XI JORNADA DE JÓVENES INVESTIGADORES DEL I3A

Instituto Universitario de Investigación en Ingeniería de Aragón Universidad Zaragoza

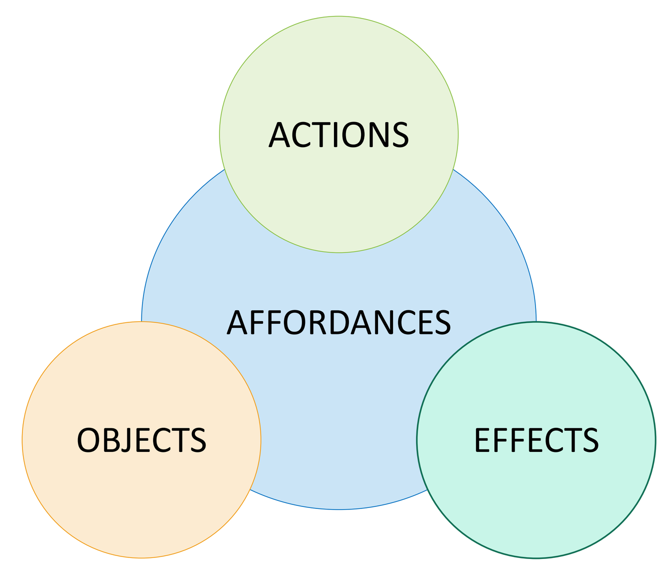
Bayesian classification of affordances from RGB images

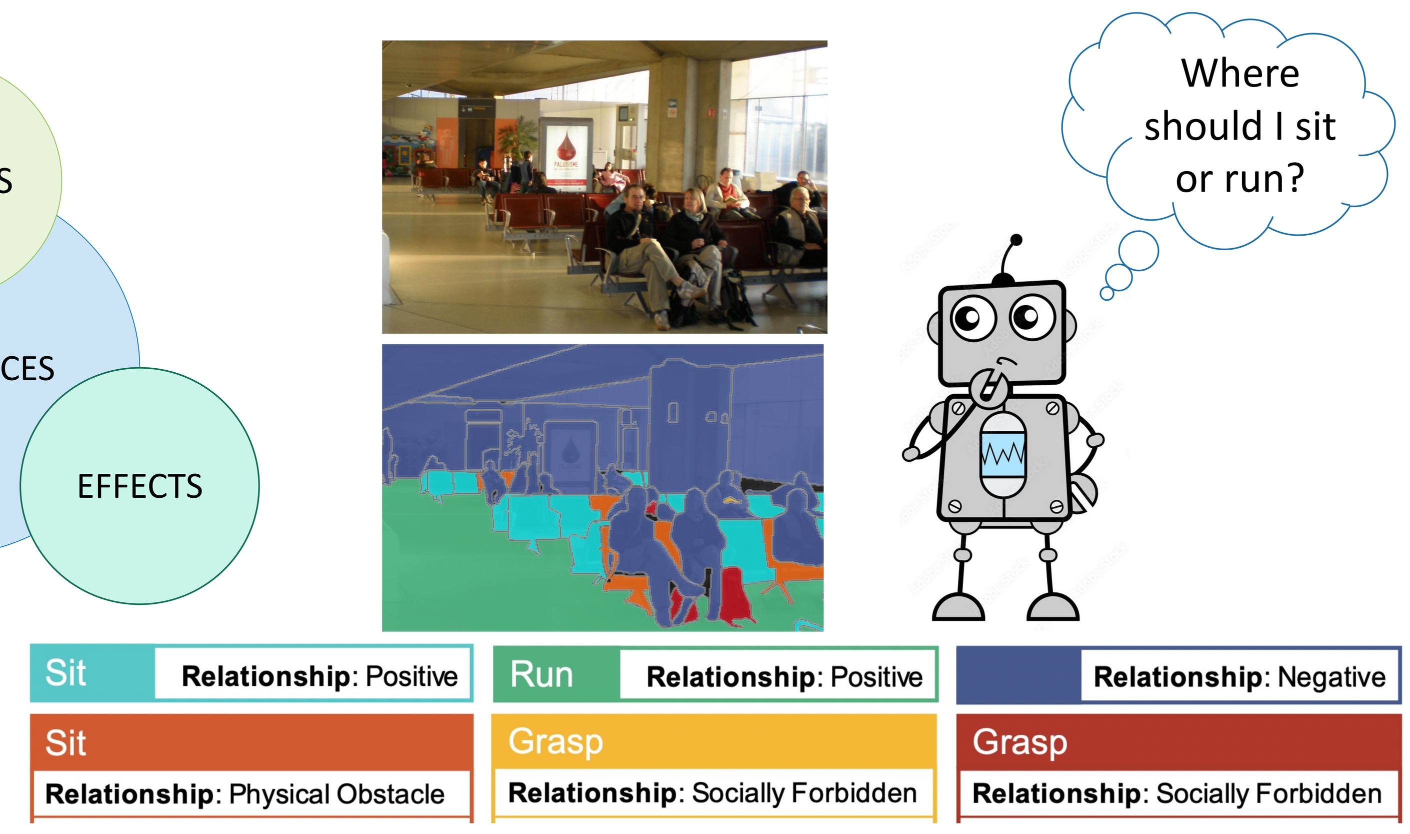
Lorenzo Mur-Labadía, Rubén Martínez-Cantín

Introduction

Affordances are the different action possibilities available in the environment depending on the motor and sensing capabilities of the individual [3]. They relate the objects, the actions and the possible effects of that actions carried on the objects [10]. Based on this, affordance prediction emerges as a powerful tool for autonomous and active agents where we need to understand the content of the scene: a cup is *graspable*, a road is *traversable* and a chair is *sitable* but it can be also *graspable* depending on the context.

Uncertainty estimation helps to discard low-confidence results, reasons about similarities, models noisy observations, analyses sources of uncertainty and serves as a basis for active learning algorithms.





Methods

Deterministic model

We use a CNN architecture as an encoder to extract the semantic features from the object and the global scene and we use the object class $\hat{c}$ of the ground-truth segmentation. Then, we build a Multi-Layer Perceptron with Fully-Connected layers to fuse the vector activations. During training, we incorporate Dropout layers before each FC to prevent overfitting. We compare three feature extractors: Resnet-50 [6], Resnet-18 and Mobilenet-v3 [7].

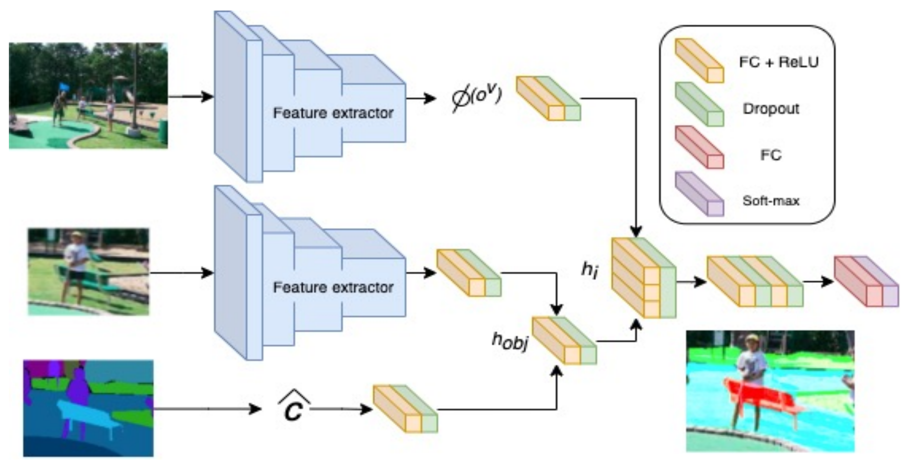


Figure 1: Architecture of our model. The CNN encoder extracts the semantic information from the object and the global scene, which are combined with the object-class

Bayesian model

Bayesian models predict the category and the degree of confidence of the prediction, providing a more robust tool for robotic applications[5, 11, 4, 1, 12]. We compare two alternatives:

- **Monte-Carlo Dropout:** approximates the posterior as the mean of the *N* forward passes during the test time with a random dropout of neurons, but we only train one single model
- **Deep Ensembles:** requires training *M* different models with random initialisation of their weights. Although we increase the training cost linearly, it works better when the posterior distribution does not follow a Bernoulli distribution.

The final prediction is the mean of the samples $\widehat{p}_m = \frac{1}{M}\sum_{m=1}^{M} p_m$

- **Aleatoric uncertainty:** it is associated with the noise inherent in the observations (motion noise, distant objects, boundaries) and it cannot be reduced by collecting more data [9].
- **Epistemic uncertainty:** related to the model knowledge, we reduce it by increasing the dataset [9].

$$\sigma_a = \frac{1}{M}\sum_{m=1}^{M} diag(p_m) - p_m {p_m}^T \qquad \sigma_e = \frac{1}{M}\sum_{m=1}^{M}(p_m - \widehat{p}_m)(p_m - \widehat{p}_m)^T$$

Dataset

We conduct our experiments in the **ADE-Affordances dataset** [2], composed of 44K objects, which was built on top of the ADE20K scenes [13], a popular semantic segmentation dataset. It divides the *object-action* relationships into 7 categories, including exceptions with social meaning:

1. Positive
2. Object non-functional
3. Physical obstacles
4. Socially awkward
5. Socially forbidden
6. The action is dangerous
7. Firmly negative

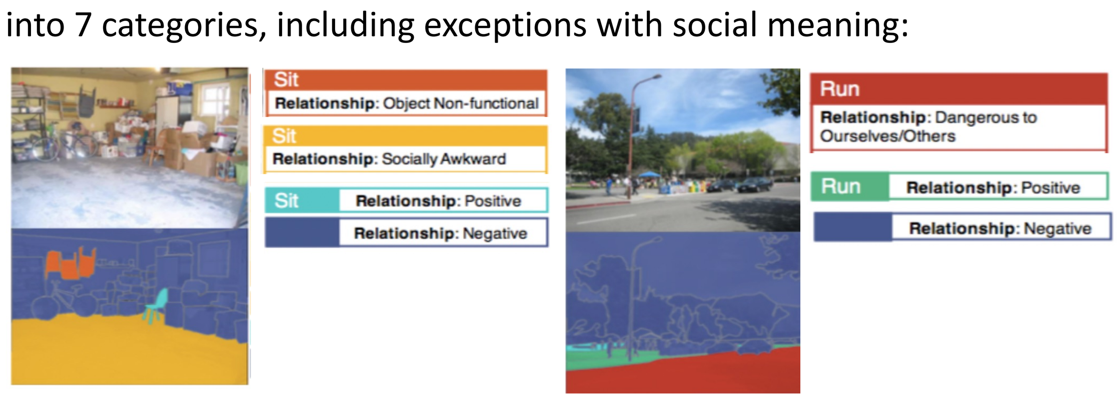


Figure 2: Examples of the ground truth annotations of the ADE-Affordances dataset

Metrics

We compute the **mean accuracy** of the predictions for the deterministic experiment

$$mAcc = \frac{1}{C}\sum_{i=1}^{c}\frac{TP + TN}{TP + FP + TN + FP}$$

For the Bayesian experiments we report:

- **Brier Score (BS):** it measures the accuracy of the model. A perfect accurate model scores BS = 0, while a BS=1 means that the model is completely inaccurate.

$$BS = \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{R}(p_{mi} - \widehat{c}_{mi})^2$$

- **Expected Calibration Error (ECE):** it reports the calibration of the model, expressed as the difference between the confidence of the prediction and its accuracy.

$$ECE = \sum_{l=1}^{L}\frac{B_l}{M}\left|acc(B_l) - conf(B_l)\right|$$

- The evolution of the components of the **covariance matrix:** components in the trace reflect the variance of that category, while components out of the trace show inter-relationship between categories.

Results

1. We surpass the baseline [2] over a wide margin
2. Feature extractors affect significantly the performance, so we select Mobilenet-v3 for the Bayesian experiments.
3. The higher generalization capability of Bayesian models increases the performance.+
4. Deep-Ensembles exceeds Monte-Carlo Dropout [8] since they approximate better the posterior distribution, which does not follow a Bernoulli distribution.
5. The *mAcc*, ECE and BS curves show that we need a minimum number of Bayesian models *M* = 20 to achieve a calibrated and accurate model
6. The components of the covariance matrix also showed convergence with the number of models *M* to the analytical expression. They also show how the model 'doubts' between challenging classes (see Minigolf example)
7. Aleatoric variance is significant in far and blue objects far away from the camera, where the motion is translated to the pixel noise
8. Epistemic uncertainty appeared in uncommon objects out of the data distribution

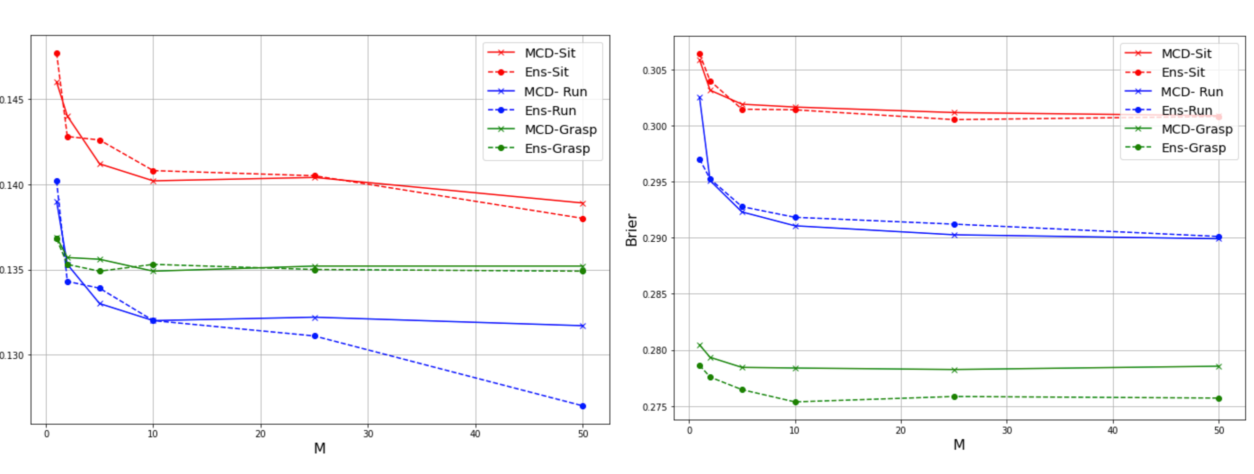


Figure 3: Evolution of the ECE metric and Brier score for the number of models

Figure 4: Qualitative examples and variance maps. Lighter colours mean a high variance in that prediction.

	Sit	Run	Grasp
Deterministic: Baseline	0.428	0.424	0.289
Deterministic: Mobilenet	0.820	0.834	**0.861**
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**	**0.836**	0.859
DE M = 50	**0.821**	0.835	**0.861**
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

Table 1: *mAcc* for the ADE-affordance dataset. Comparative between Bayesian and deterministic models

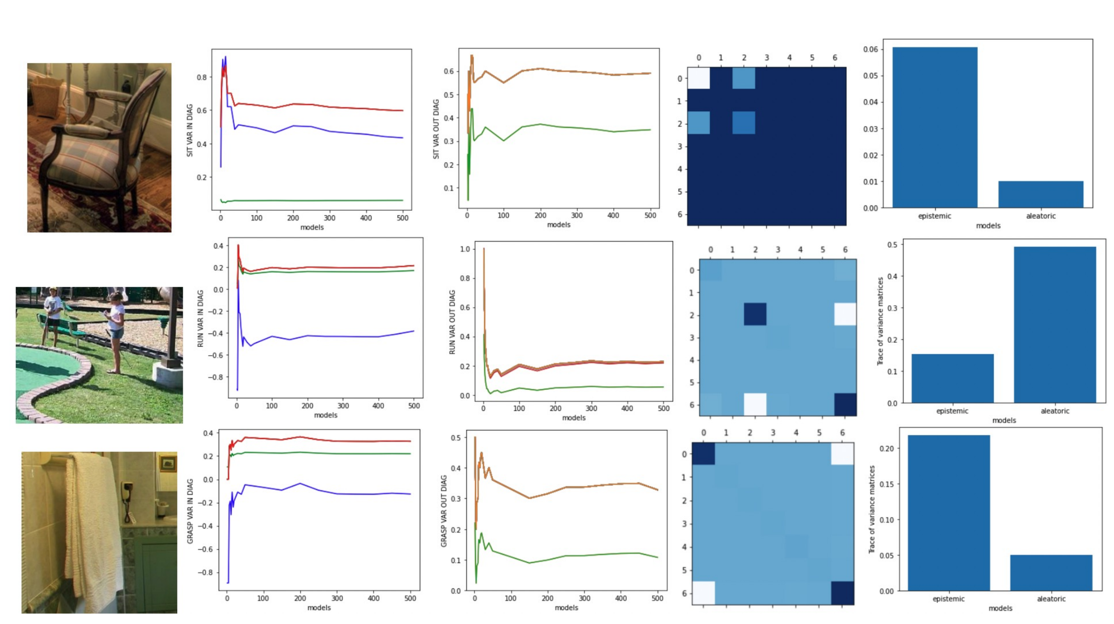


Figure 5: Evolution of the components of the covariance matrix and comparison between aleatoric and epistemic uncertainty

Conclusions

We **propose a Bayesian deep learning model for affordance prediction** directly from image data. We obtain **higher performance over previous works** and we **extend the predictions with the quantification of the uncertainty at no cost in the classification**. Comparing MC-Dropout and Deep-Ensembles as the Bayesian techniques, we show an extensive analysis of the uncertainty estimation with the Brier Score, the ECE, the evolution of the components of the covariance matrix and a comparison of the epistemic and aleatoric uncertainty.

References

[1] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.
[2] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018.
[3] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.
[4] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.
[5] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020.
[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
[7] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.
[8] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.
[9] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, 142:106816, 2020.
[10] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. Learning object affordances: from sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26, 2008.
[11] Buu Truong Phan. Bayesian deep learning and uncertainty in computer vision. 2019.
[12] Kumar Shridhar, Felix Laumann, and Marcus Liwicki. Uncertainty estimations by softplus normalization in bayesian convolutional neural networks with variational inference. *arXiv preprint arXiv:1806.05978*, 2018.
[13] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.