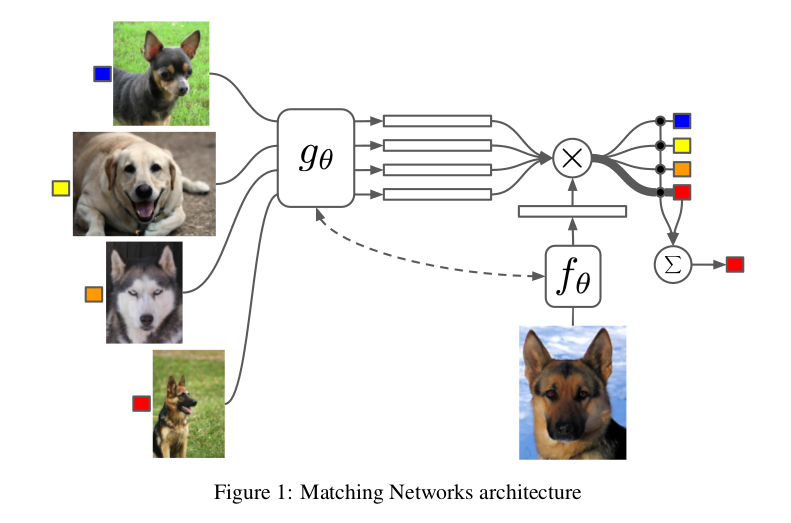
从几个例子学习仍然是机器学习的一个关键挑战。尽管最近在诸如视觉和语言的重要领域取得进展，标准监督深度学习范例不能提供令人满意的解决方案  
用于从小数据中快速学习新概念。 在这项工作中，我们运用想法从基于深度神经特征和最近进展的度量学习增加神经网络与外部记忆。 我们的框架学习一个网络将小标记的支持集和未标记的示例映射到其中标签，避免了微调以适应新类类型的需要。 然后我们定义视觉上的one-shot learning问题（使用Omniglot，ImageNet）和语言任务。 我们的算法提高ImageNet的一次性准确率从87.6％93.2％和从88.0％到93.8％。我们还演示了同一模型对语言建模的有用性在Penn Treebank引入一个一次性任务。

1 Introduction

人类在很少的监督下学习新的概念。 一个孩子可以概括概念的“长颈鹿”从一本书的单张图片 - 但我们最好的深度学习系统需要数百或数千个例子。 这激发了我们感兴趣的设置：“一次性”学习包括从单个标记的示例学习类。  
 深度学习在语言[7]，视觉[13]和语言[16]等领域取得了重大进展，但是对于需要大数据集是臭名昭着的。 数据增加和正则化技术减轻过度拟合在低数据制度，但不解决它。 此外，学习仍然缓慢和基于大数据集，需要使用随机梯度下降进行许多重量更新。 这在我们看来是主要是由于模型的参数方面，其中培训示例需要慢慢学习由模型转化为其参数。

相比之下，许多非参数模型允许新的例子被快速吸收，而不是遭受灾难性遗忘。 在这个家庭中的一些模型（例如，最近的邻居）不需要任何培训，但性能取决于所选择的度量[1]。 以前的工作量度在非参数设置中的学习[18]对我们的模型有影响，我们的目标是并入来自参数模型和非参数模型的最佳特征 - 即快速采集的新例子，同时从常见例子提供了很好的概括。  
 我们的工作的新颖性有两个方面：在建模层面和在培训程序。 我们建议Matching Nets（MN），一种使用近来在注意和记忆方面的进步的神经网络使快速学习。 其次，我们的培训程序是基于简单的机器学习原则：测试和训练条件必须匹配。 因此，为了训练我们的网络做快速学习，我们通过每个类只显示一些示例来训练它，将任务从minibatch切换到minibatch，很像在提交一个新任务的几个例子时如何进行测试。

  
 除了我们在定义一个适合one-shot learning的模型和训练标准的贡献，我们通过定义可以用于对其他方法进行基准的任务作出贡献ImageNet和小规模语言建模。 我们希望我们的结果能够鼓励他人的工作对这个具有挑战性的问题。  
 我们组织了这篇论文，首先定义和解释我们的模型，相关工作。 然后在下面的部分我们简要说明一些相关的工作到任务和我们的模型。 在第4节中，我们描述了我们的一般设置和实验表现出对各种任务和设置的one-shot learning的强大结果。

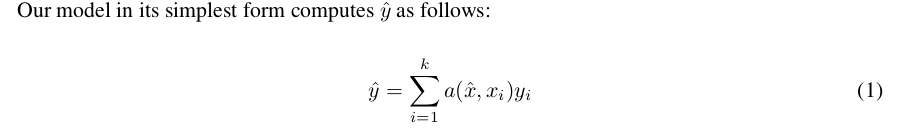
2 Model

我们的非参数方法解决one-shot learning是基于我们的两个组件请在下面的小节中描述。 首先，我们的模型架构遵循最近的神经进展网络增加了内存（如第3节所述）。 给定一个（小）支持集S，我们的模型定义了每个S的函数c S（或分类器），即映射S→c S（。）。 二，我们聘用训练策略是针对来自支持集S的one-shot learning定制的。

2.1 Model Architecture

近年来，许多团体已经研究了如何增加神经网络架构外部存储器和其他组件，使它们更“计算机”。 我们得到灵感从序列到序列（seq2seq）注意[2]，内存网络[29]和指针网络[27]。  
 在所有这些模型中，经常完全可微分的神经注意机制被定义为访问（或读）存储器矩阵，其存储有用信息以解决手头的任务。 典型用途这包括机器翻译，语音识别或问题回答。 更一般来说，这些架构模型P（B | A）其中A和/或B可以是序列（如在seq2seq模型中），或者更多有趣的是为我们，一套[26]。

我们的贡献是将one-shot learning的问题投入到设定的框架内[26]。关键点是，当训练时，Matching Nets络能够产生合理的测试标签未观察到的类，而没有对网络的任何更改。 更准确地说，我们希望从a（小）支持图像标签对的k个实例的集合S = {（x i，y i）} ki = 1到分类器c Sx）其中，给出一个测试示例x，定义了输出y上的概率分布。 我们定义映射S→c S（x）为P（y |x，S）其中P由神经网络参数化。 因此，当给定一个新的支持集的例子S从其中one-shot learning，我们简单地使用参数神经网络由P定义以对每个测试示例x做出关于适当标签y的预测：P（y |x，S）。 一般来说，我们对于给定输入的预测输出类未看见示例x和支撑件集合S变为arg max y P（y |x，S）。



其中x i，y i是来自支持集合S = {（x i，y i）} ki = 1的样本和标签，a是注意机制，我们在下面讨论。注意， 1基本上描述了一个新类的输出支持集中的标签的线性组合。其中注意机制a是内核X×X，则（1）类似于核密度估计器。其中注意机制为零b最远x根据一些距离度量和适当的常数，否则，然后（1）等价于'k-b'个最近邻（虽然这需要注意的延伸机制，我们在第2.1.2节中描述）。因此（1）包含KDE和kNN方法。（1）的另一个视图是其中a作为注意机制，y i作为存储器对应的x i。在这种情况下，我们可以将其理解为一种特定类型的关联内存其中，给定输入，我们“指向”支持集合中的相应示例，检索其标签。  
然而，不像其他注意的记忆机制[2]，（1）是非参数性质：支持集大小增长，所使用的内存也增加。因此，由分类器定义的函数形式c S（x）非常灵活，可以轻松适应任何新的支持集。

2.1.1 The Attention Kernel

我们模型的主要新颖性在于重新解释一个很好研究的框架（神经网络部存储器）做one-shot learning。与度量学习密切相关，嵌入函数 - f和g作为特征空间X的提升，以通过分类实现最大精度函数。 1。

尽管事实上分类策略完全取决于整个支持集P（x，S），我们将余弦相似性应用于“参与”，“点”或简单的嵌入计算最近邻居是近视的，即每个元素x i被嵌入g（x i）独立于支撑组S中的其他元件。此外，S应该能够修改如何我们嵌入测试图像x通过f。  
 我们建议通过一个函数来嵌入集合的元素，该函数将整个集合作为输入S除了x i之外，即g变为g（x i，S）。因此，作为整个支撑集合S的函数，g可以修改如何嵌入x i。当一些元素x j非常接近x i时，这可能是有用的，其中 情况下，改变嵌入x i的函数可能是有益的 - 这一点的一些证据在第4节中讨论。我们使用双向长短期存储器（LSTM）[8]来编码x i在支持集S的上下文中，被视为序列（参见附录中的更准确的定义）。

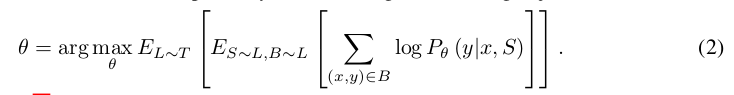
第二个问题可以通过具有注意力的整个集合S的输入的LSTM来固定等于x：



其中f（x）是输入到LSTM的特征（例如，从CNN导出）（常数每个时间步长）。 K是LSTM的展开步长的固定数目，g（S）是其上的集合我们参加，嵌入g。 这允许模型潜在地忽略中的一些元素支持集S，并且将“深度”添加到关注的计算中（更多细节参见附录）。

2.2 Training Strategy

在上一小节中，我们描述了将支持集映射到分类的Matching Nets络函数，S→c（X ）。我们通过修改set-to-set范例来实现这一点注意，其中所得到的映射具有形式Pθ（x，S），注意θ是参数（即前面描述的嵌入函数f和g）。  
 必须仔细选择训练过程，以便在测试时匹配推理。我们的模型必须与支持集S执行良好，其中包含训练期间从未见过的类。  
 更具体地，让我们将任务T定义为在可能的标签集L上的分布考虑T均匀地加权至少几个唯一类（例如，5）的所有数据集合每个类别的示例（例如，最多5个）。在这种情况下，从任务T采样的标签集合L将会通常具有5至25个实例。  
 为了形成“情节”以计算梯度并更新我们的模型，我们首先从T中采样L（例如，L可以是标签集{cats，dogs}）。然后我们使用L来对支持集合S和批次B进行采样（即，S和B都是猫和狗的实例）。Matching Nets然后被训练最小化预测在支持集S条件下的批次B中的标签的错误形式的元学习，因为训练过程明确地学习从给定的支持集学习以最小化批次的损失。更准确地说，Matching Nets训练目标如下：



训练θ与eq。 2产生在从不同的采样S〜T时工作良好的模型新标签的分布。 至关重要的是，我们的模型不需要对它具有的类进行任何微调从未看到由于其非参数性质。 显然，当T从远离我们的T发散采样以学习θ，模型将不工作 - 我们在4.1.2节进一步讨论这一点。

3 Related Work

3.1 Memory Augmented Neural Networks

最近涌现的模型超越了固定向量的“静态”分类到他们的类已重塑了当前的研究和工业应用。 这是最大的在巨大的在语言[7]，翻译[23,2]或学习程序等各种任务中采用LSTM [8][4,27]。 允许更多表达模型的关键组件是引入“内容”  
[2]中的“基于计算机的注意力”以及“神经图灵机”[4]或“类似计算机”的构  
内存网络[29]。 我们的工作采用了[21]的金属学习范式，LSTM学到了从按顺序呈现的数据中快速学习，但我们将数据视为一组。 one-shot learning我们在Penn Treebank上定义的任务[15]涉及评估技术和模型[6]，我们在第4节中讨论。

3.2 Metric Learning

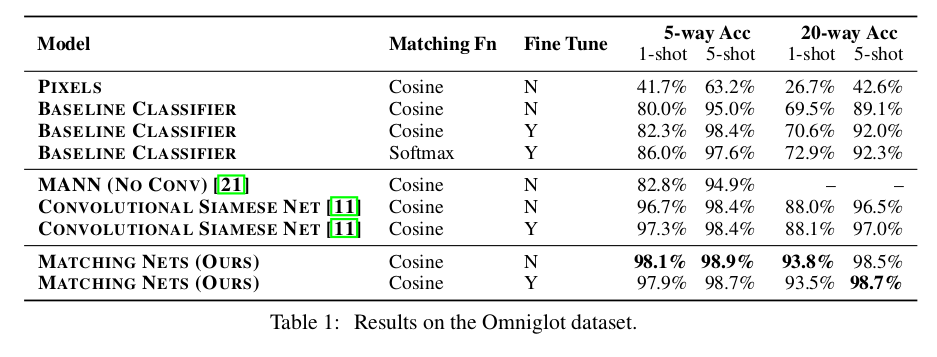
如第2节所讨论的，在基于内容的关注，基于内核的最近之间有许多链接  
邻居和度量学习[1]。 最相关的工作是邻域组件分析（NCA）[18]和随访非线性版本[20]。 损失与我们非常相似，除了我们使用整个支持集S而不是更易于一次性的成对比较学习。 后续工作以深层卷积暹罗[11]网络的形式包括了很多更强大的非线性映射。 其他损失包括集合的概念（但使用较少强大的度量）[28]。  
 最后，[14]中的one-shot learning的工作是鼓舞人心的，也为我们提供了无价的Omniglot数据集 - 称为MNIST的“转置”。 其他作品使用零点学习ImageNet，例如 [17]。 然而，ImageNet上没有太多的单向文献，我们希望通过我们的基准和任务定义在以下部分修改。

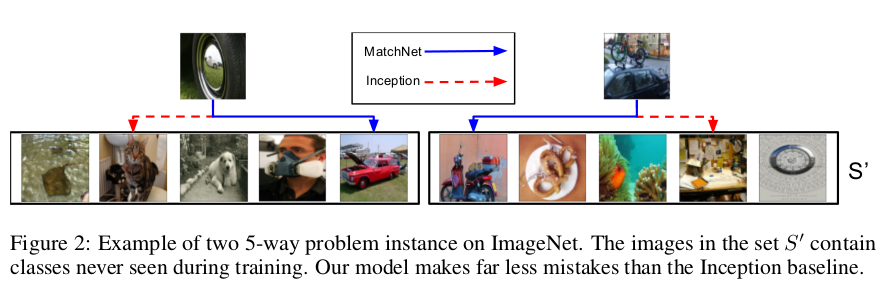
4 Experiments

在本节中，我们描述了许多实验的结果，比较了我们的Matching Nets络强基线模型。我们所有的实验都围绕着同样的基本任务：一条N路k-shot学习任务。每个方法提供来自N的每一个的一组k个标记的示例以前没有训练过的类。然后任务是对不相交的批次进行分类未标记的示例转换为这N个类中的一个。因此，这个任务的随机性能1 / N。我们将多个替代模型（作为基线）与Matching Nets络进行了比较。  
 让我们介绍一些符号。 L表示我们只使用的标签的保留子集一次性。因此，除非另有规定，训练总是开启= L，并在单次模式下测试L。  
 我们对三个数据集运行一次性实验：两个图像分类集（Omniglot [14]和ImageNet [19，ILSVRC-2012]）和一个语言建模（Penn Treebank）。实验上三个数据集包括在复杂性，大小和模态方面的多样性的一组质量。

4.1 Image Classification Results

对于视觉问题，我们考虑了四种基线：匹配原始像素，匹配来自最先进的分类器（基线分类器），MANN [21]和我们的卷积暹罗网的重新实现[11]。基线分类器被训练到将图像分类为训练数据集中存在的原始类别中的一个，但不包括N类，以便不给予它不公平的优势（即，训练以分类in = L）。我们占用这个网络，并使用最后一层（softmax之前）的最近邻居的特征匹配，计算机视觉中常用的策略[3]，这已经取得了优异的成果跨越许多任务。在[11]之后，卷积暹罗网络在相同或不同的训练任务的原始训练数据集，然后最后一层用于最近邻匹配。  
 我们还尝试使用仅从L采样的支持集合S进一步微调特征。这个产生大规模过度拟合，但鉴于我们的网络高度规则化，可以产生额外的收益。注意，即使在微调时，设置仍然是一次性的，因为每个类只有一个示例L。



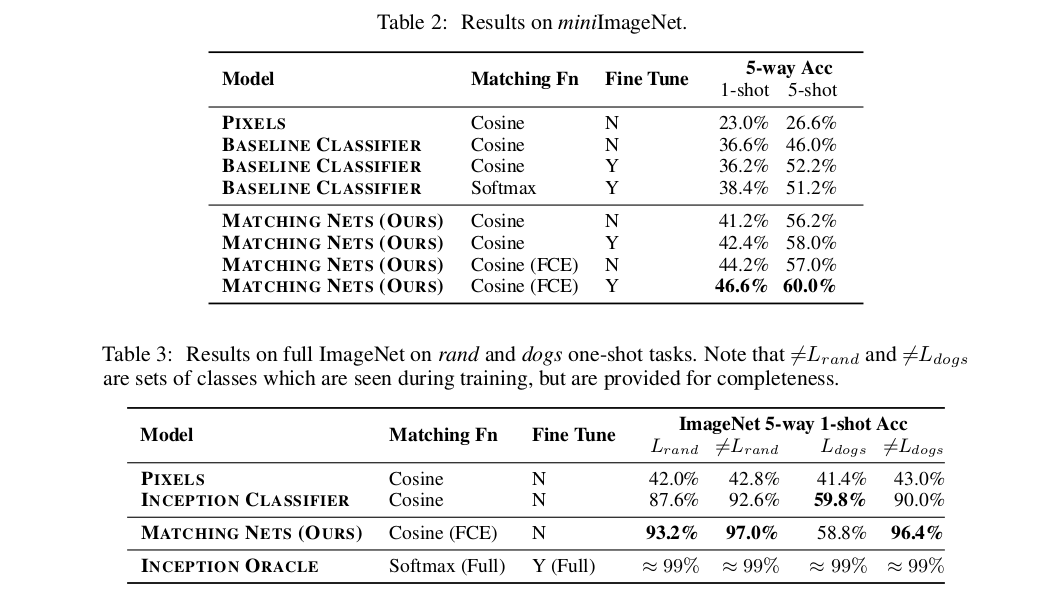


4.1.1 Omniglot

Omniglot [14]由来自50个不同字母的1623个字符组成。每一个都是手绘的由20个不同的人。大量的类（字符），每个类只有相对较少的数据（20），使其成为测试小规模单炮分类的理想数据集。 N路全能任务设置如下：选择N个看不见的字符类，独立于字母表，为L.提供该模型具有N个字符中的每一个的一个图形作为S〜L和批次B〜L[21]，我们用90度的倍数随机旋转增强了数据集，并使用了1200用于训练的字符，以及用于评估的剩余字符类。  
 我们使用一个简单而强大的CNN作为嵌入函数 - 包括一堆模块，每个都是一个3×3的卷积64滤波器，然后批次标准化[10]，一个Relu非线性和2×2最大合并。我们调整了所有的图像大小为28×28，所以，当我们堆叠4模块，所得到的特征图是1×1×64，导致我们的嵌入函数f（x）。 A完全连接层之后是softmax非线性用于定义基线分类器。  
 将基线与我们在Omniglot上的模型进行比较的结果示于表1中和5-shot，5-way和20-way，我们的模型胜过基线。没有什么重大的惊喜这些结果：使用更多的k-shot分类示例有助于所有模型，而5-way比较容易20路。我们注意到，当对S进行微调时，基线分类器提高了一点，并且使用余弦距离与训练来自小训练集的小softmax（因此需要微调）表现良好。暹罗网与我们的Matching Nets相比，每个类使用5个例子，但它们的性能在一次性中快速降低。完全条件嵌入（FCE）没有似乎帮助很多，由于空间的限制，被遗弃在表中。  
 像[11]中的作者，我们也测试我们的方法训练在Omniglot一个完全不相交的任务 - 单次，10路MNIST分类。基线分类器的精确度约为63％（如他们的论文中所报道的）暹罗网行动70％。我们的模型达到72％。

4.1.2 ImageNet

我们的实验遵循与Omniglot相同的设置进行测试，但我们考虑了rand和a狗（更硬）设置。在rand设置中，我们从训练集中随机删除了118个标签只测试这118个类（我们表示为L rand）。对于狗的设置，我们删除所有在ImageNet中的类从狗（总共118）下降，然后在所有非狗类上进行训练测试狗类（L狗）。 ImageNet是一个众所周知的大数据集，这可以是相当的壮举工程和基础设施在其上运行实验，需要许多资源。因此，以及使用完整的ImageNet数据集，我们设计了一个新的数据集 - miniImageNet - 由60,000组成彩色图像尺寸84×84，具有100个等级，每个具有600个实例。这个数据集更多复杂比CIFAR10 [12]，但适合在现代机器上的记忆，使它非常方便快速原型和实验。我们使用80个类进行训练，并对其余的进行测试20类。因此，总的来说，我们有randImageNet，dogsImageNet和miniImageNet。  
 miniImageNet实验的结果如表2所示。与Omniglot，Matching一样网络性能优于基准。然而，miniImageNet是一个比Omniglot更难的任务这使我们能够明智地评估完全上下文嵌入（FCE）（在Omniglot上它没有区别）。正如我们所看到的，FCE提高了Matching Nets络的性能，有和没有微调，通常将性能提高约两个百分点。

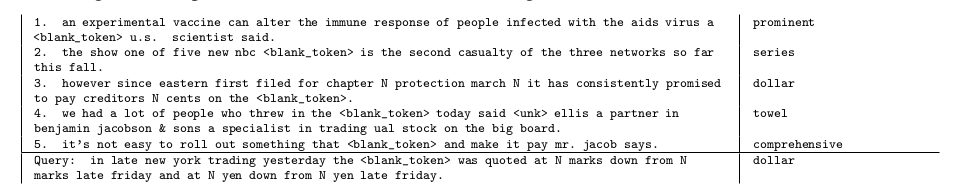


接下来，我们转向基于全尺寸，全规模ImageNet的实验。我们的基线分类器这个数据集被训练以对所有类进行分类，除了在测试集合中的那些（对于randImageNet）或那些关于狗（对于dogsImageNet）。我们还比较了功能在ImageNet中的所有类上训练的Inception Oracle分类器作为上限。我们的基线  
分类器是最强的发布ImageNet模型之一，标准的精度为79％ImageNet验证集。我们不是从头开始训练这些大型任务的Matching Nets络使用来自Inception分类器的参数（预训练）初始化它们的特征提取器f和g对数据的适当子集），然后进一步在随机5路上训练得到的网络来自训练数据集的单击任务，包括完整上下文嵌入和我们的Matching Nets络和培训战略。  
 randImageNet和dogsImageNet实验的结果如表3所示Oracle（在所有类上训练）在仅限于5个类时几乎完美地执行，这是不令人惊讶的给予其令人印象深刻的前1准确性。当训练完全在= L rand，匹配当在L rand上测试时，Nets在开始时提高了近6％，将错误减半。如图2所示两个五向one-shot learning的实例，其中Inception失败。看看所有的错误，开始似乎有时喜欢一个图像高于所有其他（这些图像往往是混乱像示例在第二列中，或者颜色更常量）。Matching Nets，另一方面，管理以从这些离群值中恢复，有时出现在支持集合S中。

Matching Nets络设法改进在互补子集= L狗的初始（虽然此设置不是一次性的，因为已经在这些标签上训练了特征提取）。但是，在更有挑战性的L狗子集，我们的模型降低1％。我们假设这个事实训练期间的采样集S来自标签的随机分布（来自= L个狗），而来自L狗的测试支持集S包含类似的类，更类似于细粒分类。因此，我们相信，如果我们的训练策略适应来自细粒度的样本S.集合标签而不是从ImageNet类树的叶子均匀采样，改进可以实现。我们把这作为未来的工作。

4.1.3 One-Shot Language Modeling

我们还引入了一个新的单向语言任务，类似于为图像检查的那些任务。任务如下：给定一个带有一个缺失单词的查询语句，以及一组支持的句子每个都有一个缺少的单词和一个相应的1热标签，从支持中选择标签设置最匹配的查询语句。 这里我们展示一个单例，虽然注意词在右侧不提供，并且集合的标签被给出为1-hot-of-5向量。



句子取自Penn Treebank数据集[15]。在每次试验，我们确保集和批处理用不重叠的句子填充。这意味着我们不用具有非常低的频率计数的字;例如如果对于给定的单词我们只有一个句子不使用此数据，因为句子将需要在集合和批处理中。与图像一样任务，每个试验包括在集合中可用的类之间的5路选择。我们用了一批在整个句子匹配任务（SMT）中大小为20，并且在k = 1,2,3时改变集大小。我们确保对于集合中的每个类可用相同数量的句子。  
 我们将这些词分成随机抽样9000个训练和1000个测试。同样，我们训练和测试使用从标准PTB训练集中取得的句子并用句子测试从标准测试集。因此，在测试期间使用的词语和句子都没有看见在训练期间。  
 我们将我们的一次性匹配模型与oracle LSTM语言模型（LSTM-LM）[30]训练所有的词。在这个设置中，LSTM有一个不公平的优势，因为它不是one-shot learning，但看到所有的数据 - 因此，这应该被视为一个上限。为了这样做，我们检查类似的设置，其中句子被呈现给具有用5填充的单个单词的模型不同的可能词（包括正确答案）。对于模型给出的这5个句子中的每一个对数似然和这些的最大值被认为是模型的选择。  
 与其他5路选择任务一样，此任务的机会性能为20％。 LSTM语言模型oracle在测试集上实现了72.8％的上限。Matching Nets络与简单编码模型实现32.4％，36.1％，38.2％的任务精度k = 1，2，3实例中的集合，分别。未来的工作应探索结合参数模型如具有非参数组件的LSTM-LM，例如这里探讨的Matching Nets络。  
 两个相关的任务是CNN QA测试的实体预测从新闻文章[5]和儿童的书测试（CBT）[6]。在CBT中，例如，提供来自书的句子序列作为上下文。在最后一句中，在前一句中出现的一个词是失踪。任务是选择正确的单词从一小组单词填充这个空白作为可能的答案，所有这些都出现在前面的句子中。在我们的句子匹配任务在集合中提供的句子从PTB语料库中随机抽取，并且与在查询批处理中的句子只通过它们共享一个单词的事实。与CBT和CNN相反数据集，它们只提供通用而不是特定的顺序上下文。

5

Conclusion

在本文中，我们介绍了Matching Nets，一种新的神经结构，通过它的方式相应的培训制度，能够在各种一次性的最先进的表现分类任务。这项工作有几个关键的见解。首先，one-shot learning是很多更容易，如果你训练网络做one-shot learning。其次，非参数结构在a神经网络使网络更容易记住和适应新的训练集在同一个任务。将这些观察结合在一起产生Matching Nets络。此外，我们已经定义ImageNet上的新的一次性任务，ImageNet的简化版本（用于快速实验）和a语言建模任务。我们的模型的一个明显的缺点是，随着支持集S的增长在大小上，每个梯度更新的计算变得更昂贵。虽然稀疏和基于抽样的方法来缓解这一点，我们的许多未来的努力将集中在这一点上局限性。此外，如ImageNet狗子任务中所示，当标签分布具有时明显的偏差（如细粒度），我们的模型受到损害。我们觉得这是一个令人兴奋的领域我们希望在未来的工作中不断改进的挑战。