Bayesian Classification of Affordances from RGB Images

Lorenzo Mur-Labadía, Rubén Martínez-Cantín

Afiliación: Robotics and Perception in Real Time (RoPeRT)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: lmur@unizar.es, rmcantin@unizar.es

Abstract

We propose a Bayesian classification model using Deep Ensembles (DE) and MC-Dropout (MC-D) to predict affordances from RGB images. Our Bayesian model obtains a higher performance than previous works and captures the aleatoric and epistemic uncertainty, showing consistency with the type of objects and scenarios.

Introduction

Affordances are the different action possibilities available in the environment depending on the motor and sensing capabilities of the individual **[1]**. They relate the objects, the actions, and the possible effects of that actions carried on the objects. Based on this, affordance prediction emerges as a powerful tool for autonomous and active agents where we need to understand the scene content.

Methods

Based on the model from **[2]**, we build a deterministic classification model that predicts the effect of taking an action on an object. We combine local information from the object's bounding box h_{obj}, with the context of the scene $\phi(I)$ and the ground-truth class $\hat{c}$ of the respective object. We create a MLP with Fully Connected (FC) layers with ReLU non-linear activation and intermediate dropout layers to prevent the overfitting. Finally, we output the class-probabilities y_i for each of the three action-object affordances with a Soft-max layer. We expand this architecture to a Bayesian model that also predicts the degree of confidence of the prediction. MC-D approximates the posterior distribution $p(w|x,y)$ as the mean of the M forward passes during the test time with a random dropout of neurons, thus it only increases linearly the inference cost. DE require to train M different models with randomly initialized weights to approximate $p(w|x,y)$, so it increases linearly both the training and inference time.

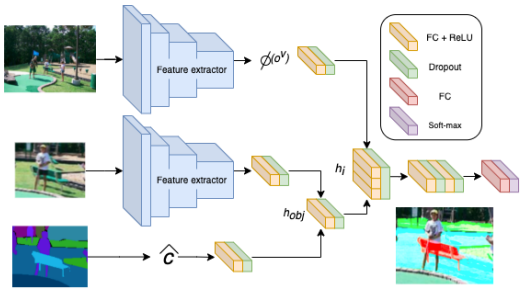

Figure 1. Our architecture combines local and contextual information to predict an affordance for each object-action.

The epistemic uncertainty σ_e is related to the model knowledge and reduces as the training dataset increases, while the aleatoric uncertainty σ_a is associated with the noise inherent to the observations. We calculate them y_i and their average $\hat{y}_i = \frac{1}{N}\sum_{i=1}^{n} y_i$ and following Eq. 1 and 2 [3].

$$\sigma_a = \frac{1}{N}\sum_{i=1}^{N} diag(y_i) - y_i y_i^T \quad \text{(Eq.1)}$$

$$\sigma_e = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)(y_i - \hat{y}_i)^T \quad \text{(Eq. 2)}$$

Experiments

The ADE-Affordances dataset [2] was built on top of the ADE20K [4] dataset for three different actions: *sit*, *run* and *grasp*. It not only indicates if we can take that action, but it also includes 5 different exceptions with social meaning (i.e, object non-functional, the action is dangerous, there is a physical obstacle) that allow the model to reason about the acceptance of that action. We compare three different backbones as feature extractors: Resnet-18, Resnet-50 [5] and Mobilenet-v3 [6]. We report the Mean Accuracy with exceptions (mAcc-E) as our main metric. For the Bayesian models, we show the evolution of the components in the covariance matrix with the M and the differences between the epistemic and aleatoric variance and the Expected Calibration Error (ECE) and the Brier Score metrics.

Results

Table 2. Performance on ADE-Affordance Dataset (mAcc-E)

	Sit	Run	Grasp
Deterministic: Baseline	0.428	0.424	0.289
Deterministic: Mobilenet	0.820	0.834	0.860
Deterministic: Resnet-18	0.787	0.796	0.839
Deterministic: Resnet-50	0.819	0.835	0.859
DE M = 5	0.819	0.834	0.859
DE M = 10	0.820	0.834	0.859
DE M = 25	0.821	0.835	0.859
DE M = 50	0.821	0.835	0.860
MC-D = d_r 0.1	0.818	0.834	0.860
MC-D = d_r 0.3	0.821	0.835	0.860
MC-D = d_r 0.5	0.778	0.780	0.798

The Mobilenet configuration obtained the highest performance and improve the baseline **[2]** over a wide margin due to its quick convergence, as Table 1 shows. The model distinguishes properly between positive and firmly negative classes, although in some cases it fails to classify the type of exception. Our results show that the non-deterministic configurations achieve better performance due to the higher generalization capability of the model. Excessive dropout rates d_r decrease the prediction capacity of the model.

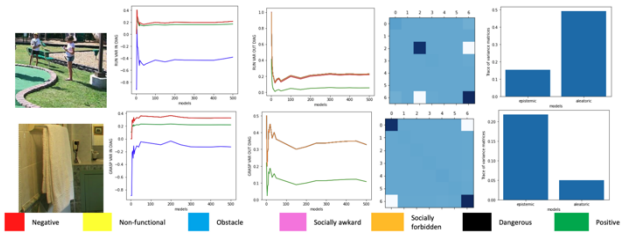

Figure 2. Evolution of the covariance matrix components and comparative between the epistemic (left) and aleatoric (right) variances.

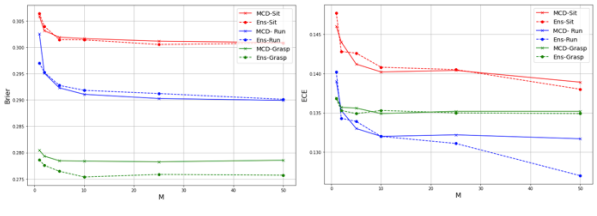

Figure 3. Evolution of the Brier Score and ECE metrics with *M*.

The evolution of the components ins Figure 2 shows the convergence of the variance to the analytical expression as *M* increases, where the components of the most uncertain categories present larger values. It also shows correlations between the different classes and adds a new level of reasoning. For instance, although the model fails to predict that in the *grass* object there is a *physical obstacle* exception, analysing the components of the covariance matrix shows that the model truly doubts between these two categories (components in the trace reflect the variance of that category, while components out of the trace show inter relationship between classes). Following the qualitative analysis, we appreciate that the aleatoric uncertainty is significant in far and blur objects where camera noise is translated to pixels level and that the epistemic uncertainty shows that the sample is out of the distribution.

Finally, the evolution of the Brier-Score and ECE with *M* shows the convergence and that *M=10* is enough to capture the uncertainty of the sample. DE models present higher mAcc-E, ECE and Brier score metrics than MC-D. We suggest that it is because $\boldsymbol{p(w|x,y)}$ does not follow a Bernouilli distribution and it is better approximated by DE rather than MC-D. However, DE is an expensive technique that increases linearly the training time since we need to train *M* different models.

Conclusions

We have proposed a Bayesian method to classify affordances from an RGB image with higher results than previous works. Using Monte-Carlo Dropout and Deep Ensembles techniques, our model combines local with contextual content to predict the class and the aleatoric and epistemic variances which make the prediction more robust and informative.

References

[1] J. J. Gibson, «The ecological approach to visual perception: classic edition,» 2014.

[2] C.-Y. a. L. J. a. T. A. a. F. S. Chuang, «Learning to act properly: Predicting and explaining affordances from images,» 2018.

[3] Y. a. W. J.-H. a. K. B. J. a. P. M. C. Kwon, «Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation,» vol. 142, nº 2020.

[4] B. a. Z. H. a. P. X. a. F. S. a. B. A. a. T. A. Zhou, «Scene parsing through ade20k dataset,» 2017.

[5] K. a. Z. X. a. R. S. a. S. J. He, «Deep residual learning for image recognition,» 2016.

[6] A. a. S. M. a. C. G. a. C. L.-C. a. C. B. a. T. M. a. W. W. a. Z. Y. a. P. R. a. V. V. a. o. Howard, «Searching for mobilenetv3,» 2019.