An Inference Network Model for Goal-directed Attentional Selection

Yang Chu (y.chu16@imperial.ac.uk)
Department of Electrical and Electronic Engineering, Imperial College London
South Kensington Campus, London SW7 2AZ, United Kingdom

Dan F. M. Goodman (d.goodman@imperial.ac.uk)
Department of Electrical and Electronic Engineering, Imperial College London
South Kensington Campus, London SW7 2AZ, United Kingdom

Abstract

"Listen to the cello in this symphony!" How can we direct selective attention according to different goals, even in distracting environments which we haven't experienced before? It is an essential cognitive ability of the brain, but remains challenging for machines. We developed a computational model that can identify individual digits in images containing multiple overlapping digits, without ever having seen overlapping digits during training. The goal-driven attentional selection is modelled as inferring the posterior distribution of latent variables (the attended target) in a generative model, conditioned on both sensory input and different semantic goals. A neural network model has been build to efficiently carry out the the inference process by predicting the most likely results, instead of using classic per-sample based iterative optimization methods which may not naturally map onto neural structures. Our model also help to understand how top-down and bottom-up attention are combined during perception in the brain.

Keywords: goal-driven attention; amortized inference; computational models; neural network model

Introduction

The brain can easily switch attention between different conversations and the background music in a noisy cocktail party. It may also mistake a vine branch as a snake while searching in the forest. In both cases, the brain directs selective attention differently to disentangle the compound sensory input, according to the different goals it is pursuing at that moment. This is a critical cognitive skill of humans but the question of the underlying neural mechanisms remains unresolved.

In vision, images of objects typically interfere with each other by occlusion, while sounds interfere by superposition. Models of visual segmentation and saliency typically assume, therefore, that each pixel can be assigned to precisely one object (e.g. Long, Shelhamer, & Darrell, 2015; Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2016). This approach can also be applied to sounds but works poorly at low signal-to-noise levels. However, visual stimuli can also superpose in certain situations such as transparent images, reflections or overlapping line drawings or writing. We decided to focus on modelling the problem of detecting and reconstructing individual hand-written digits in an overlapping mixture of digits, as it encompasses several of these problems. (fig 1)

The capsule network (Sabour, Frosst, & Hinton, 2017) is able to very accurately identify pairs of digits in a mixture, but is specifically trained on images of pairs of overlapping digits. The brain, however, is able to solve this and many other tasks without having ever previously seen examples of these types of images. We therefore added a constraint that our model should be able to carry out the task without having previously seen any pairs of overlapping digits.

We can formulate this problem as a generative model: which set of pixels are most likely to be the result of the presence of a given digit in the image? One approach to solving this would be to use an iterative optimization method such as Markov Chain Monte Carlo (MCMC). However, it is not clear if such sophisticated methods could plausibly be implemented by the brain, and we therefore added a second constraint that our model should be implemented via a purely feed-forward neural network.

The ideal would be a network that could solve the overlapping digits problem having only been trained on images of single digits. This proved too difficult. A straightforward approach that can work based only on seeing isolated digits in the training data is template matching or prototype matching. These methods have been used for top-down models of attention in visual object detection. However, they cannot easily handle deformations of objects, or very diverse classes of objects such as digits which can be drawn in many different ways. We therefore opted to train our model using images of overlapped letter-digit pairs. Only the digits were labelled, and therefore the letters function purely as structured noise. This proved sufficient to train our model to carry out this task.

Methods

We assume that the entangled sensory input signal i is the observable variable of a generative process(fig 2), in which foreground target objects a are sampled based on the object classes x and then combined with background distractions b. Then the goal-directed attentional selection can be understand as:

1. clamp the bottom-up sensory input as i
2. clamp the top-down object class goal as x
3. infer the posterior distribution of latent variable a, which is the perception results of attended foreground object

Classic iterative optimization methods like Markov Chain Monte Carlo (MCMC) can be used in step 3 for each individual sample, but have high computational complexity. Instead,

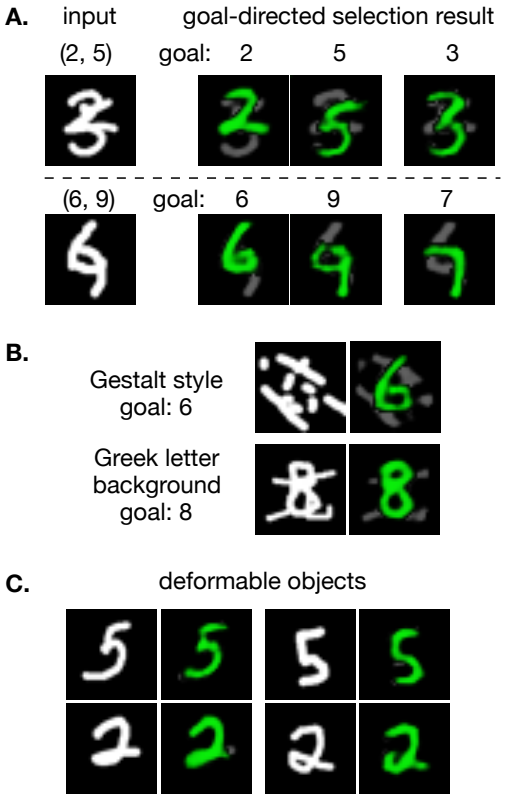

Figure 1: Original input in grey style. Disentangled results after goal-directed selection are reconstructed in green, with the original image overlaid in grey color for visualisation purpose only. **A**: Our model can selectively disentangle individual digits in overlapping images according to different top-down semantic goals. First, similar to human, the model can even find "non-existed" digits like 3 and 7 in these example. Second, note the difference between our selectively disentangle approach and semantic segmentation, which can not handle transparently overlapping objects. Third, note that our model has never seen multiple-overlapping digits during training. **B**: Our model can disentangle the attended object from Gestalt style noisy background or novel background distraction that it has never seen before. **C**: Deformed objects in different shapes are all correctly reconstructed by our model, overperforming simple template matching.

we choose to use a feedforward neural network as an amortized inference(Kingma & Welling, 2013) function here. This inference network learns to map current i and x directly to a single optimal $\hat{a}$ as a regression model. (fig 3) In other words, the network learns to carve and enhance the target's neural representation from the entangled representation of sensory input with current semantic goal.

The background variable b is not constrained in this model, allowing diverse identification results in more general backgrounds, which is comparable to allowing the brain to imagine

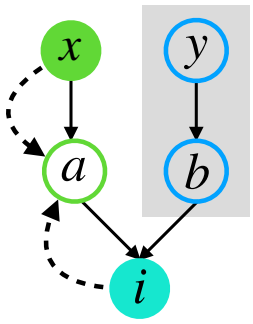

Figure 2: Generative model of entangled sensory stimulus. x is foreground object classes, corresponding to the high-level semantic goal, e.g. digit label; a is foreground object image, corresponding to the attentional selection target, which need to be recovered by posterior inference; b is background object image and y is its prior, both are not necessarily defined; i is the observable compound stimulus. The generative process is indicated by arrows with solid line. Inference process is indicated by arrows with dashed line.

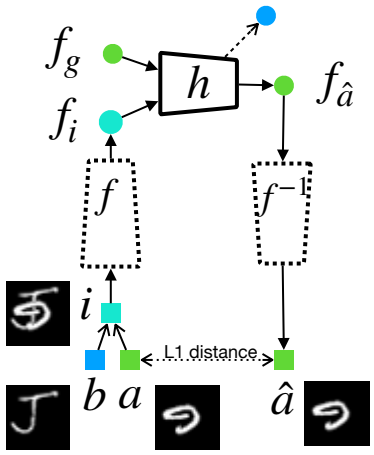

Figure 3: Model Structure. **Networks**: inference network h; encoder f; decoder f^{-1}. **Images**: foreground digit a; background letter b; compound image i as training input; reconstructed digit image $\hat{a}$. **Neural representations**: f_i is original entangled neural representation of image i; $f_{\hat{a}}$ is disentangled neural representation of attended object a; f_g is top-down semantic goal.

a piece of cloud looks as a sheep. Task-relevant prior of the background variable can be added into the inference process to improve prediction accuracy, similar to the case of multi-label recognition in (Sabour et al., 2017).

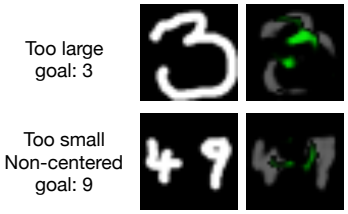

Figure 4: Failure cases

Experiments and Results

We tested our model in a handwritten digit recognition task.(fig 3) The training dataset uses 60k images and labels for digits 0-9 from MNIST (LeCun, Bottou, Bengio, Haffner, et al., 1998) training set, and 54k images for 18 capital English letters[1] from EMNIST(Cohen, Afshar, Tapson, & van Schaik, 2017) without any label. Each training input image is generated by randomly overlaying a digit on top of a letter. A 4-layer convolutional autoencoder network is pretrained to translate between 32×32 images and 1×128 neural representation vector. The inference network (a 4 layer feed-forward network with width [128+10; 500; 500; 128] in each layer) is then trained as a regression model: taking as input the neural representation of overlapped image i and 1×10 one-hot goal vector, output the disentangled neural representation of attended object $\hat{a}$. A simple point estimation is used for current task. The model is trained by back-propagation with L1 loss between $\hat{a}$ and ground-truth a and L1 loss between reconstructed image using $f^{-1}(f_{\hat{a}})$ and the ground-truth foreground image.

When tested in overlapping digits and other noisy background, our model successfully direct selective attention to different objects according to different goals, mimicking human behaviours. Note that the model has never seen multi-digit images or noisy background during training, but generalizes well in these new test cases.

However, this model may fail in cases(fig4) where targeted objects have large variations in size or position. This can be resolved by learning hierarchically disentangled representations for shape, size and position.

Discussion

We designed a fully feed-forward neural network model that is able to selectively separate individual digits from overlapping mixtures of digits according to different goals, even without ever having seen pairs of digits before. Top-down attention at the conceptual level are combined with bottom-up stimulus to disentangle target object by posterior inference. A feed-forward neural network learns to perform this optimization instead of classic numerical method, providing a more biologically plausible solution.

[1] Excluding 8 letters 'BILOSTZY' to avoid uncontrolled confusion.

This approach may also prove fruitful for the cocktail party problem of attending to a single voice in a complex mixture where the signal to noise ratio is near 0 dB, a situation which is particularly challenging for current speech recognition algorithms but relatively easy for normal hearing human listeners.

Acknowledgements

This work was partly supported by a Titan Xp donated by the NVIDIA Corporation, and The Royal Society (grant RG170298).YC is supported by China Scholarship Council and EPSRC.

References

Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017). Emnist: an extension of mnist to handwritten letters. *arXiv preprint arXiv:1702.05373*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3431–3440).

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. In *Advances in neural information processing systems* (pp. 3856–3866).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2921–2929).