Understanding Hinton’s Capsule Networks. Part I: Intuition.

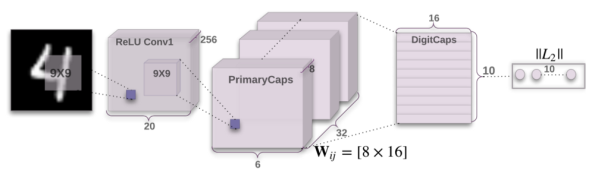
1. Introduction

Last week, Geoffrey Hinton and his team published two papers that introduced a completely new type of neural network based on so-called capsules. In addition to that, the team published an algorithm, called dynamic routing between capsules, that allows to train such a network.

For everyone in the deep learning community, this is huge news, and for several reasons. First of all, Hinton is one of the founders of deep learning and an inventor of numerous models and algorithms that are widely used today. Secondly, these papers introduce something completely new, and this is very exciting because it will most likely stimulate additional wave of research and very cool applications.

In this post, I will explain why this new architecture is so important, as well as intuition behind it. In the following posts I will dive into technical details.

However, before talking about capsules, we need to have a look at CNNs, which are the workhorse of today’s deep learning.



1. 引言

上周，Geoffrey Hinton和他的团队发表了两篇论文，介绍了一种基于所谓的“胶囊”新的神经网络。另外，该团队还发表了一种算法，称为*ynamic routing between capsules*，用来训练该神经网络。

这对于深度学习领域的每一个人来说，都是一个爆炸性的新闻。因为这么几个原因：首先，Hinton是深度学习领域的奠基人之一，同时，也是现在广泛使用的众多模型和算法的创造者。其次，该论文介绍的内容完全是新的，这是非常令人激动的，因为这有可能会激起额外的研究浪潮和开发很多很酷的应用。

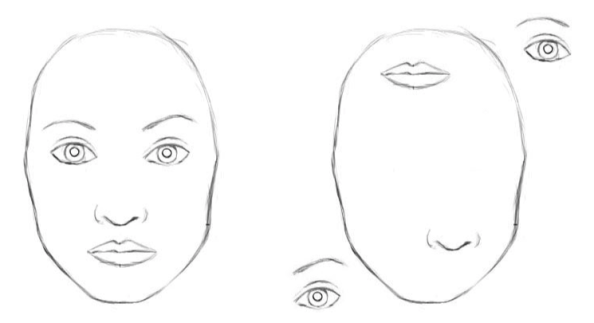
本文中，我将解释为什么这种新的架构是如此重要的，以及它背后的直觉。后面的文章，我将对技术的细节展开深入的分析。

但是，在讨论“胶囊”网络之前，我们需要先来看看卷积神经网络（CNN），它是现在深度学习的主力。

Architecture of CapsNet from the original paper.

CNNs (convolutional neural networks) are awesome. They are one of the reasons deep learning is so popular today. They can do amazing things that people used to think computers would not be capable of doing for a long, long time. Nonetheless, they have their limits and they have fundamental drawbacks.

Let us consider a very simple and non-technical example. Imagine a face. What are the components? We have the face oval, two eyes, a nose and a mouth. For a CNN, a mere presence of these objects can be a very strong indicator to consider that there is a face in the image. Orientational and relative spatial relationships between these components are not very important to a CNN.



To a CNN, both pictures are similar, since they both contain similar elements.

How do CNNs work? The main component of a CNN is a convolutional layer. Its job is to detect important

原论文中CapsNet的架构

卷积神经网络（CNNs）是非常了不起的，促成了今天深度学习如此受欢迎。它们能够做许多神奇的东西，人们曾经长久地认为计算机都无法完成的工作。尽管如此，它们有许多限制和缺点。

让我们考虑一个非常简单且非技术性的问题。想象一张脸，都有什么组成？有一个椭圆形的脸框，两个眼睛，一个鼻子和一张嘴。对于CNN，这些物体的存在就是非常强大的特征，据此判断一张图上有一张脸。对于CNN来说，这些组件之间的取向和相对空间关系不是那么重要。

对于一个卷积神经网络（CNN），这两幅图片是相似的，因为它们都包含相似的元素。

CNNs怎样工作？CNN重要的部分就是卷积层。它的工作就是检测图像像素里的重要

features in the image pixels. Layers that are deeper (closer to the input) will learn to detect simple features such as edges and color gradients, whereas higher layers will combine simple features into more complex features. Finally, dense layers at the top of the network will combine very high level features and produce classification predictions.

An important thing to understand is that higher-level features combine lower-level features as a weighted sum: activations of a preceding layer are multiplied by the following layer neuron’s weights and added, before being passed to activation nonlinearity. Nowhere in this setup there is pose (translational and rotational) relationship between simpler features that make up a higher level feature. CNN approach to solve this issue is to use max pooling or successive convolutional layers that reduce spacial size of the data flowing through the network and therefore increase the “field of view” of higher layer’s neurons, thus allowing them to detect higher order features in a larger region of the input image. Max pooling is a crutch that made convolutional networks work surprisingly well, achieving superhuman performance in many areas. But do not be fooled by its performance: while CNNs work better than any model before them, max

特征。更深层的卷积层（closer to the input）将会学习检测简单的特征，诸如边缘和颜色梯度，而更高的卷积层将会组合这些简单的特征为更为复杂的特征。最后，卷积网络顶层的致密层（dense layers）将会结合出更高水平的特征，产生分类预测。

pooling nonetheless is losing valuable information.

Hinton himself stated that the fact that max pooling is working so well is a big mistake and a disaster:

Hinton: “The pooling operation used in convolutional neural networks is a big mistake and the fact that it works so well is a disaster.”

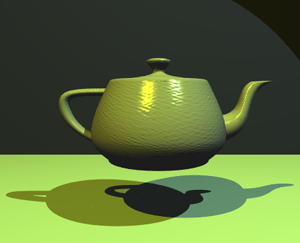
Of course, you can do away with max pooling and still get good results with traditional CNNs, but they still do not solve the key problem:

Internal data representation of a convolutional neural network does not take into account important spacial hierarchies between simple and complex objects.

In the example above, a mere presence of 2 eyes, a mouth and a nose in a picture does not mean there is a face, we also need to know how these objects are oriented relative to each other.

3. Hardcoding 3D World into a Neural Net: Inverse Graphics Approach

Computer graphics deals with constructing a visual image from some internal hierarchical representation of geometric data. Note that the structure of this representation needs to take into account relative positions of objects. That internal representation is stored in computer’s memory as arrays of geometrical objects and matrices that represent relative positions and orientation of these objects. Then, special software takes that representation and converts it into an image on the screen. This is called rendering.



Computer graphics takes internal representation of objects and produces an image. Human brain does the opposite. Capsule networks follow a similar approach to the brain. Source.

Inspired by this idea, Hinton argues that brains, in fact, do the opposite of rendering. He calls it inverse graphics: from visual information received by eyes, they deconstruct a hierarchical representation of the world around us and try to match it with already learned patterns and relationships stored in the brain. This is how recognition happens. And the key idea is that representation of objects in the brain does not depend on view angle.

So at this point the question is: how do we model these hierarchical relationships inside of a neural network? The answer comes from computer graphics. In 3D graphics, relationships between 3D objects can be represented by a so-called pose, which is in essence translation plus rotation.

Hinton argues that in order to correctly do classification and object recognition, it is important to preserve hierarchical pose relationships between object parts. This is the key intuition that will allow you to understand why capsule theory is so important. It incorporates relative relationships between objects and it is represented numerically as a 4D pose matrix.

When these relationships are built into internal representation of data, it becomes very easy for a model to understand that the thing that it sees is just another view of something that it has seen before. Consider the image below. You can easily recognize that this is the Statue of Liberty, even though all the images show it from different angles. This is because internal representation of the Statue of Liberty in your brain does not depend on the view angle. You have probably never seen these exact pictures of it, but you still immediately knew what it was.



Your brain can easily recognize this is the same object, even though all photos are taken from different angles. CNNs do not have this capability.

For a CNN, this task is really hard because it does not have this build-in understanding of 3D space, but for a CapsNet it is much easier because these relationships are explicitly modeled. The paper that uses this approach was able to cut error rate by 45% as compared to the previous state of the art, which is a huge improvement.

Another benefit of the capsule approach is that it is capable of learning to achieve state-of-the art performance by only using a fraction of the data that a CNN would use (Hinton mentions this in his famous talk about what is wrongs with CNNs). In this sense, the capsule theory is much closer to what the human brain does in practice. In order to learn to tell digits apart, the human brain needs to see only a couple of dozens of examples, hundreds at most. CNNs, on the other hand, need tens of thousands of examples to achieve very good performance, which seems like a brute force approach that is clearly inferior to what we do with our brains.

4. What Took It so Long?

The idea is really simple, there is no way no one has come up with it before! And the truth is, Hinton has been thinking about this for decades. The reason why there were no publications is simply because there was no technical way to make it work before. One of the reasons is that computers were just not powerful enough in the pre-GPU-based era before around 2012. Another reason is that there was no algorithm that allowed to implement and successfully learn a capsule network (in the same fashion the idea of artificial neurons was around since 1940-s, but it was not until mid 1980-s when backpropagation algorithm showed up and allowed to successfully train deep networks).

In the same fashion, the idea of capsules itself is not that new and Hinton has mentioned it before, but there was no algorithm up until now to make it work. This algorithm is called “dynamic routing between capsules”. This algorithm allows capsules to communicate with each other and create representations similar to scene graphs in computer graphics.



The capsule network is much better than other models at telling that images in top and bottom rows belong to the same classes, only the view angle is different. The latest papers decreased the error rate by a whopping 45%. Source.

5. Conclusion

Capsules introduce a new building block that can be used in deep learning to better model hierarchical relationships inside of internal knowledge representation of a neural network. Intuition behind them is very simple and elegant.

Hinton and his team proposed a way to train such a network made up of capsules and successfully trained it on a simple data set, achieving state-of-the-art performance. This is very encouraging.

Nonetheless, there are challenges. Current implementations are much slower than other modern deep learning models. Time will show if capsule networks can be trained quickly and efficiently. In addition, we need to see if they work well on more difficult data sets and in different domains.

In any case, the capsule network is a very interesting and already working model which will definitely get more developed over time and contribute to further expansion of deep learning application domain.

This concludes part one of the series on capsule networks. In the second, more technical part, I will walk you through the CapsNet’s internal workings step by step.

Machine LearningDeep LearningGeoffrey HintonArtificial IntelligenceFuture Technology