Explorations at the vision-language interface. Architectures and mechanisms for language generation and inference from visual data

Albert Gatt, Institute of Linguistics and Language Technology, University of Malta

How do linguistic symbols acquire their reference in the perceptual world and to what extent can artificial systems model this process? For decades, this so-called "grounding" problem has been at the heart of AI research. One way in which this question has recently been addressed is through deep neural network architectures. For example, many image captioning systems use a recurrent network conditioned on visual features extracted from a convolutional neural net. In this way, they make linguistic choices conditional on perceptual data.

However, there are several ways in which this can be done: How early or late should perceptual features be combined with linguistic features? Should the linguistic and perceptual modalities be kept separate as far as possible? Answers to these questions have implications for vision-language generation architectures. This talk will discuss a number of ongoing experiments comparing these various architectural possibilities and their implications for the relationship between language and the world in artificial systems.

We will then broaden the scope of the discussion beyond descriptive captioning, by considering the related problem of Natural Language Inference (NLI, or Textual Entailment). Here, the problem is to determine the semantic relationship between a premise (say, 'The cat is on the mat') and a second sentence (say, 'The cat is outside'). In the standard definition of the task, the second sentence may be entailed by, contradict, or be neutral with respect to the premise. Can a vision-language architecture improve our systems' ability to detect semantic relationships between sentences? In other words, can perceptual information act as useful second modality in the inference task? Some recent experiments suggest that, by incorporating image data in a state-of-the-art NLI architecture, we can observe measurable improvements under certain conditions.

The talk will seek to place these results within the broader context of discussions in AI and the Cognitive Sciences concerning the interface between language and perception.