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

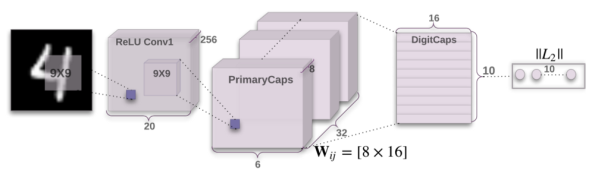
1. Introduction

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。

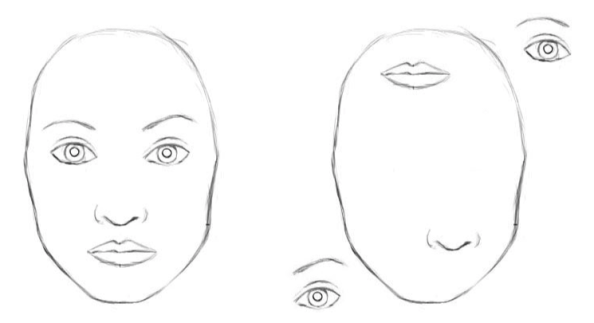
本文中，我将解释为什么这种新的架构是如此重要的，以及它背后的直觉。后面的文章，我将对技术的细节展开深入的分析。

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。

需要了解的重要一点是，较高级别的特征将较低级特征组合为加权和：前一层的激活与下一层神经元的权重相乘并相加，然后传递到激活非线性。 在这个设置中，没有任何地方在构成更高级别特征的简单特征之间存在姿态（平移和旋转）关系。 CNN解决这个问题的方法是使用最大汇集或连续卷积层来减少流经网络的数据的空间大小，因此增加了上层神经元的“视野”，从而使他们能够在输入图像的较大区域检测到更高阶的特征。最大池化是一个使卷积网络工作得非常好的拐杖，在许多领域实现了超人的表现。 但是不要被它的表现所迷惑：虽然CNN比之前的任何模式都工作得更好，但是最大限度的集中却正在失去有价值的信息。

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

Hinton自己说，max pooling运作良好的事实是一个很大的错误和灾难：

Hinton：“卷积神经网络中使用的汇集操作是一个很大的错误，它运行得很好的事实是一场灾难。”

当然，你可以用传统的CNNs来取消最大的池化，但是仍然不能解决关键问题：

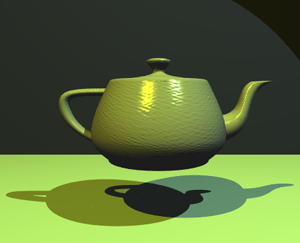
卷积神经网络的内部数据表示不考虑简单和复杂对象之间的重要空间层次。

在上面的例子中，图片中仅存在2只眼睛，嘴巴和鼻子并不意味着有脸，我们也需要知道这些物体是如何相对的。

3.将3D世界硬编码为神经网络：逆向图形方法

计算机图形学涉及从几何数据的内部分层表示来构造可视图像。 请注意，这种表示的结构需要考虑对象的相对位置。 该内部表示作为表示这些对象的相对位置和方向的几何对象和矩阵的阵列存储在计算机的存储器中。 然后，特殊软件将该表示转换成屏幕上的图像。 这就是所谓的渲染。

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

计算机图形学采用对象的内部表示并产生图像。 人脑做相反的事情。 胶囊网络遵循类似于大脑的方法。 资源。

受这个想法的启发，Hinton认为，大脑事实上与渲染相反。 他称之为逆向图形：从眼睛接收到的视觉信息中，他们解构了我们周围世界的层次表征，并试图将其与已经学习的模式和存储在大脑中的关系相匹配。 这是如何识别的发生。 而关键的思想是大脑中物体的表示不依赖于视角。

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.

所以现在的问题是：我们如何建模神经网络中的这些层次关系？ 答案来自计算机图形学。 在3D图形中，3D对象之间的关系可以用所谓的姿势来表示，其本质上是平移加旋转。

Hinton认为，为了正确地进行分类和对象识别，重要的是保持对象部分之间的分层姿态关系。 这是让你理解胶囊理论为何如此重要的关键直觉。 它结合了对象之间的相对关系，并以数字形式表示为4维姿态矩阵。

当这些关系被构建到数据的内部表示中时，模型变得非常容易理解，它所看到的只是另一种以前看到的东西。 考虑下面的图片。 你可以很容易地认识到，这是自由女神像，即使所有的图像从不同的角度显示。 这是因为你脑中的自由女神像的内部表现并不取决于视角。 你可能从来没有见过这些照片，但你仍然立即知道它是什么。



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.

即使所有的照片是从不同的角度拍摄的，你的大脑可以很容易地识别出这是同一个对象。 CNN没有这个能力。

对于一个CNN来说，这个任务确实很难，因为它没有这个3D空间的内建理解，但是对于一个CapsNet来说，这是非常容易的，因为这些关系是明确建模的。 使用这种方法的论文相比于先前的技术水平能够将错误率降低45％，这是一个巨大的改进。

胶囊方法的另一个好处是它能够通过仅使用CNN将使用的一小部分数据来学习获得最先进的性能（Hinton在他着名的关于CNN错误的着名的讲话中提到了这一点）。 从这个意义上说，胶囊理论更接近人脑在实践中所做的事情。 为了学会把数字分开，人脑只需要看几十个例子，最多只有几百个例子。 另一方面，CNN需要成千上万的例子才能取得非常好的成绩，这看起来像是一种暴力方式，显然比我们的大脑差。

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.

这个想法非常简单，没有办法没有人提出过它！ 而事实是，辛顿几十年来一直在思考这个问题。 之所以没有出版物，是因为以前没有技术上的办法。 其中一个原因是，在2012年左右之前，在GPU之前的时代，计算机还不够强大。另一个原因是，没有一种算法可以实现并成功地学习胶囊网络（同样， 人造神经元自1940年代就开始出现，但是直到1980年代中期，反向传播算法才出现并允许成功训练深度网络。

以同样的方式，胶囊本身的想法并不是那么新，而Hinton之前也提到过，但是到目前为止，还没有一种算法能够实现。 这个算法被称为“胶囊之间的动态路由”。 该算法允许胶囊相互通信，并在计算机图形中创建与场景图相似的表示。



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

胶囊网络比其他模型要好得多，可以告诉上下行的图像属于同一类，只有视角不同。 最新的论文使错误率降低了45％。 资源。

胶囊引入了一个新的构建模块，可以在深度学习中使用，以更好地模拟神经网络内部知识表示内部的层次关系。 他们身后的直觉非常简单而优雅。

Hinton和他的团队提出了一种培训这种由胶囊组成的网络的方法，并成功地在一个简单的数据集上进行了培训，实现了最先进的性能。 这是非常令人鼓舞的。

尽管如此，还是有挑战的。 目前的实现比其他现代深度学习模型慢得多。 时间将显示胶囊网络是否可以被快速有效地训练。 另外，我们需要看看它们是否适用于更困难的数据集和不同的领域。

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

无论如何，胶囊网络是一个非常有趣而且已经有效的模型，随着时间的推移，它肯定会得到更多的发展，并有助于深度学习应用领域的进一步扩展。

这就结束了胶囊网络系列的第一部分。 在第二个技术部分，我将逐步介绍CapsNet的内部工作。

关键词：机器学习，深入学习，Geoffrey Hinton，人工智能，未来技术