Structured evidence accumulation in deep neural architectures

Résumé du projet de recherche (Langue 1)

Convolutional Neural Networks (ConvNets) have been used with significant success in recent years for tasks related to image classification and object recognition [GBC16]. Indeed, many state of the art results for these tasks are today obtained by using ConvNets [KSH12, HZRS16, HLvdMW17]. However, ConvNets have several characteristics that inherently limit their effectiveness for more complex vision tasks such as scene understanding and precise object extraction/segmentation: first, sub-sampling loses the precise spatial relationships between higher-level object parts (these relationships are needed for precise recognition) and, second, they cannot extrapolate their understanding of geometric relationships to new viewpoints. While sub-sampling tries to make the neural activities invariant to small translations in the image, a better goal is equivariance, i.e. changes in the image should lead to corresponding changes in neural activations.

Recently, Capsule Neural Networks (CapseNets) have been introduced by Sabour, Frost and Hinton [SFH17, HKW11] as a way to address these drawbacks of ConvNets and also to formalize the intuition about how the human brain processes visual information. They implement two ideas: the first is to package multi-dimensional properties of a feature together into “capsules”, the second is to activate higher-level features by agreement between lower-level features (“routing by agreement”). A “capsule” is seen as a subset of neurons within a layer that outputs both an “instantiation parameter” indicating whether an entity is present within a limited domain and a vector of “pose parameters” specifying the pose of the entity relative to a canonical version. The parameters output by low-level capsules are converted into predictions for the pose of the entities represented by higher-level capsules, which are activated if the predictions agree and output their own parameters. Instead of modeling the co-existence, disregarding the relative positioning, capsule-nets try thus to model the global relative transformations of different sub-parts along a hierarchy according to the equivariance vs. invariance trade-off.

Taking as a starting point the ideas behind ConvNets and the more recent CapseNets, this thesis proposes to explore the more general idea of “structured evidence” for object detection and recognition: while component subnetworks should be triggered/activated by the presence of local patterns in the input, the sequence of decision should be based on the idea that an object is more than a (non-)linear combination of features (as in ConvNets) or a “bag of vector features and poses” (as in CapseNets). The goal is to explore new network architectures that allow the integration of topological and geometrical relationships, the motivation stemming from the fact that current architectures need better modeling of the spatial relationships of the parts. One idea is to use the flow of information through the network to assemble low level evidence into high level concepts meanwhile learning the information pathways which maximize the agreement. Another idea is to use local recurrent connections to evolve specialized units and to implement topological invariance through negative feedback. This approach tries to mimic the apparent hierarchical organization of information processing in the brain, the aim being to understand learning as self-organization (how a network evolves its internal architecture as a graph of units while gradually specializing these units to detect patterns that increase the overall response). We expect such a graph topology approach to increase both the effectiveness and the interpretability of the results.