Imprecise Probability as a Linking Mechanism Between Deep Learning, Symbolic Cognition and Local Feature Detection in Vision Processing

Ben Goertzel

Novamente LLC

Abstract. A novel approach to computer vision is outlined, involving the use of imprecise probabilities to connect a deep learning based hierarchical vision system with both local feature detection based preprocessing and symbolic cognition based guidance. The core notion is to cause the deep learning vision system to utilize imprecise rather than single-point probabilities, and use local feature detection and symbolic cognition to affect the confidence associated with particular imprecise probabilities, thus modulating the amount of credence the deep learning system places on various observations and guiding its pattern recognition/formation activity. The potential application to the hybridization of the DeSTIN, SIFT and OpenCog systems is described in moderate detail. The underlying ideas are even more broadly applicable, to any computer vision approach with a significant probabilistic component which satisfies certain broad criteria.

1 Introduction

One key aspect of vision processing is the ability to preferentially focus attention on certain positions within a perceived visual scene. Another key aspect is the ability for abstract, symbolic cognition, based on various forms of long-term memory, to module visual perception. In principle, these two aspects of vision can be incorporated within a deep learning based vision architecture such as HTM [6],[3] or DeSTIN [1]. In current practice, however, neither of these aspects is a strength of deep learning vision systems. So from the perspective of an integrative approach to AGI, it is interesting to explore the hybridization of deep learning vision systems with other approaches, such as for local feature detectors like SIFT [7], and general cognitive engines like OpenCog [5]. Such hybridization may be carried out in many different ways; here we suggest a novel approach based on imprecise probabilities, which applies to deep learning based vision systems that are probabilistic in their foundations.

In fact the basic idea suggested here applies to any probabilistic sensory system, whether deep-learning-based or not, and whether oriented toward vision or some other sensory modality. However, for sake of concreteness, we will focus here on the case of deep learning and vision.

1.1 Visual Attention Focusing

Since visual input streams contain vast amounts of data, it's beneficial for a vision system to be able to focus its attention specifically on the most important parts of its input. Sometimes knowledge of what's important will come from cognition and long-term memory, but sometimes it may come from mathematical heuristics applied to the visual data itself.

In the human visual system the latter kind of "low level attention focusing" is achieved largely in the context of the eye changing its focus frequently, looking preferentially at certain positions in the scene [2]. This works because the center of the eye corresponds to a greater density of neurons than the periphery.

So for example, consider a computer vision algorithm like SIFT (Scale-Invariant Feature Extraction) [7], which (as shown in Figure 1) mathematically isolates certain points in a visual scene as keypoints which are particularly important for identifying what the scene depicts (e.g. these may be corners, or easily identifiable curves in edges). The human eye, when looking at a scene, would probably spend a greater percentage of its time focusing on the SIFT keypoints than on random points in the image.

The human visual system's strategy for low-level attention focusing is obviously workable (at least in contexts similar to those in which the human eye evolved), but its also somewhat complex, requiring the use of subtle temporal processing to interpret even static scenes. We suggest here that there may be a simpler way to achieve the same thing, in the context of vision systems that are substantially probabilistic in nature, via using imprecise probabilities. The crux of the idea is to represent the most important data, e.g. keypoints, using imprecise probability values with greater confidence.

Similarly, cognition-guided visual attention-focusing occurs when a mind's broader knowledge of the world tells it that certain parts of the visual input may be more interesting to study than others. For example, in a picture of a person walking down a dark street, the contours of the person may not be tremendously striking visually (according to SIFT or similar approaches); but even so, if the system as a whole knows that it's looking at a person, it may decide to focus extra visual attention on anything person-like. This sort of cognition guided visual attention focusing, we suggest, may be achieved similarly to visual attention focusing guided on lower-level cues – by increasing the confidence of the imprecise probabilities associated with those aspects of the input that are judged more cognitively significant.

1.2 Imprecise Probabilities

But what precisely are these "imprecise probabilities" that keep getting mentioned? Broadly speaking an "imprecise probability" is a representation of probability that uses more than one number, and that tries to represent the "uncertainty associated with a certain probability estimate." For instance, one may be very sure that a certain probability is 50%, or one may be only moderately sure

that it's 50%, figuring it might actually be 80% or 20% and one will only know more certainly one gathers more data.

There are multiple forms of imprecise probabilities [4], e.g. one may use

- (probability, confidence) = (s,c) pairs
- (L,U) intervals as introduced by Peter Walley [4], representing lower and upper bounds on the means of probabilities in an envelope
- PLN-style [4] indefinite probabilities of the form ((L,U), b , k), with the interpretation that after k more observations are made, the odds are b that the mean of the estimated distribution describing the event in question will lie in the interval (L,U)

We will speak here in terms of the confidence of an imprecise probability, but this doesnt embody a commitment regarding representation, since essentially any imprecise probability can be used to generate a confidence value. In the case of Walley probabilities, one can simply use the negation interval width, i.e. $c = 1 - (U - L)$, as a confidence value. In the case of indefinite probabilities there is a more complex formula, previously calculated and tested.

We will also assume here that there is a method for taking any calculation done using ordinary single-number probabilities as inputs and outputs, and transforming it into a calculation to be done using imprecise probabilities as inputs and outputs. Straightforward methods of this nature exist for both Walley-style and indefinite probabilities, for example.

2 Using Imprecise Probabilities to Guide Vision Processing

Suppose one has a vision system that internally constructs probabilistic values corresponding to small local regions in visual input (these could be pixels or voxels, or something a little larger), and then (perhaps via a complex process) assigns probabilities to different interpretations of the input based on combinations of these input-level probabilities. For this sort of vision system, one may be able to achieve focusing of attention via appropriately replacing the probabilities with imprecise probabilities. Such an approach may be especially interesting in hierarchical vision systems, that also involve the calculation of probabilities corresponding to larger regions of the visual input. Examples of the latter include deep learning based vision systems like HTM or DeSTIN, which construct nested hierarchies corresponding to larger and larger regions of the input space, and calculate probabilities associated with each of the regions on each level, based in part on the probabilities associated with other related regions.

In this context, we now state the basic suggestion of the paper:

1. Assign higher confidence to the low-level probabilities that the vision system creates corresponding to the local visual regions that one wants to focus attention on (based on cues from visual preprocessing or cognitive guidance)

2. Carry out the vision system's processing using imprecise probabilities rather than single-number probabilities
3. Wherever the vision system makes a decision based on the most probable choice from a number of possibilities, change the system to make a decision based on the choice maximizing the product (expectation * confidence).

2.1 Sketch of Application to DeSTIN

An example of a vision system to which this approach could be applied is Itamar Arels DeSTIN system [1]. Internally to DeSTIN, probabilities are assigned to pixels or other small local regions (according to equations to be detailed below). If a system such as SIFT is run as a preprocessor to DeSTIN, then those pixels or small regions corresponding to SIFT keypoints may be assumed semantically meaningful, and internal DeSTIN probabilities associated with them can be given a high confidence. A similar strategy may be taken if a cognitive system such as OpenCog [5] is run together with DeSTIN, feeding DeSTIN information on which portions of a partially-processed image appear most cognitively relevant. The probabilistic calculations inside DeSTIN can be replaced with corresponding calculations involving imprecise probabilities. And critically, there is a step in DeSTIN where, among a set of beliefs about the state in each region of an image (on each of a set of hierarchical levels), the one with the highest probability is selected. In accordance with the above recipe, this step should be modified to select the belief with the highest probability*confidence.

3 Conceptual Justification

What is the conceptual justification for the approach presented?

One justification is obtained by assuming that each percept has a certain probability of being erroneous, and those percepts that appear to more closely embody the semantic meaning of the visual scene are less likely to be erroneous. This follows conceptually from the assumption that the perceived world tends to be patterned and structured, so that being part of a statistically significant pattern is (perhaps weak) evidence of being real rather than artifactual. Under this assumption, the proposed approach will maximize the accuracy of the systems judgments.

A related justification is obtained by observing that this algorithmic approach follows from the consideration of the perceived world as mutable. Consider a vision system that has the capability to modify even the low-level percepts that it intakes i.e. to use what it thinks and knows, to modify what it sees. The human brain certainly has this potential [2]. In this case, it will make sense for the system to place some constraints regarding which of its percepts it is more likely to modify. Confidence values semantically embody this a higher confidence being sensibly assigned to percepts that the system considers should be less likely to be modified based on feedback from its higher (more cognitive) processing levels. In that case, a higher confidence should be given to those

percepts that seem to more closely embody the semantic meaning of the visual scene which is exactly what we're suggesting here.

4 Particulars of Application to DeSTIN

DeSTIN [1] is a holistic AGI architecture comprising three crosslinked hierarchies, handling perception, action and reinforcement. Here we will be concerned only with the perceptual hierarchy (also called the "spatiotemporal inference network"), which is the best-developed of the three to date.

The hierarchical architecture of DeSTIN's spatiotemporal inference network comprises an arrangement into multiple layers of "nodes" comprising multiple instantiations of an identical cortical circuit. Each node corresponds to a particular spatiotemporal region, and uses a statistical learning algorithm to characterize the sequences of patterns that are presented to it by nodes in the layer beneath it. More specifically,

- At the very lowest layer of the hierarchy nodes receive as input raw data (e.g. pixels of an image) and continuously construct a belief state that attempts to characterize the sequences of patterns viewed.
- The second layer, and all those above it, receive as input the belief states of nodes at their corresponding lower layers, and attempt to construct belief states that capture regularities in their inputs.
- each node also receives as input the belief state of the node above it in the hierarchy (which constitutes "contextual" information)

DeSTIN's basic belief update rule, which governs the learning process and is identical for every node in the architecture, is as follows. The belief state is a probability mass function over the sequences of stimuli that the nodes learns to represent. Consequently, each node is allocated a predefined number of state variables each denoting a dynamic pattern, or sequence, that is autonomously learned. We seek to derive an update rule that maps the current observation (o), belief state (b), and the belief state of a higher-layer node (c), to a new (updated) belief state (b'), such that

$$b'\left(s'\right) = \Pr\left(s'|o,b,c\right) = \frac{\Pr\left(s' \cap o \cap b \cap c\right)}{\Pr\left(o \cap b \cap c\right)}, \tag{1}$$

alternatively expressed as

$$b'\left(s'\right) = \frac{\Pr(o|s',b,c)\Pr\left(s'|b,c\right)\Pr\left(b,c\right)}{\Pr\left(o|b,c\right)\Pr\left(b,c\right)}. \tag{2}$$

[1] This section is pasted with minor modifications from the article *World Survey of Artificial Brains: Part II, Biologically Inspired Cognitive Architectures* published in *Neurocomputing* in December 2010, coauthored by Ben Goertzel and colleagues including Itamar Arel; this section was largely written by Itamar Arel.

Under the assumption that observations depend only on true state, or $\Pr(o|s', b, c) = \Pr(o|s')$, we can further simplify the expression such that

$$b'(s') = \frac{\Pr(o|s') \Pr(s'|b, c)}{\Pr(o|b, c)}, \tag{3}$$

where $\Pr(s'|b, c) = \sum_{s \in S} \Pr(s'|s, c)\, b(s)$, yielding the belief update rule

$$b'(s') = \frac{\Pr(o|s') \sum_{s \in S} \Pr(s'|s, c)\, b(s)}{\sum_{s'' \in S} \Pr(o|s'') \sum_{s \in S} \Pr(s''|s, c)\, b(s)}, \tag{4}$$

where S denotes the sequence set (i.e. belief dimension) such that the denominator term is a normalization factor. One interpretation of (4) would be that the static pattern similarity metric, $\Pr(o|s')$, is modulated by a construct that reflects the system dynamics, $\Pr(s'|s, c)$. As such, the belief state inherently captures both spatial an temporal information. In our implementation, the belief state of the parent node, c, is chosen using the selection rule

$$c = \arg\max_s b_p(s), \tag{5}$$

where b_p is the belief distribution of the parent node. A closer look at eq. (4) reveals that there are two core constructs to be learned, $\Pr(o|s')$ and $\Pr(s'|s, c)$. We show that the former can be learned via online clustering while the latter is learned based on experience by adjusting of the parameters with each transition from s to s' given c. The result is a robust framework that autonomously (i.e. with no human engineered pre-processing of any type) learns to represent complex data patterns, such as those found in real-life robotics applications.

Based on these equations, the DeSTIN perceptual network serves the critical role of building and maintaining a model of the state of the world. In a vision processing context, for example, it allows for powerful unsupervised classification. If shown a variety of real-world scenes, it will automatically form internal structures corresponding to the various natural categories of objects shown in the scenes, such as trees, chairs, people, etc.; and also the various natural categories of events it sees, such as reaching, pointing, falling.

4.1 Enabling Visual Attention Focusing in DeSTIN via Imprecise Probabilities

Given the above outline of DeSTIN, the application of imprecise probability based attention focusing to DeSTIN is almost immediate.

The probabilities $P(o|s)$ may be assigned greater or lesser confidence depending on the assessed semantic criticality of the observation o in question. So for instance, if one is using SIFT as a preprocessor to DeSTIN, then one may assign probabilities $P(o|s)$ higher confidence if they correspond to observations o of SIFT keypoints, than if they do not.

These confidence levels may then be propagated throughout DeSTIN's probabilistic mathematics. For instance, if one were using Walley's interval probabilities, then one could carry out the probabilistic equations using interval arithmetic.

Finally, one wishes to replace Equation 5 above with

$$c = \arg\max_s \left((b_p(s)).\text{strength} * (b_p(s)).\text{confidence} \right), \tag{6}$$

or some similar variant. The effect of this is that hypotheses based on high-confidence observations are more likely to be chosen, which of course has a large impact on the dynamics of the DeSTIN network.

Preliminary results from the application of this approach to the hybridization of DeSTIN and SIFT have been obtained and appear promising, and will be described in a later publication. The integration of DeSTIN and OpenCog, along the lines described here, is planned for implementation during the next year.

References

1. Arel, I., Rose, D., Coop, R.: Destin: A scalable deep learning architecture with application to high-dimensional robust pattern recognition. Proc. AAAI Workshop on Biologically Inspired Cognitive Architectures (2009)
2. Changizi, M.: The Vision Revolution. BenBella Books (2009)
3. George, D., Hawkins, J.: Towards a mathematical theory of cortical micro-circuits. PLoS Comput Biol 5 (2009)
4. Goertzel, B., M. Ikl, I.G., Heljakka, A.: Probabilistic Logic Networks. Springer (2008)
5. Goertzel, B.: Opencog prime: A cognitive synergy based architecture for embodied artificial general intelligence. In: ICCI 2009, Hong Kong (2009)
6. Hawkins, J., Blakeslee, S.: On Intelligence. Times (2004)
7. Lowe, D.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision. pp. 1150–1157 (1999)

Fig. 1. The SIFT algorithm finds keypoints in an image, i.e. localized features that are particularly useful for identifying the objects in an image. The top row shows images that are matched against the image in the middle row. The bottom-row image shows some of the keypoints used to perform the matching (i.e. these keypoints demonstrate the same features in the top-row images and their transformed middle-row counterparts). SIFT keypoints are identified via a staged filtering approach. The first stage identifies key locations in scale space by looking for locations that are maxima or minima of a difference-of-Gaussian function. Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space co- ordinate frame. The features achieve partial invariance to local variations, such as affine or 3D projections, by blurring image gradient locations.

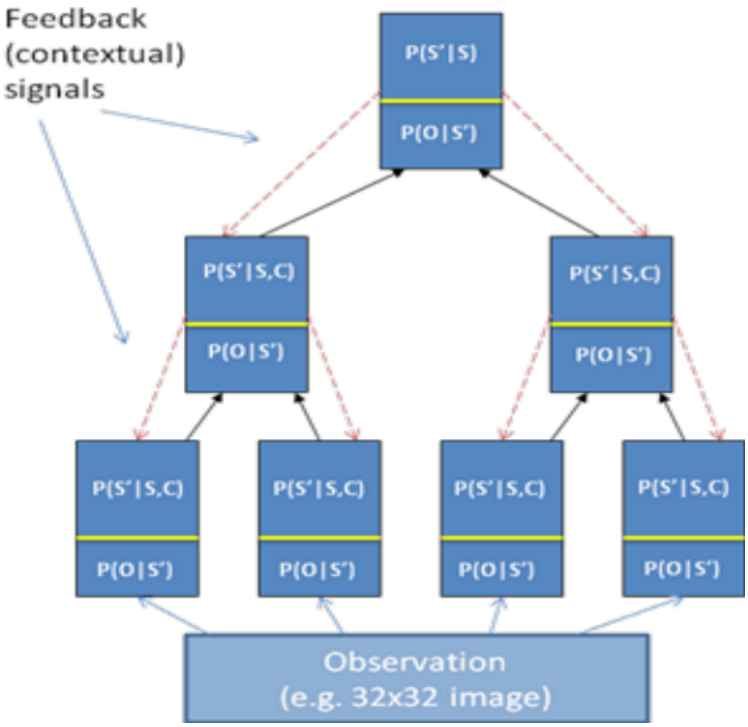

Fig. 2. Small-scale instantiation of the DeSTIN perceptual hierarchy. Each box represents a node, which corresponds to a spatiotemporal region (nodes higher in the hierarchy corresponding to larger regions). O denotes the current observation in the region, C is the state of the higher-layer node, and S and S‘ denote state variables pertaining to two subsequent time steps. In each node, a statistical learning algorithm is used to predict subsequent states based on prior states, current observations, and the state of the higher-layer node.