Throughout my research career, I’ve always been looking for the ‘optimal’ image representation: a representation that captures all relevant information for a given task, including scene composition, 3D information, illumination and other cues; a representation that can easily generalize and adapt to new tasks; a representation that can be updated over time with new information, without forgetting what was learned before; a representation that is explicit in the sense that it can easily be interpreted or explained; a representation, in short, that supports true understanding of the image content, ultimately allowing the machine to reason and communicate about it in natural language. In this talk, I will describe a few recent efforts in this direction.