Review of "A Joint Model of Language and Perception for Grounded Attribute Learning"

Vedant Mishra(10792)

February 26, 2013

1 Synopsis

This paper present an approach for joint learning of language and perception models for grounded attribute induction. The perception model includes learning different physical characteristics like colour,shape and a language model includes learning representations of meanings of natural languages.Joint model is the combination of these 2 models. It first relates all the logical meanings of the sentences with the classifiers. Then using the classifier,it determines what sort of objects would be selected.

For eg. if given a sentence "Pick all the yellow blocks", then the robot analyzes the complete sentenses using semantic analysis model and represents the meanings of different words .It then classify the appropriate blocks(in our case all the yellow blocks). Thus the required set of object is the set of all yellow blocks as shown in the right side of Figure 1.

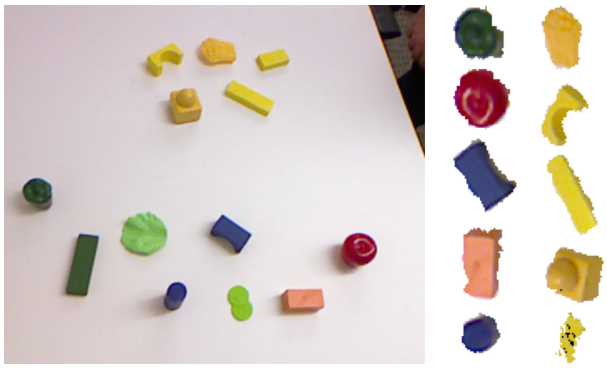

Figure 1: Taken from ref[1]

2 Overview of Methodology

The language and the perception models used in the learning process includes

- **A semantic parsing (language model)** : FUBL algorithm is used for semantic parsing which gives logical meaning(usually high order functions like lamda calculus) to the sentences. This algorithm is used for learning factored Combinatory Categorial Grammar. FUBL parse the trees and creates a log linear model which yields the probability of the parse.For eg. given a sentence " Pick all the yellow blocks", algorithm parse the tree and gives logical meaning to the words like pick,yellow etc.
- **Visual attributes classifier (perception model)** : These includes sets of classifier. Each classifier returns true for all possible objects in object scene.For eg. given a sentence "Pick all the yellow blocks", the robot after getting the logical meaning of the words like yellow, it actually states true for all yellow blocks from the given set of objects.
- **Joint model is the combination of the above 2 models**.Given a natural language sentence x and a given set of scene objects O , this model identifies subset G of object described by O . We evaluate w as set of possible classifier output and z as all the possible logical forms of a sentence.

$$\text{P (G|x,O)} = \sum_{x} \sum_{w} \text{P (G,z,w|x,O)}$$

$$\Rightarrow \qquad \text{P(G,z,w|x,O)} = \text{P(z|x)} * \text{P (w|O)} * \text{P(G|z, w)}$$

Here P(z | x) defines the language model, P(w | O) defines the vision model and P(G|z,w) defines conditional probability. For finding that which set does the object belongs we evaluate arg $\max_G$ P(G|x,O). This gives us the maximum probability set to which that object belongs to.

Model Learning process is done in 2 phases

- **Initialization phase** We will use a small,supervised data to construct initial language and perceptual model. This contains basic words or objects. For any new or unseen word,we use second phase.
- **Assigning unseen words to classifier** : We create a new classifier for each unseen word/object. For eg. for a new word "colour blue", we will create a new classifier. A new lexeme is created by combining the unseen word with a logical constant. Adding a new classifier would result in reestimating of language and perception parameter. For reestimating the parameters, we use expected maximization(EM) algorithm in which we maximize the marginal probabilities.

3 Experiments and Results (Taken From ref[1])

Two features of every objects are extracted from the given set of objects(input) using kernel descriptors.First is the depth value that corresponds to the shape and second is RGB value for the colour. Thus we classify the objects on the basis of their shape and colour. Thus a red triangular block is classified in different group than blue square block.

In order to measure the task performance, we train the input data using 6 objects. The other 6 objects are used as test data set. The system performs well with an average precision of 82%. In order to find the need of joint model, performance of vision, language and joint model are measured and its result is shown in Table 1 ([1]). The table tells us that the precision,recall and F1 Score of the joint model is greater than either of the language or vision(perception) model.

	Precision	Recall	F1 Score
Vision	0.92	0.41	0.55
Language	0.52	0.09	0.14
Joint	0.82	0.71	0.76

Table 1: Results of Experiments (Taken from [1])

4 References

[1] Matuszek fitzgerald zettlemoyer 12icml joint language and perception learning (Main reference Paper).

[2] Kwiatkowski, T., Zettlemoyer, L.S., Goldwater, S., and Steedman, M. Lexical generalization in CCG grammar induction for semantic parsing.

[3] Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. Object detection with discriminatively trained part based models for local image descriptors.

[4] CCG Grammar Wikipedia, http://en.wikipedia.org/wiki/Combinatory_categorial_grammar