Current captioning approaches can describe images using black-box architectures whose behavior is hardly controllable and explainable from the exterior. As a