1 摘要

深度学习工具已经在应用机器学习领域引起了巨大的注意。然而，这些用于回归和分类的工具不能捕捉模型不确定性。相比之下，贝叶斯模型提供了给予数学的网络以分析模型不确定性的原因，但是经常会导致过高的计算代价。在这篇论文里，我们发展了一个新的技术性的网络，捕捉深度神经网络训练过程中的dropout。这个理论的一个直接的结果就是给了我们带有衰落N的模型不确定性的工具——从之前被抛弃的已有的模型中提取信息。这缓解了在深度学习中没有牺牲或是计算复杂度或是测试准确度的表示不确定性的问题。我们执行了关于dropout的不确定性的广泛的（extensive）研究。各种网络架构和非线性以回归和分类的任务被评估，以MNIST作为一个例子。通过使用dropout的不确定性来。我们展示了在预测对数似然性和RMSE领域的一个显著的提升，和已存在的最优的方法。

2 介绍

深度学习已经吸引了来自众多领域，如物理学，生物学和建筑学等，的研究院的注意[1]。一些工具，如深度网络，dropout，卷积神经网络和别的已经广泛应用。然而，这是在表示性模型不确定性是很关键的领域。随着最近许多领域向bayesian不确定性转移，对深度学习工具提出了新的要求。

标准的为回归和分类准备的深度学习工具不能捕捉模型不确定性。在分类领域，在pipeline（softmax输出）的最后获得的预测概率经常被错误地解释为模型confidence。（图1）[2]通过对由softmax（实线1b）输出的函数的点估计（实线1a）形成推断[3]：未被纠正的高confidence对于原理训练数据的点。例如，x^*将以概率1被分类成类1.然而，通过softmax（阴影区域1b）的分布（阴影区域1a）更好地反映出远离训练数据的不确定性。

模型不确定性对于深度学习的使用者也是必不可少的。利用在手边的模型confidence[4]，我们可以显示地处理输入和特殊情形。例如，在分类的场景，一个模型可能以高不确定性返回一个结果。在这种情况下，我们可以决定传递这个输入给人用来分类。这可能会发生在邮局，根据他们的zip码排序，或者在带有重要基础设施的原子能基地。[5]不确定性在强化学习中也很重要。利用不确定性的信息，一个agent能够决定何时去exploit，何时去explore环境。进来RL已经利用NN来进行Q函数的近似。这是能够估计一个agent能采取的不同action的质量的函数。ϵ贪婪搜索在agent以一定概率选择它的最好的action和探索其他action时使用。利用对于q函数不确定性估计，例如汤姆逊采样这样的技巧可以用来学习更多。

贝叶斯概率理论提供给我们基础的数学工具，来推出[6]模型不确定性，但是经常会带来比较高的计算代价。令人惊讶的是，一些深度学习工具如贝叶斯模型可能是可以计算的，不改变模型或是最优化。[7]我们将展示dropout及其变体在NN的使用可以被解释为一个众所周知的概率模型——高斯过程的贝叶斯最优化。dropout在深度学习中被用于很多模型，作为一种避免过拟合的方式，并且我们

[1] to name a few ,列举几个

[2] 图一，一个对于理想的二进制分类问题的softmax输入和输出的概览。训练数据在灰色虚线之间的区域被给出。函数点估计以实线显示出。函数的不确定性在阴影区域被展示出。标红色虚线的是离训练数据远的点x^*。忽略函数不确定性，点x^*以概率1被分成类型1

[3] extrapolation,外推法，推断

[4] at hand 在手边，即将到来

[5] infrastructure, 基础设施

[6] reason about,推出

[7] cast,投，抛，计算

的解释建议dropout近似集成了模型权重。我们发展了表示已存在dropoutNN的模型不确定性的工具——提取到目前位置被抛弃的信息。这缓解了在深度学习中表示模型不确定性的问题，而不用牺牲计算复杂度或是测试准确度。 在这篇论文中，我们给出了关于高斯过程和dropout之间关系的完整的理论方法。[8]我们对在回归和分类任务中dropout NN和卷积神经网络拥有的不确定性的属性进行了广泛的探索性的评估。我们比较了在回归任务中不同的模型架构和非线性拥有的不确定性，并且显示出模型不确定性对于分类任务是必不可少的，以MNIST为例。我们接着展示了与已存在的最后的方法，我们的方法在预测对数似然性和RMSE领域的显著提升。最后，我们给出了在一个和深度强化学习中与具体任务相似的实际的任务中，在强化学习设定中模型不确定性的定量评估。

3 相关的研究

众所周知，无限宽（单隐层）带权重的NN，收敛到高斯过程 。这个已知的关系是通过一个有限的论证，并不允许我们把高斯过程的属性容易地转移到NN。有限的带有权重的NN已经被作为贝叶斯神经网络进行了广泛的研究。这些给过拟合提供了健壮性，但是会给inference造成困难，并造成计算代价。variational inference已经被应用到这些模型，但是只带来有限的成功。这些已经被用来拥有新的贝叶斯网络的近似，表现得和dropout一样好。然而，这些模型带来过高的计算代价。为了表示不确定性，在这些模型中，参数的数量是相同模型的两倍。更进一步，他们要求更多时间收敛，并且在已有的技巧上不提升。好的不确定性估计可以更廉价地从一般的dropout模型中获得，但是这可能会导致不必要的计算。一个代替variational inference的方法利用了expectation propatation,并且在VI方法对于RMSE和不确定性估计取得了显著提升。在结果那一节，我们会将dropout和这些方法进行比较，并且显示出dropout在RMSE和不确定性估计的显著提升。

4 droptout作为贝叶斯近似

我们显示了对于一个带有任意深度和非线性的神经网络，如果在每层应用了dropout,在数学上等于一个概率深度高斯过程的近似。我们想强调的是，在这篇文献中，在利用dropout上，没有简化的假设，并且得出的结果可以利用到任何网络架构上，利用dropout，正如在实际应用中一样。更进一步，我们的结果也带来了dropout的其他变体（如drop-connect），乘性高斯噪声。我们展示了dropout目标函数，实际上[9]最小化了一个近似的分布和深度高斯过程（在有限秩方差函数参数上的边缘概率）的后验之间的KL距离。由于空间限制，我们推荐读者参考文献对dropout，高斯过程和variational inference进行深入的了解。结果总结在这里，下一届我们将获得对于dropout NN的不确定性估计。
令$\hat{y}$是L层，loss function为$E(,)$(softmax loss或者欧几里得loss)NN模型的输出我们把NN的$K_i * K_{i-1}$维的带权重矩阵记为W_i,bias向量b_i对于每层$i = 1,...,L$。将观察到的相对于输入x_i的输出记为y_i。输入，输出设为X,Y。在NN最优化过程中，通常会添加一个正规项。我们经常使用L_2正规项，带有一些权重腐蚀λ，导致一个最小化的目标函数（经常被引作代价），
$\mathrm{L}_{dropout} := \frac{1}{N}\sum_{i=1}^{N} E(y_i, \hat{y}_i) + \lambda \sum_{i=1}^{L}(\| W_i \|_2^2 + \| b_i \|_2^2)$
有了dropout，我们在每层layer（除了最后一层）为每个输入点和每个network采样二进制变量。每个二进制变量以概率p_i取值1对于第i层。一个unit被

[8]treatment,治疗，疗法，处理，对待
[9]in effect,实际上

dropped（如它的值设为0）当它相应的二进制变量取值0.我们在后向传递参数导数的时候采样相同的值。
和非概率的NN相比，深度高斯过程在统计上是一个有力的工具，它允许我们建模函数的分布。假设我们被给定一个一下形式的协方差函数：
$K(x,y)=\int p(w)p(b)\sigma(w^Tx+b)\sigma(w_Ty+b)dwdb$
其中，$\sigma(.)$为元素级的非线性函数，$p(w),p(b)$为分布。在第3和4节，在附录里，我们会展示一个一个L层，协方差函数$K(x,y)$的深度高斯过程可以通过在高斯过程的协方差函数的频谱分解[10]的每个部分放置一个variational分布来近似。这个频谱分解将深度高斯过程的每一层映射到显式表达的隐层单元的一层。下面简要介绍。
令W_i为一个（现在是随机的）K_i*K_{i-1}维（对于每一层layer i）矩阵，写$w=\{W_i\}_{i=1}^L$.一个先验，我们令每行W_i根据$p(w)$分布。额外的，假设每个GP层维度K_i向量m_i，深度高斯模型的预测概率（关于有限秩协方差参数w集成），在给定一些准确度参数$\tau>0$的情况下，可以被参数化作为：
$p(y|x,X,y)=\int p(y|x,w)p(w|X,Y)dw,$
$p(y|x,w)=\mathrm{N}(y;\hat{y}(x,w),\tau^{-1}I_D),$
$\hat{y}(x,w=W_1,...,W_L)=\sqrt{\frac{1}{K_L}}W_L\sigma(...\sqrt{\frac{1}{K_L}}W_2\sigma(W_1x+m_1)...)$
后验分布$p(w|X,Y)$是棘手的。[11]。我们使用$q(w)$，一个基于矩阵的分布，该矩阵的列随机设为0，去近似这个难处理的后验。我们定义$q(w)$:
$W_i=M_idiag([z_{i,j}]_{j=1}^{K_i})$
$z_{i,j}\ \ Bernoulli(p_i)fori=1,...,L,j=1,...,K_i-1$ 给定一些概率p_i和矩阵M_i作为variationl参数。二进制变量$z_{i,j}=0$和单元j在第i-1层被dropped作为第i层的输入。variational分布$q(w)$高度多峰[12]，在矩阵W_i推导出强关节关系。（和稀疏频谱高斯过程近似相关）。
我们最小化近似的后验$q(w)$和整个深度高斯过程的后验$p(w|X,Y)$之间的KL距离。KL距离是我们的最小化目标函数
$-\int q(w)\log p(Y|X,w)dw+KL(q(w)\parallel p(w))$
我们将第一项重写为一个和式
$-\sum_{n=1}^N\int q(w)\log p(y_n|x_n,w)dw$
并且通过和一个单独的样本$\hat{w}_n\sim q(w)$来获得一个无偏估计$-\log p(y_n|x_n,w)$。
我们进一步估计等式3的第二项然后获得$\sum_{i=1}^L(\frac{p_il^2}{2}\parallel M_i\parallel_2^2+\frac{l^2}{2}\parallel m_i\parallel_2^2)$带有先验长度比例系数$l$。给定模型准确度$\tau$我们将结果[13]乘以比例系数$\mathcal{L}_{GP-MC}\propto\frac{1}{N}\sum_{n=1}^N\frac{-\log p(y_n|x_n,\hat{w}_n)}{\tau}+\sum_{i=1}^L(\frac{p_il_2}{2\tau N}\parallel M_i\parallel_2^2+\frac{l_2}{2\tau N}\parallel m_i\parallel_2^2)$
设$E(y_n,\hat{y}(x_n,\hat{w}_n))=-\log p(y_n|x_n,\hat{w}_n)/\tau$
我们恢复等式1为一个合适的准确度超参数τ和长度比例l的设定。采样得到的$\hat{w}_n$导致从伯努利分布$z_{i,j}^n$到dropout情况下二进制变量的实现。

4.1 获得模型不确定性

我们接下来推导上述呈现的结果模型不确定性可以通过dropout NN模型获得。
根据附录的2.3，我们的近似预测分布由下式给出：
$q(y^*|x^*)=\int p(y^*|x^*,w)q(w)dw$
其中$w=\{W_i\}_{i=1}^L$是我们对于L层layer模型的随机变量的集合。
我们将执行瞬间匹配和经验性地估计预测分布。更具体地，给定$\{W_1^t,...,W_L^t\}_{t=1}^T$，我们利用$z_i^t=[z_{i,j}^t]_{j=1}^{K_i}$，从伯努利分布$\{z_1^t,...,z_L^t\}_{t=1}^T$采样T个实现的向量的集合。

[10] spectral,频谱的，光谱的
[11] intractable,棘手的，难解；难处理的
[12] multimodal,多峰的
[13] scale,衡量；规模，比例；测量，依比例决定

我们估计
$\mathrm{E}_{q(y^*|x^*)}(y^*) \simeq \frac{1}{T}\sum_{t=1}^{T}\hat{y}^*(x*, W_1^t, ..., W_L^t)$
按照附录里的提议C。我们更喜欢蒙特卡洛估计作为MC输出。实际上，这等于通过网络执行T个随机前向传播然后对结果取平均。
这个结果之前作为模型平均已经被研究过了。我们已经给出了对于这个结果的新的推导，允许我们能够推导基于数学的不确定性估计。Srivastava也经验性地推论MCdropout能够通过对网络的权重取平均进行近似（在测试时间对每个权重乘以概率，参考标准dropout）。
我们以相同的方式估计第二个初始的瞬间：
$\mathrm{E}_{q(y^*|x^*)}((y^*)^T(y*)) \simeq \tau^{-1}I_D + \frac{1}{T}\sum_{t=1}^{T}\hat{y}^*(x^*, W_1^t, ..., W_L^t)^T\hat{y}^*(x^*, W_1^t, ..., W_L^t)$
按照附录的提议D。为了获得模型的预测方差，我们有： $Var_{q(y^*|x^*)} \simeq \tau^{-1}I_D + \frac{1}{T}\sum_{t=1}^{T}\hat{y}^*(x^*, W_1^t, ..., W_L^t)^T\hat{y}^*(x^*, W_1^t, ..., W_L^t) - \mathrm{E}_{q(y*|x*)}(y^*)^T\mathrm{E}_{q(y^*|x^*)}(y^*)$
这个式子等于T个随机前向传播的采样方差加上反向模型准确度。这里y^*是一个列向量，因此和是外积。给定权重腐蚀系数λ以及我们的先验长度比例系数l，我们能够从以下等式得到模型准确度：
$\tau = \frac{pl^2}{2N\lambda}$
我们能够通过等式2的蒙特卡罗集成估计我们的预测对数似然性。这是一个对于模型能够拟合平均值和不确定性多好的估计。对于回归，这是由下式给定的：
$\log p(y^*|x^*, X, Y) \simeq \mathrm{logsumexp}(-\frac{1}{2}\tau \parallel y - \hat{y}_t \parallel^2) - \log T - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \tau^{-1}$
带有一个对数-求和-指数的T项以及通过网络的随机前向传播。 我们的预测分布$q(y^*|x^*)$被期望成高度多峰的，并且上述近似仅仅给出它的属性的一瞥。这是因为近似放置在每个权重矩阵列的variational 分布是双峰的，因此每层权重的joint分布也是多峰的。 dropout NN模型本身是不变的。为了估计预测均值和不确定性，我们仅仅手机了随机前向传播的结果。因此，这个信息可以用于用dropout训练的NN模型。更进一步地，前向传播可以同时计算[14]，导致常数运行时间等于标准的dropout。

[14] concurrently, 同时地