

Deep Compositional Networks: Overview

- ▶ We described earlier that Deep Networks are sensitive to context and occlusion (money with guitar, penguin with TV). We also discussed how Deep Networks contain internal representations of object parts, although these representations are implicit and hard to categorize. We also discussed pictorial structure and compositional models where objects are represented in terms of parts and spatial relationships between them.
- ▶ This lecture will go further in this direction. We will investigate the use of compositional models that exploit deep network features and have explicit representations of parts and their spatial relationships. In particular, we will show that these types of models are better than standard deep networks for detecting and classifying objects in the presence of extreme occlusion.
- ▶ These models are unsupervised, in the sense that they only know the object identity but do not know the identity/position of the parts, or the different viewpoints (which correspond to the different spatial patterns of the objects). These models are currently only for vehicles on the PASCAL3F+ dataset. We deal with two types of occlusion datasets: (i) PASCAL3D+Occlusion, where occluders are photoshopped onto images in PASCAL3D+, and (ii) CoCo-occluders, where we select images from the CoCo dataset and classify the amount of occlusion.

Occluders and CAPTCHAs: Human Observers

- ▶ Psychophysics studies (Hongru Zhu et al. Cog Sci. 2019) show that humans can classify objects correctly despite extreme occlusion, but deep networks make many misclassifications. A simple compositional model does better than deep networks, showing similar error patterns to human observers (but still works worse than humans).
- ▶ In these studies, we train the compositional models on unoccluded data but test it on occluded data. Why? This is because we want to show that these models can learn to generalize outside their training domain. Also the number of ways that an object can be occluded is exponential (N possible occluders in M possible positions) so there will not be enough training data for the model. In other words, the model must be able to generalize to deal with occluders.
- ▶ This problem can be illustrated by CAPTCHA's which are designed to be ambiguous and very hard for AI systems to read correctly (hence their use for distinguishing between humans and bots). D. George et al. (Science 2017 – HANDOUT) showed that compositional models (similar to those described in this lecture) were able to perform well on CAPCHA's even though they were trained on a limited amount of data (but deep networks worked much worse). The challenge of reading CAPCHA's is particularly hard in situations where there are multiple possible interpretations based on bottom-up cues, but where consistency in the image rules out almost all of them (e.g., the simuli AA gives bottom-up cues for letters A and V, but consistency in the image means that the best interpretation is two A's).

Dictionaries of Visual Concepts and Spatial Patterns

- ▶ We use clustering algorithms to learn a dictionary of visual concepts (see earlier lecture). This clustering can be thought of as a complex variant of Hebbian learning (L. Valiant's Circuits of the Mind). Technically, this is like the EM Algorithm (for mixtures of Gaussians, or for von Mises-Fisher).
- ▶ These visual concepts can be used to encode the object. So that at each lattice position (at level four) the feature vector of the object can be encoded by a visual concept. This encoding can be "hard" or "soft". Hard encoding means that one, or more, visual concepts are activated and yield a binary coding of the object. Soft encoding means that a few visual concepts are activated with a probability. Visualization (earlier lecture) shows that the visual concepts correspond roughly to semantic parts of the objects (e.g., wheels, or parts of wheels).
- ▶ Objects can be encoded by spatial patterns of visual concepts. But these spatial patterns will vary depending on the viewpoint of the object (e.g., a car seen from the side will have two wheels visible, but a car seen from a three-quarters view will have three wheels visible).
- ▶ This means that an object must be represented by a set of different spatial patterns (one for each distinct viewpoint). These distinct spatial patterns must be learnt in an unsupervised way. This can be done by clustering and using the EM algorithm. Finally, we must learn a generative model for each spatial pattern specifying the probability that a specific visual concept is activated in a specific spatial position.

Compositional Models and Occlusion

- ▶ To summarize the model. Each object is represented as a mixture of models (one for each viewpoint) where each mixture component is a spatial pattern of visual concepts (corresponding to one viewpoint of the object). The visual concepts correspond roughly to parts of the object.
- ▶ It is straightforward to make these compositional models robust to occlusion. We simply specify a probability that the spatial patterns are corrupted by having incorrect visual concepts in some spatial positions. These corruptions correspond to occluders. Observe that it is possible to make this model robust to occluders because we are encoding parts explicitly (by visual concepts) which means that they can be automatically switched on or off (this is not possible for deep networks because they do not have explicit representations of parts, so the occluders cannot be switched off but instead will confuse the deep network).
- ▶ The compositional model described so far has three types of learning; (i) the original deep network learning to learn the weights of the deep network, (ii) the clustering to learn the visual concepts, (iii) the clustering and EM algorithm to learn the spatial patterns (for different viewpoints of the objects).
- ▶ This can be extended to a compositional deep network that is learnt end-to-end and minimizes a loss function that includes several different terms (classification loss for objects, loss for learning the visual concepts, and loss for learning the mixtures of spatial patterns).

Compositional Deep Networks: Results

- ▶ We can compare Compositional Nets and Compositional Deep Nets (trained end-to-end) with Deep Networks on datasets with occlusion – PASCAL3D+Occlusion and CoCoOcclusion.
- ▶ Both compositional methods outperform Deep Networks when there is significant occlusion (even if the Deep Network is given some occluded data to train on). The Deep Network slightly outperform the Compositional Net is there is no occlusion, but the Deep Compositional Net performs as well as the Deep Net in this no-occlusion case.
- ▶ Both the compositional net and compositional deep net are able to locate the occluders, but the compositional deep net does better (the deep network cannot do this task).
- ▶ These compositional nets can be extended to include visual concepts at several different levels of the network, which improves their performance. But note that compositional networks of this type can only work on vehicles (e.g., this approach would not work for humans and animals, because their appearance patterns are too variable and so the clustering algorithms used will not work).