Understanding Capsule Networks

Zhenyue Qin

July 31, 2018

1 Introduction

In the end of year 2017, Geoffrey Hinton and his team published two papers that introduced the concepts of capsule networks [SFH17], which differ significantly from traditional neural networks. Additionally, the team also published a paper so called *Dynamic Routing Between Capsules* [SFH17], with which they conceived an algorithm on how to train capsule networks and managed to reach state-of-the-art performance on the MNIST dataset. As a consequence, their capsule networks have demonstrated better performance than traditional CNNs on highly overlapping digits.

Hinton (Figure 1) and his team have been brewing the ideas on capsule networks for many years since (at least) 2011. He has been giving speeches titled *What is wrong with CNNs?*. Therefore, it is exciting for everyone to witness a preliminary outcome for their journey on discovering capsule networks. Expectedly, this became a huge news for the whole deep learning community since it will likely stimulate additional wave of research and cool applications.

Hinton is known to be one of the founders of artificial neural networks and deep learning. It also requires great courage to challenge the work in which he has spent more than half of his life. Nonetheless, although Hinton has used the word *reconstruction of neural networks*, he was not denying all the achievements that traditional neural networks have obtained. In this post, I try to explain the motivations behind Hinton and the basic mechanism of capsule networks. Of course, my understanding can be opposite to what you may think. I warmly welcome all the criticisms and suggestions.

2 CNNs can be problematic

Although CNNs have demonstrated their powerful performances among various applications, their usage of pooling layers concerns people. In order to understand the limits and essential disadvantages of CNNs, we first need to understand how CNNs work.

CNNs consist of convolutional layers. These basic components of CNNs aim to detect important features within the pixels of an image. Lower-level layers detect simple features whereas higher-level layers combine those simpler features to shape more complex features. However, higher-levels interpret lower-level features simply by combining them as weighted sum, without learning their **pose** relationship. That is, how do simpler features shape more complex features in the space.

For example, CNNs may regard Figure 2 as a face, since it contains two eyes, a nose and a mouth, which are basic components of a face. However, they fail to learn how do basic components shape a face. This differs from the learning process of human brains. In other words, human brains can learn a spacial hierarchical relationship between lower level and higher level features.

3 The origin of a capsule

One branch to achieve artificial intelligence is to simulate the structure of our brain, since it is the unequivocal seat of intelligence. The architecture of our brains is exceedingly powerful [Sta17], which is highly hierarchical. However, the number of layers within our brain is much less than the current deep learning architecture, as Figure 3 indicates.

One interesting factor is that a layer of the brains for the most mammal animals consist of mini-columns. That is, a layer of human brains has intrinsically complicated structures. Hinton tries to mimic these mini-columns with the structures so called *capsules*.

Figure 1: Hinton and one of his team members

Figure 2: A fake face contains all the basic components

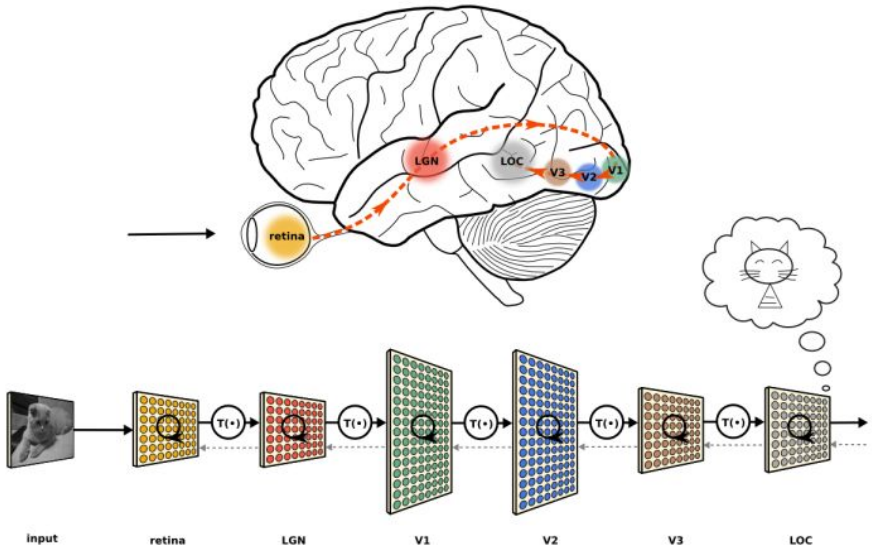


Figure 3: A comparison between human brain and ANN layers (By Jonas Kubilias)

Great! The question now becomes what these capsules do. Before we tackle this question, let's consider some cognitive questions. Hinton believes when we are observe objects, we will naturally build coordinate systems for those being observed objects. For example, we cannot judge whether two Rs in Figure 4 until we have mentally rotated one R.

Another example (perfect one for Australians :-)) is what is the continent in Figure 5? It actually is Africa. However, a lot of people (including Zhenyue) initially thought it was Australia due to the wrong mental coordinate.

As such, Hinton suspects that there exists a group of neurons acting as a whole to represent an object. Only in this way mental coordinates are feasible. This group of neurons is what we think as a capsule.

4 Equivariance and invariance

Hinton also mentions that CNNs are solving wrong questions, mainly coming from the ideas of pooling. Hinton thinks what CNNs do is *invariance*. That is, when the contents of images slightly change, CNNs should still be able to robustly recognize the object. However, we should aim to better represent the object, not to recognize it. This is what Hinton thinks as *equivariance*. With this means, traditional data augmentation, including image rotation and flipping, should become trivial, since capsule networks should realize those are the same objects with different coordinates. Therefore, the foundation of coordinating systems is essential.

5 Routing by agreement

Now the main question lays on *how do higher-level capsules interpret the outputs from lower-level capsules?*. In order to answer the question, we first need to know what are the outputs from a single capsule.

Hinton defines the outcomes from a capsule as instantiation parameters, which is a high-dimensional vector:

- Its norm represents the existence probability of an object.
- Its direction represents the generalized pose of an object, including locations, orientations, sizes, speed, color, etc.

Afterwards, we decide which higher-level capsules to activate based on the agreement from lower-level capsules. For example, if there exist a bunch of lower-level capsules recommend the

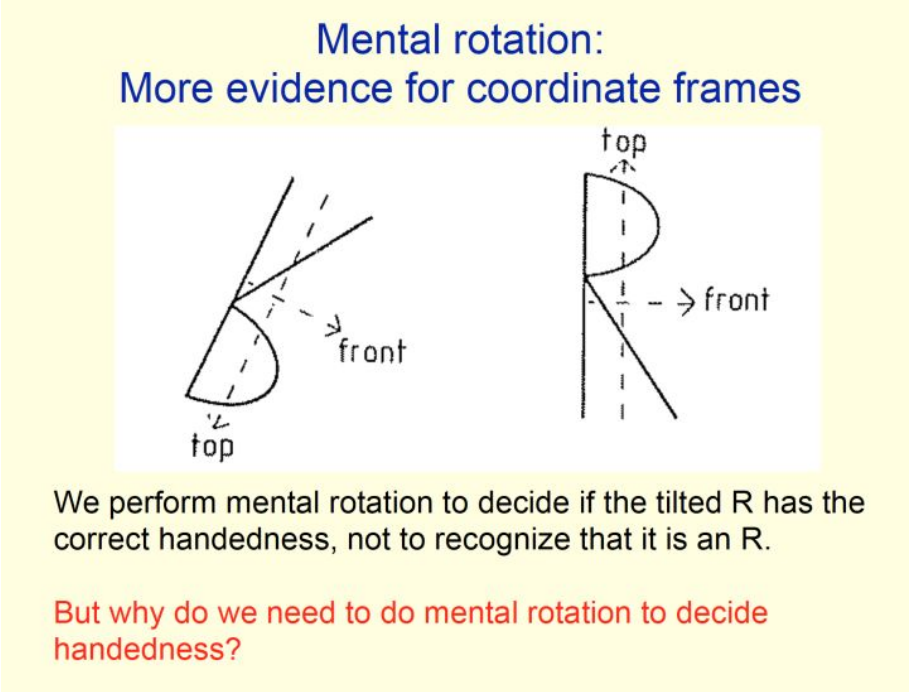
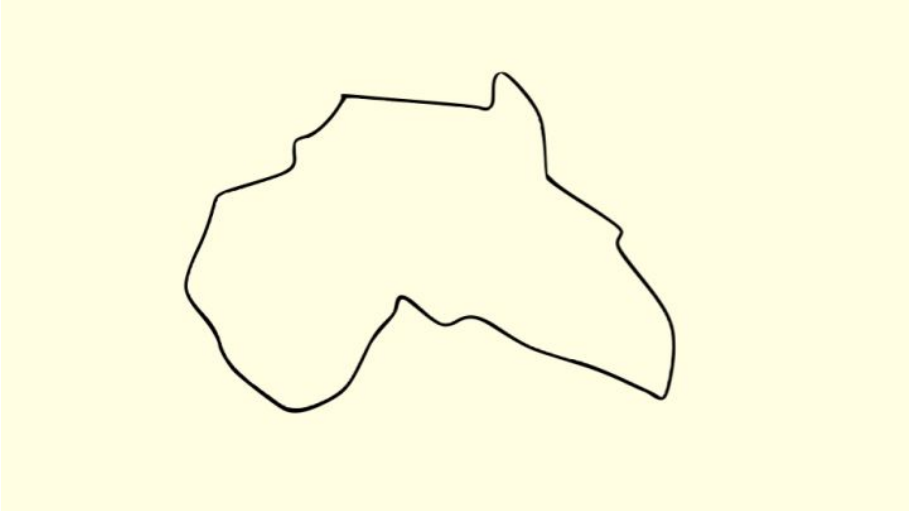
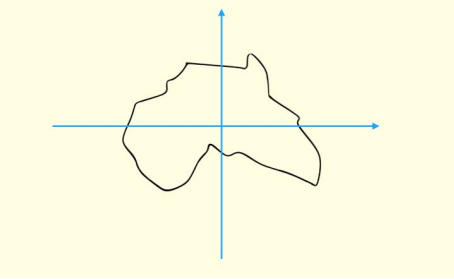


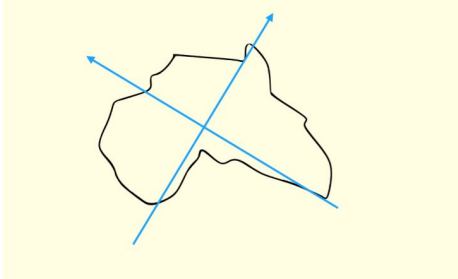
Figure 4: R Handedness (Geoffrey Hinton)



(a) Australia?



(b) Australian coordinate



(c) African coordinate

Figure 5: Different mental coordinates can correspond different representations

existence of a face, then it should have a descent chance of a face indeed existing. This may make the recognition of objects more robots. This agreement is determined by clusters, specifically, the k-means algorithm. This is also why capsule networks are expensive to train. Therefore, a major training of capsule networks locate on *which capsules to send from lower capsules.*

6 Conclusion

In conclusion, capsule networks have the following advantages:

- Reaches high accuracy on MNIST, and promising on CIFAR10
- Requires less training data
- Position and pose information are preserved
- Promising for image segmentation and object detection
- Routing by agreement is great for overlapping objects
- Capsule activations nicely map the hierarchy of parts
- Offers robustness to affine transformation
- Activation vectors are easier to interpret

In contrast, it has the following disadvantages:

- No state of art on CIFAR10
- Not tested yet on large dataset (e.g., ImageNet)
- Slow training, due to clustering

7 Acknowledgments

The author is highly grateful of the work by Aurélien Géron and SIY.Z.

References

[SFH17] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3856–3866, 2017.

[Sta17] Kenneth O Stanley. Neuroevolution: A different kind of deep learning, 2017.