

Literature Study Report

Dynamic Routing Between Capsules

冯浩哲

2017 年 10 月 29 日

提纲

论文作者是提出反向传播算法并证明了多隐层网络优越的学习能力的 Hinton, 他针对神经网络本身的一些弱点, 提出了一种新的网络结构与连接方式: Capsules with Dynamic Routing. 在这份报告中, 我们主要从以下几个方面对论文内容做初步介绍:

- Capsules with Dynamic Routing 结构的提出背景
- 什么是 Capsules, 什么是 Dynamic Routing, 以及它们是如何工作的
- 新结构如何进行参数优化, 与传统网络结构有何异同
- 新结构比老结构在哪些方面具有优越性, 这些优越性具体体现在哪些方面
- 作者用了什么样的具体例子与结构来解释新结构的优越性
- 新结构有什么缺陷以及下一步的发展方向
- 总结

Capsules with Dynamic Routing 结构提出背景

该结构是一个完全基于仿生学角度所提出的结构, 它的背景是人类视觉的工作原理。一个人类视觉的机制是可以忽略不相关的细节, 而这个机制是通过一系列精心设计的“**Fixation Point**”来完成的。

具体来说, 人的视觉处理系统是一个分层结构, 底层的“**Fixation Point**”以神经元簇的形式聚集, 每一簇负责捕捉相同视觉输入不同位置的信息, 然后将信息传递给自己的父节点, 然后父节点以相似的机理继续向自己的父节点传播, 因此最后到达顶层的信息只是原输入信息的一部分, 然后顶层以高分辨率来呈现这些筛选过的信息, 而这些筛选过的信息就是我们的视觉焦点。

本文正是基于这种视觉传递的树状结构所提出了新的网络结构。

什么是 Capsules, 什么是 Dynamic Routing

与背景介绍中的簇类似, 本文将分层提取特征过程中的每一层划分为很多 **Capsules**, 每一个 Capsules 由一群神经元构成。

同样与背景介绍中的树型结构类似, 每一层的众多 Capsules 应该与上一层的 Capsules 进行连接, 而**层之间的 Capsules 并不连接**。初始模型中, 我们并不知道一个 Capsule 的 Parent Capsule 具体是哪一个, 我们需要在不断训练中来建立这种连接, 使得每一层的 Capsules 都能唯一地 (这里唯一可以理解为连接权重明显占优) 连接到更高层的 Parent Capsules。这个通过更新连接权重从而寻找父节点的过程被称为 **Dynamic Routing**。

具体到细节, Capsules 是这样利用 Dynamic Routing 工作的 (这里仅取第二层以上的 Capsule 进行介绍, 第一层的 Capsules 会在后面具体实例中进行构造分析):

Capsules 是如何工作的

首先, 假设当前层的 $Capsule_j$ 作为父节点与子层的所有 $Capsule_i$ 进行连接, 记子层的 $Capsule_i$ 的输出向量为 u_i 。每一个 $Capsule_j$ 都有一个权重矩阵 $W_{i,j}$, 因此对于每一个 $Capsule_i$ 的输出向量 u_i , 做运算

$$\begin{aligned}\hat{u}_{j|i} &= W_{i,j}u_i \\ s_j &= \sum_i c_{i,j}\hat{u}_{j|i}\end{aligned}$$

以得到输入的加权输出, 中间过程 $\hat{u}_{j|i}$ 表示 u_i 对 $Capsule_j$ 的条件输入。最后对 s_j 作“squashing”处理以使得其范数小于 1:

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} * \frac{s_j}{\|s_j\|}$$

注意这里 $c_{i,j}$ 代表子层 $Capsule_i$ 与当前层 $Capsule_j$ 的连接强度。

如何使用 Dynamic Routing 策略更新连接强度

总体来说, 本文认为如果当前层 $Capsule_j$ 的输出 v_j 与子层 $Capsule_i$ 的条件输入 $\hat{u}_{j|i}$ 的内积 $\langle \hat{u}_{j|i}, v_j \rangle$ 越大, 那么可以认为子层 $Capsule_i$ 与当前层 $Capsule_j$ 的连接强度 $c_{i,j}$ 越大。具体可以通过以下的算法进行计算 (注意我们的前提条件为 $\sum_j c_{i,j} = 1$):

Algorithm 1 Routing Algorithm

Initialize: 引入辅助参数 $\{b_{i,j}\}$, 使用 $c_{i,j} = \frac{\exp(b_{i,j})}{\sum_k \exp(b_{i,k})}$ 计算连接强度; 初始化 $\{b_{i,j}\} \leftarrow 0, \forall i \in layer(l), j \in layer(l+1)$

- 1: 开始权重 $\{c_{i,j}\}$ 更新迭代循环:
 - 2: **for all** r iterations **do**
 - 3: for all capsule j in $layer(l+1)$ calculate: $s_j = \sum_i c_{i,j} \hat{u}_{j|i}, v_j = \text{squash}(s_j)$
 - 4: for all capsule i in $layer(l)$ refresh: $b_{i,j} = b_{i,j} + \langle \hat{u}_{j|i}, v_j \rangle$
 - 5: Update $\{c_{i,j}\}, \forall i \in layer(l), j \in layer(l+1)$
 - 6: **end for**
 - 7: **Return** $\{c_{i,j}\}$
-

新结构如何进行参数优化

在参数优化方面, 文章主要改动了输出形式, 目标函数与优化方法三个方面。

- 输出形式

文章指出, 我们利用 Capsule 输出向量的范数来表示输入存在某类实例的概率, 同时利用输出向量的方向来编码这类实例的具体参数信息, 包括实例位置, 大小等具体信息。

我们以分类任务为例, 文章的最终输出仍然为 Capsule 向量的方式, 而具体分为哪一类依据最后一层输出的 Capsule 向量的范数。比如, 在 MNIST 数据集手写数据分类的任务中, 最后一层由 10 个 Capsule 组成, 最终输出为 10 个向量

$v_k, k = 1, 2, \dots, 10$, 我们用 $\|v_k\|_p$ 来预测输入属于第 k 类的概率。注意到 Capsule 输出向量的范数总是小于 1 的, 因此这种预测在数值范围上是可行的。类实例的具体参数信息, 包括实例位置, 大小等具体信息。

同时, 文章根据 Capsule 输出向量的方向进行解码训练, 重构了整个输入信息并取得了显著成功, 具体内容我们将在新老模型结构对比中详细阐释。

新结构如何进行参数优化

在参数优化方面, 文章主要改动了输出形式, 目标函数与优化方法三个方面。

- 目标函数

文章针对输出形式重构了目标函数, 以分类任务为例, 对于已知分类种类的分类任务而言, 损失函数为:

$$Loss = \sum_{c \in classes} T_c * \max(0, m^+ - \|v_c\|)^2 + \lambda(1 - T_c) * \max(0, \|v_c\| - m^-)^2$$

其中, $T_c = \begin{cases} 1 & \text{if the class of input is } c, \\ 0 & \text{if the class of input isn't } c. \end{cases}$, $m^+ = 0.9, m^- = 0.1, \lambda = 0.5$,

$\|v_c\|$ 为第 c 个 Capsule 的输出向量。

该损失函数一方面使得在正确的分类 c 中的 $\|v_c\|$ 尽可能接近于 m^+ , 同时让不在正确分类的输出尽可能接近于 m^- 。超参数 λ 为错误分类在整个损失函数的权重, 适当调整 λ 可以调整模型训练的侧重点 b 。

新结构如何进行参数优化

在参数优化方面, 文章主要改动了输出形式, 目标函数与优化方法三个方面。

- 优化方法

注意到 Capsule 模型主要有两部分需要优化的参数, 一部分是每一个 Capsule 对于输入 u_i 的权重矩阵 $W_{i,j}$, 另一部分是上下两层的 Capsule 之间的连接强度 $\{c_{i,j}\}$ 。

对于连接强度, 文章使用了 Routing Algorithm 进行优化。而对于权重矩阵的优化部分, 文章仍然与传统深度学习一样, 使用反向传播算法利用损失函数最小化下的梯度来进行优化, 优化过程中固定连接强度。

新结构比传统网络的优越性

在文章中, 作者主要在以下 5 个方面展现了新结构的优越性。

- Dynamic routing 对比 Maxpool
- 对仿射变换的鲁棒性
- 模型参数数目
- 对 Capsule 输出向量进行解码, 还原输入
- 重合目标识别问题

我将详细解释前面 2 部分, 并用 MNIST 手写数字识别例子来阐释后面 3 部分。

新结构比传统网络的优越性

- Dynamic routing 对比 Maxpool

Hinton 认为, Dynamic Routing 机制通过对人类视觉的模拟以对底层的输入特征进行筛选, 从而获取对应标签的目标特征这一过程与传统网络结构中的 Maxpool 结构类似。Maxpool 通过最大值提取令网络保留最重要的特征而略去不重要的细节。但是 Dynamic Routing 机制通过强化连接来直接保留最重要的特征, 通过控制强度来摒弃不重要的特征, 这种理念更加直接。

新结构比传统网络的优越性

- 对仿射变换的鲁棒性

卷积神经网络的一大优越之处在于对于特征的空间位置变化不敏感, 即只要特征存在于输入中且该特征被卷积核所学习到, 那么不管特征位于哪里, 我们都能通过卷积操作激活该特征。但是, 卷积神经网络对于仿射变换 (即对输入 x 做 $Ax + b$ 的变换) 的鲁棒性极差, 这将导致卷积神经网络对于在原图基础上做扭曲, 翻转, 拉伸的图像泛化能力极差, Hinton 将它解释为卷积神经网络对于新视觉点的识别能力弱。

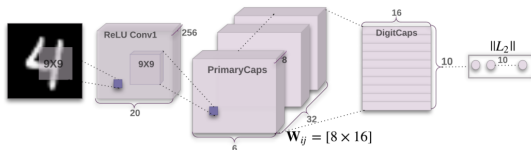
我们一般通过人为对训练数据增加扭曲变化, 并将变化后的数据纳入训练数据集来弥补这项弱点, 但是这种做法平白增添了几倍计算量, 同时我们所施加变化的种类也不够多。Hinton 指出 Capsule 模型可以解决对于仿射变换的鲁棒性问题。以 MNIST 数据集为例, Hinton 对于输入的图片先进行卷积, 使其变成 256 个 Channel 的高维特征块, 然后将这些卷积结果作为第一层 Capsule 的输入, 这种做法使得 Capsule 模型对于仿射变换非常鲁棒。

但是我觉得这种解释是有问题的。因为 Hinton 的做法相当于先对原图施加多种仿射变换再将结果作为输入, 这种做法并不是 Capsule 模型自身的特性, 而卷积网络也照样可以用这种方法来增加对仿射变换的鲁棒性, 因此这个部分我还没有完全弄懂。一个我可以接受的解释是 Hinton 在 Discussion 中提出的传统网络结构需要用 "Normalization" 方法来消除仿射变换的差异, 但是 Capsule 模型中 Capsule 自身的神经元活动会随着仿射变化而变化, 但是这个解释还是有些抽象。

Capsules with Dynamic Routing 结构提出背景
 什么是 Capsules, 什么是 Dynamic Routing, 以及它们是如何工作的
 新结构如何进行参数优化, 与传统网络结构有何异同
 新结构比老结构在哪些方面具有优越性, 这些优越性具体体现在哪些方面
 作者用了什么样的具体例子与结构来解释新结构的优越性
 新结构有什么缺陷以及下一步的发展方向
 总结

模型构造

作者在手写数字识别数据集 MNIST 上构造了 Capsule 模型, 如图所示:



我对这个模型做如下解释:

模型构造

首先, 模型对 $28 * 28$ 尺寸的图片采用 $Kernel = 9 * 9$ 的 256 个卷积核依次进行了卷积运算 + ReLU 激活, 生成了 $20 * 20 * 256$ 的特征图作为 Capsule 模型的输入, 其中卷积核参数通过反向传播学习。

然后, 我们构造第一层初始的 Capsule (Primary Capsules), 采用上面所生成的高维特征图作为输入生成 $32 * 6 * 6$ 个 Capsules, 具体操作中, 我们先采用 $8 * 9 * 256$, $stride = 2$ 的卷积核对输入特征图做卷积, 生成 $6 * 6 * 8$ 的卷积输出, 我们把该输出看作为 $6 * 6$ 个 $8 * 1$ 的向量, 每一个向量是一个 Capsule 的输出, 将上述做法重复 32 次, 就得到了 $32 * 6 * 6$ 个 Capsule 输出, 这样就构成了第一层的 Capsule 输出。

最后, 我们构造由 10 个 Capsules 组成的 DigitCaps, 每一个 Capsule 接受上一层 $32 * 6 * 6$ 个 $8 * 1$ 的 Capsule 输入, 并用矩阵 $W_{i,j} = [8 * 16]$ 做上文所提的变换, 生成 $1 * 16$ 的输出, 输出的大小表示模型推断当前类是输入的分类标签的概率。

模型优越性

通过该模型, 我们继续介绍剩下的三个优越性。

- 模型参数数目

与 Capsule 模型对比的是一个早期应用于 MNIST 数据集 3 层卷积神经网络, 各有 256, 256, 128 层, 每一层有 Kernel 为 5×5 , 步长为 1 的卷积核, 最后由 3 个 size 为 328, 192, 10 的卷积核进行连接。该卷积网络的参数数目为 2.7 亿个, 而 Capsule 模型参数数目则为 530 万个。在 Hinton 的实验中, 后者比前者得出了更好的分类结果, 同时在仿射变换下也达到了更高的稳定性 (79% VS 66%) 但是仅用了约 $\frac{1}{50}$ 的参数。

模型优越性

通过该模型, 我们继续介绍剩下的三个优越性。

- 模型参数数目

但是, 这个结果也有所诟病之处。

首先, 它是与具有全连接层的卷积神经网络进行对比的, 而全连接层现在已经是先进网络结构所避免去用的部分, 去掉全连接层后网络参数也仅仅为 230 万个, 因此很有可能问题出在前者的训练难度上。

其次, 这个 Capsule 模型的 Capsule 参数仅仅为 3.2 万个, 大部分参数仍然用在卷积部分, 也就意味着模型的拟合能力主要依赖于卷积, 这不得不让人思考 Capsule 模型到底起了多大的作用。

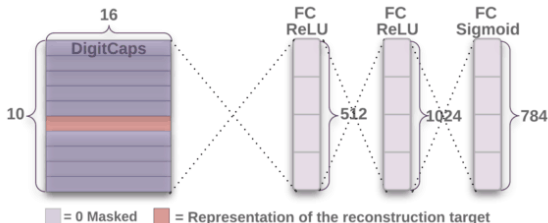
但是, 虽然有上述几个尚不明确的部分, 基于 Capsule 的模型确实在 Dynamic Routing Algorithm 以及反向传播的作用下收敛并表现出优越的预测性质。

模型优越性

通过该模型, 我们继续介绍剩下的三个优越性。

- 对 Capsule 输出向量进行解码, 还原输入

作者为了展示为什么 Capsule 输出向量的方向能够编码实例参数, 对最后的 DigitCaps 的输出向量做了反向的全连接处理, 通过输出向量生成了一张图片, 如图示:















模型优越性

通过该模型, 我们继续介绍剩下的三个优越性。

- 对 Capsule 输出向量进行解码, 还原输入

生成的图像如图所示:

(l, p, r)	(2, 2, 2)	(5, 5, 5)	(8, 8, 8)	(9, 9, 9)	(5, 3, 5)	(5, 3, 3)
Input						
Output						






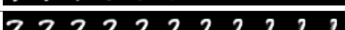
(l, p, r) 依次表示 (label, prediction, reconstruction) 结果, 最后两列是 l, p, r 不一样的地方。这些图片初步证明了 Capsule 的输出向量确实有整合底部的碎片输入并获得目标特征的能力, 同时也表现了 Capsule 模型确实是有效的。

模型优越性

通过该模型, 我们继续介绍剩下的三个优越性。

- 对 Capsule 输出向量进行解码, 还原输入

但是, Hinton 的假设是输出向量的每一个方向都编码了实例信息, 因此为了验证这个假设, Hinton 对输出的 16 维向量的每一个维度进行变换, 并研究变换前后的生成图像的差别。变换的细节为对每一个维度的标量依次从 $[-0.25, 0.25]$ 以 0.05 的间隔选取截距加在原数值上, 生成对比图像如下:

Scale and thickness	
Localized part	
Stroke thickness	
Localized skew	
Width and translation	
Localized part	

结果表明, Capsule 输出向量的 16 个维度概括了最左边一列的 6 类实例特性, 同时对每一个维度数值的修改也直接反映到这些实例特性的变化上来, 这直接证明了**输出向量的每一个方向都编码了实例信息这一重要模型假设。**

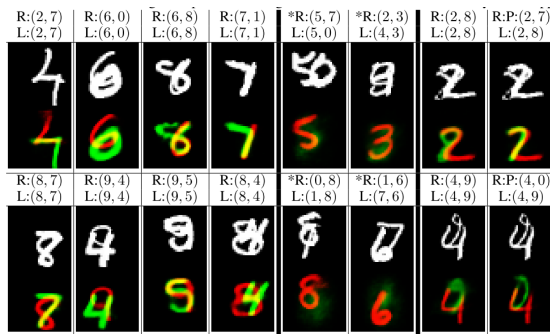
Capsules with Dynamic Routing 结构提出背景
 什么是 Capsules, 什么是 Dynamic Routing, 以及它们是如何工作的
 新结构如何进行参数优化, 与传统网络结构有何异同
 新结构比老结构在哪些方面具有优越性, 这些优越性具体体现在哪些方面
 作者用了什么样的具体例子与结构来解释新结构的优越性
 新结构有什么缺陷以及下一步的发展方向
 总结

模型优越性

通过该模型, 我们继续介绍剩下的三个优越性。

- 重合目标识别问题

Hinton 同时认为, Capsule 结构通过每一层之间互不连接的 Capsules 可以对重叠部分进行良好识别, 并在多分类任务上取得很好的结果。文章在重叠数字的 MNIST 数据集上验证了这一假设, 如图所示:



模型优越性

通过该模型, 我们继续介绍剩下的三个优越性。

- 重合目标识别问题

重合结果中的 L,R,P 依次表示 2 个重叠数字的 Label,Prediction 与 Construction。注意这里 R 仅仅是表示从第 (i, j) 个 Capsule 输出向量进行输出还原, 仅仅当 R, P, L 都出现的时候意味着预测错误。Hinton 的 Capsule 模型对于重合度达 80% 以上的数据达到了 5% 的错误率, 对于其它数据错误率则小于 4%。同时, 图像重构也表现出 Capsule 模型对于信息的高效利用。观察 (R:(5,7),L:(5,0)) 这组图片, 图像没能重构 7 因为模型知道 5 与 0 才是最合适的, 因此重构 7 的信息不足, 同时观察 (R:(0,8),L:(1,8)) 这组图片, 图像对 0 也没有良好重构因为模型已经在 label8 对应 Capsule 输出向量中解释了 0 的信息。

Capsules with Dynamic Routing 结构提出背景
什么是 Capsules, 什么是 Dynamic Routing, 以及它们是如何工作的
新结构如何进行参数优化, 与传统网络结构有何异同
新结构比老结构在哪些方面具有优越性, 这些优越性具体体现在哪些方面
作者用了什么样的具体例子与结构来解释新结构的优越性
新结构有什么缺陷以及下一步的发展方向
总结

缺陷

Hinton 认为, Capsule 模型的主要缺陷是该模型试图对图片中背景的每一部分都进行解释, 这导致模型对于背景种类单一的图片识别任务能取得更好的结果, 但是对于 CIFAR-10 这样背景包含大量不重复出现的各类物体时表现很差。

下一步的发展方向

Hinton 最后指出, Capsule 结构是建立在一个非常强的假设之上的: 在图片的每一个局部区域, Capsule 最多能表现最多一个实例特征。这个假设简单来说就是重要特征是稀疏分布的。这个假设意味着我们可以利用 Capsule 的输出向量的方向来对给定位置编码实例特征。

Hinton 认为, 相比于传统神经网络通过高维特征空间上的点对实例参数进行编码, 这种利用输出向量来对实例特征进行编码更为有效, 而这些特征可以通过矩阵乘法而最终还原到整个空间中去。使用这些 Capsule 模型的基本假设与特性, 我们可以有目的地将这

个模型扩展到更多计算机视觉任务。

总结

在我看来, Hinton 所提出的模型是对层与层之间以及层内神经元之间关系的一个颠覆。Hinton 并没有否定反向传播算法在拟合参数方面的作用, 并在模型中大量使用了反向传播算法, 但是他对于层之间的连接, 以及神经元的输出形式提出了新的模型与连接强度拟合算法。这些模型被证明是有重要作用的, 而同时 Hinton 也将“模型输出的应该是有范数与方向的向量, 而不是一个标量”这一理念通过 Capsule 传递了出来, 并证明了这一理念的有效性。

Hinton 的结构中也包含了很多经典的神经网络结构, 如何将现有神经网络的最新进展融入 Hinton 的模型是一个很广阔的天地。

但是受眼界所限, Hinton 文章的很多理念我还没有完全理解, 因此可能会在某些地方产生自以为是的解释, 希望大家予以补充。

Tips: Hinton 论文中向量长度 (Length) 与向量范数 (Norm) 同义, 不是指向量的维数。