mathbb{1}$ an indicator variable that denotes whether or not $w_t$ is a bias-prone word, $N$ the batch size, and $T$ the number of words in the given sentence. As we want the model to be confident about its prediction when there is an appropriate information on the image, this time we use complete images $I$ as input instead of masked ones $I'$. With $j$ the index of word $b$ in list $B_{word}$, we have the confidence function $F^j$.

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

$F^j$ will tend towards zero if $p(\tilde{w}_t = b_j)$ dominates the sum of the predicted distribution of every other *bias-prone* word.
We use $F^j$ to define the confident loss $\mathcal{L}^{Confidence}$:

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

By adding a standard cross-entropy loss $\mathcal{L}^{CE}$ to non-bias-prone words, we obtain a model able to use context priors when there is no interchangeable word for the predicted one and to be confused/confident when the question arises thanks to the loss $\mathcal{L}$,, with $\alpha$, $\beta$ and $\mu$ hyper-parameters.:

$$\mathcal{L} = \alpha \mathcal{L}^{CE} + \beta \mathcal{L}^{Confidence} + \mu \mathcal{L}^{Confusion} \tag{5}$$

## Reasoner

The reasoner block gathers info from the KB and the neural network. If the prediction is a sub-class, it means that the model is confident. If the network's prediction is a class, it means that the model is confused and could not find enough evidences to chose a sub-class.

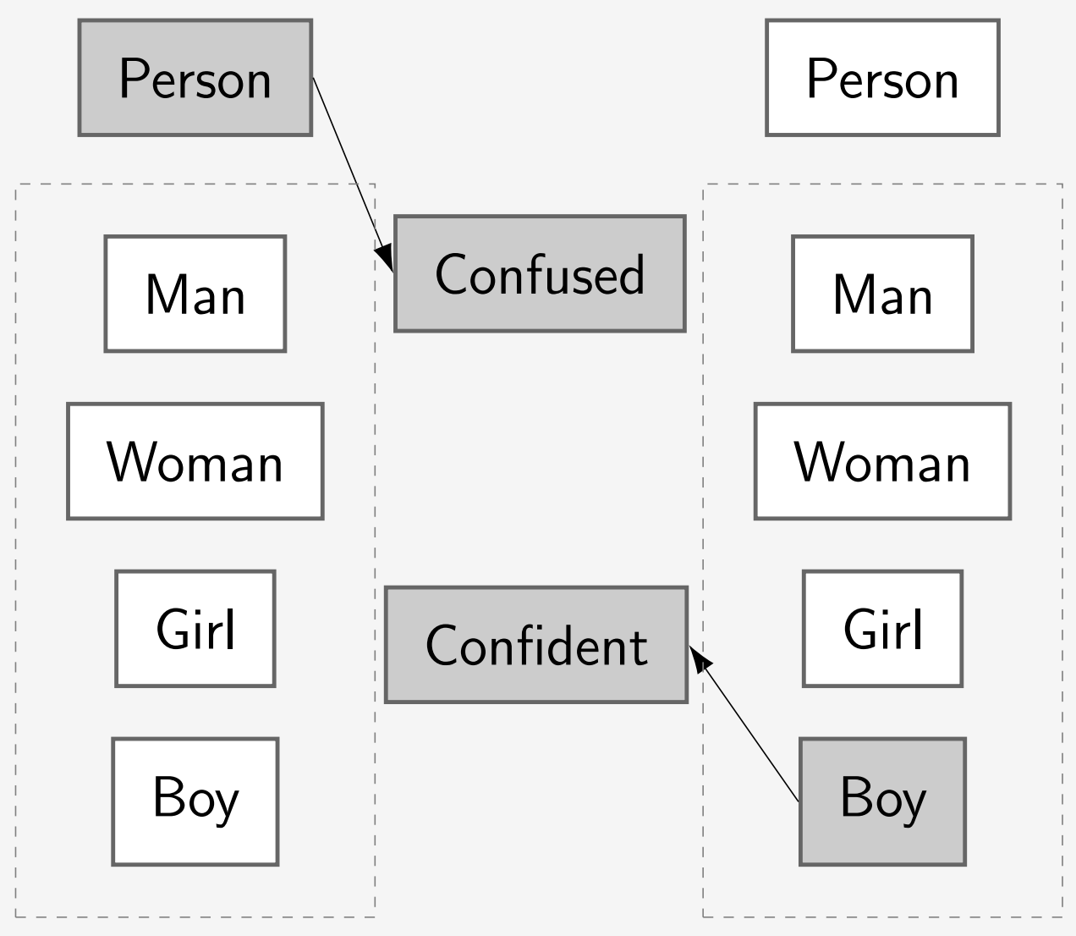

Figure 4: *2 Examples of reasoning*: Case 1 - *boy* is predicted, it means that the model is confident Case 2 - *person* is predicted, it means that the model is confused. In grey the predicted word by the model, dashed rectangle is a sub-class

## References

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

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

## Conclusion

We propose a model endowed with a non-external KB, i.e., directly built on the learning data of a neural network, that allows to influence its learning and to correct bias thoroughly, while giving a fair explanation from its predictions. As the user or expert external knowledge does not interfere the predictions in the explanation process, it constitutes a truly explainable model [2].