# 3

## 3.2

由于我们不能直接定性分析每个训练样本x被记忆多少，我们通过计算每个样本的average loss. T步SGD更新后的之前样本点x的loss梯度

Loss-sensitivity 损失感知 T步gx的平均值

使用Maclaurin来进行实验

对于realdata部分数据集有高gx拔，randomdata则几乎所有有高-

哪怕一个样本一个类别，在real random的表现也不同

通过计算x’s的GINI，我们可以将gx拔作为训练进程。Gini可以度量频率分布值之间不平等性。 0 完全一样 1 最大性不一样

在上述中random数据（意味着网络需要去记忆）real我们去进一步理解做了啥，这让我们相信gini能对记忆很敏感。因此说明realdata 网络可能不去记忆，或者至少不是以randomdata所需要的方式来记忆

（real random 记忆方式是不同的）

定义了具体类别损失感知

图4说明 对于真实数据，网络能更倾向学习跨类别模式

## 3.3

在不同随机输入或随机标签下能力影响和有效性

3.3.1 容量及数据集在验证性能的影响

Fig5 更优的验证性能需要更高的容量 这与传统概念(容量应该被限制)违背。我们假设高容量允许网络以一种不会干扰真实数据学习的方式来拟合噪点样本。作为对比，如果我们能直接删除噪点样本，产生更小更纯净的数据集，更低容量模型应该能达到最优的性能。

3.3.2 容量及数据集在训练时间上的影响

接下来实验侧重时间收敛 降低容量或增加数据集同事在real random 影响严重

# 4

DNN如何在real random训练上的学习假说的复杂性

为了实现这个目标，我们建立在这样环境，输入空间被划分的不同决策区域的数量反映了所选择的假设的复杂性。这个概````````````````念在精神上类似于函数可以散布随机标签的程度：数据空间中较高的决策边界密度允许分散更多的样本。

因此，我们通过测量模型决策边界周围数据流形上密集点的数量来估计复杂性。 直观地说，如果要随机地从数据分布中抽样点，那么靠近决策边界的一小部分点表明学习假设更简单。

## 4.1 临界采样率CRS

临界样本同上用来预测上述提到的临界边界的密度。临界样本是数据集的子集， 对样本X 在其相近位置，存在至少一个对抗样本点(Adversarial sample)x拔。搜索X拔使用LASS

将Langevin应用与FGSM 此方法我们称为LASS

一阶梯度导数(如FGSM)有下面问题：可能存在这些点，梯度为0 但有很大的二阶导数，能在临域上的决策产生很大影响。通过LASS算法搜索添加的样本点，能够过滤这些点。

## 4.2 整个训练中临界样本

噪声数据和真实数据相比，对深度网络(CNN)的训练来说有更高的临界样本数量。我们测量了验证集，整个训练中的临街样本。F9

F9中 randX randY是噪声数据 所以在真实数据CIFAR上学习的决策面更复杂。同时可以看到，CSR都是逐渐增加并趋于稳定，这说明了网络在三个训练集训练过程中是逐渐学习更加复杂的假说的。

我们接下来实验，评估了在替换了20%-80%输入或类别噪声的MNIST 和CIFAR-10数据集上的性能和CSR。对于randX或randY，都有噪声越大 CSR越大，反应了学习到的预测函数更高层的复杂性。在验证集的准确率上，噪声升越复杂准确率越低(RandX Y都是)，这表明噪声实例对网络学习真实数据的能力有一定的干扰。

更明显的，对于randY （F7b 8b） 神经网络在验证集上准确率达到最高比在训练准确率上达到最高要早。这是因为模型首先在真是数据上学习简单且通用的模式，然后才去拟合噪点(这就导致了验证集的准确率的下降)。进一步的，当模型从拟合真实数据像你和噪点转变时候，CSR突然增加，表明需要学习更复杂的假说来去解释噪点。结合这个观点和3.1结论，我们总结真实数据比噪点更容易拟合。

# 5 学习中正则化的影响

在这里，我们演示正则化降低随机标签数据训练性能的能力，同时保持真实数据的泛化性能。 Zhang等人 （2017）认为明确的正则化不是优秀泛化性能的主要解释，而是基于SGD的优化主要负责它。 我们的发现扩展了他们的观点，并且表明明确的正则化可以大大限制memorization数据的速度，而不会对真实数据产生重大影响。

结果总结在图10中。对于数据集和正则化技术的每个组合，randY（x轴）上的最终训练精度与CIFAR-10上的最佳验证精度相比，在用不同正则化参数（y轴）。 平坦的曲线表明，相应的正则化技术可以减少应用于随机标记时的记忆，同时在干净的验证集上产生相同的验证准确度。 我们的研究结果表明，不同的正规化者在不同程度上针对记忆行为 - 辍学是最有效的。 我们发现，辍学，特别是加上对抗性训练，最好是在不降低模型学习能力的情况下阻止记忆。 图11还示出了就训练损失而言对于所选实验（即所选超参数值）的这种影响。

# 相关工作

我们的工作建立在实验的基础上，并挑战Zhang等人的解释。（2017年）。 我们在噪声数据集的背景下大量使用他们研究DNN训练的方法。 Zhang等人 （2017）表明，DNN可以完美地融合噪声，因此它们的泛化能力不能通过传统的统计学习理论来解释（例如参见（Vapnik＆Vapnik，1998; Bartlett等人，2005））。 我们同意这一发现，但是还表明，DNN中记忆和泛化的程度不仅取决于体系结构和训练程序（包括显式正则化），还取决于训练数据本身8。

我们研究的另一个方向是正规化与记忆之间的关系。 Zhang等人 （2017）认为明示和暗示的正则化（包括SGD）可能无法解释或限制随机数据的破碎。 在这项工作中，我们显示正规化（特别是dropout）确实控制了DNN记住的速度。 这很有趣，因为dropout也可以防止灾难性遗忘（Goodfellow等，2013），因此总的来说，它似乎有助于DNN保留模式。

各种以前的作品分析了DNN泛化能力的解释。 Montavonetal。（2011）使用内核方法分析深度学习架构的复杂性，并发现网络先验（例如由CNN或MLP的网络结构实现）控制每层学习的速度。 Neyshabur等人 （2014）注意到参数的数量并不能控制DNN的有效容量，DNN泛化的原因尚不清楚。 我们通过展示表征能力的影响随噪声水平的变化来补充这一结果。 在探索噪音样本学习动力学的效果与传统相一致时（Bishop，1995; An，1996），我们首先研究了噪声样本的分数与学习算法的其他属性之间的关系，即容量，训练时间和数据集大小。

以前已经提出了多种用于分析DNN训练的技术，包括查看泛化误差，轨迹长度演化（Raghu等，2016），分析与不同层次相关的雅可比矩阵（Wang; Saxe等，2013）或者损失最小化的形状通过SGD（Imetal。，2016; Chaudharietal。，2016; Keskar等，2016）。我们不是通过测量学习假设的损失锐度，而是通过训练和不同数据集和正规化者的学习假设的复杂性来衡量临界样本比。关键样本指的是附近有对抗例子的实际数据点（Szegedy等，2013; Goodfellow等，2014）。敌对的例子主要提供了一些不易察觉的受到错误分类的数据集。 （Miyato等，2015）通过改变预测分布来定义虚拟敌对的例子，从而将定义扩展到未标记的数据点。 Kurakin等人（2016）建议在训练敌对案例时使用此定义，这是我们使用的定义。

两本当代作品深入探索与我们作品有关的主题。 Bojanowski＆Joulin（2017）表明，预测随机噪声目标可以在无监督学习中产生最先进的结果，这证实了我们在3.1节中的发现，特别是图2. Koh和Liang（2017）使用影响参数变化的影响函数来衡量培训期间的变化，如3.2节。 他们探索了这种技术的几个有前景的应用，包括对抗性训练示例的生成。

# 7 结论

我们的实证研究表明，在噪点与真实数据的DNN优化方面存在质的差异，所有这些都支持这样的说法，即使用SGD变体训练的DNN首先使用模式而不是蛮力记忆来拟合真实数据。 然而，由于DNN具有证明能够消除噪点的能力，因此不清楚为什么他们能找到真实数据的泛化能力; 我们认为包括分布式和分层表示在内的深度学习先验可能起着重要作用。 我们的分析表明DNN中的记忆和泛化取决于网络架构和优化器，但也取决于数据本身。 我们希望能够鼓励未来研究数据集的属性如何影响深度学习算法的行为，并建议将DNN能力作为研究目标的数据依赖理解。