Page 1:

各位老师下午好，我是何老师19级的硕士生高也，我硕士论文的题目是基于自注意力机制的自知识蒸馏研究。

Page 2:

在第一部分介绍知识蒸馏和自知识蒸馏的概念。

Page 3:

本文从大类上属于图像分类模型，目的是尽可能提高图像的分类准确率。比如当把图中四张图像输入分类模型之后，希望模型对四张图像的预测标签分别是猫标签、虎标签、狗标签、狼标签。

Page 4:

典型分类模型的网络结构如图所示。输入图像经过多层神经网络和softmax得出预测结果。

它的损失函数为预测结果和真实标签的交叉熵，通过梯度下降法将其最小化。

为了追求越来越优异的分类性能，神经网络的规模与日俱增。由于大型神经网络的训练需要昂贵的计算资源和时间成本，研究者提出知识蒸馏技术对大型神经网络进行压缩。

Page 5:

知识蒸馏的主要思想是利用相似标签的信息，比如猫标签和虎标签、狗标签和狼标签。

它的网络结构如论文第2页的图1.1所示。红框里面的部分等同于刚才提到的典型的图像分类模型，称之为学生模型。红框外上方的神经网络称为教师模型，它的规模远远大于学生模型。红框外的教师模型和学生模型都使用了softmax(T)函数，表达式如论文第3页的式(1.1)所示，T称为蒸馏温度，其目的是让相似标签（比如猫和虎、狗和狼）的概率分布更为接近。

知识蒸馏模型的损失函数如论文第3页的式(1.2)所示，它由学生损失函数和蒸馏损失函数相加而成。通过这种方法把大型教师模型中的暗知识压缩到小型学生模型，提升学生模型的分类准确率。

Page 6:

虽然传统的知识蒸馏模型取得了优异的成果，但是它依然有不足之处。它还是需要训练一个高容量的教师模型，这个训练过程需要大量的计算和存储资源。

Page 7:

那么能否有一种知识蒸馏它不需要外部的教师模型？这就是自知识蒸馏模型。自知识蒸馏模型的特点是让神经网络蒸馏自己内部的知识，而不需要借助外部的教师模型。这里介绍的自蒸馏模型是BYOT模型，它的主要思想是考虑到深度神经网络的深层部分比浅层部分包含更高阶、更抽象的信息，因此可以利用神经网络的深层部分向浅层部分蒸馏暗知识。

BYOT网络结构如论文12页的图2.1所示。将神经网络根据其深度划分为几个浅层块和一个最深层块，在每个浅层块之下设置一个瓶颈层，一个全连接层和softmax(T)。每个浅层块和相应的瓶颈层、全连接层、softmax(T)组成一个浅层分类器，最深层块和其后的全连接层、softmax(T)组成最深层分类器。BYOT模型将所有浅层分类器视为学生模型，将最深层分类器视为教师模型，将最深层分类器的暗知识蒸馏到每一个浅层分类器。

BYOT模型的损失函数如论文14页的式(2.5)所示，它是三种损失函数的加和。第一种损失函数是真实标签的独热向量与每个分类器输出概率分布的交叉熵，第二种损失函数是最深层分类器输出的概率分布和每个浅层分类器输出的概率分布的相对熵，第三种损失函数是最深层块特征图和每个浅层块特征图的损失函数。

Page 8:

让我们思考一下，能否进一步改进BYOT模型？它将作为学生模型的各个浅层块一视同仁直接相加，忽略了各个浅层块对最深层块的不同影响。这可能会造成一些暗知识的损失。

Page 9:

那么，怎样才能有效区分和量化各个浅层块对最深层块的不同影响？

Page 10:

本篇论文提出基于自注意力机制的自知识蒸馏模型，简称SKDSAM模型。它的主要思想是给BYOT模型增加自注意力机制，使自注意力机制通过训练“学习”得出各个浅层块对最深层块的不同影响，从而使不同深度网络层的信息能够更有效地聚合。

SKDSAM模型的网络结构如论文20页的图3.2所示，和BYOT模型相比，它在每一个浅层块和最深层块之间增加了自注意力机制。

SKDSAM模型的损失函数如论文26页的式(3.13)所示，它也是三种损失函数的加和。第一种损失函数是真实标签的独热向量与最深层分类器输出概率分布的交叉熵，第二种损失函数是最深层分离器输出概率分布和每个浅层分类器输出概率分布的相对熵，第三种损失函数是最深层块特征图和每个浅层块特征图的归一化损失函数。需要注意的是，相对熵和归一化损失函数相加以后，还需要乘以相应的注意力权重再与交叉熵相加。红框里为自注意力机制部分的损失函数。那么，怎样计算红框里的，图中自注意力的内部结构是什么？我们在下一页介绍。

Page 11:

我们知道，计算自注意力需要、、和，计算流程如论文22页的图3.3所示。第个浅层块的特征图和最深层分类器的特征图，经过相应的非线性投影层，得到自注意力机制的和，如式(3.6)所示。再经过softmax(T’)层，计算的点积经过softmax即可得到注意力权重，如式(3.7)所示。每一个浅层块第二种损失函数和第三种损失函数的加和，即可得到。注意力权重和对应的相乘后累加，即可得到注意力部分的损失函数，如式(3.12)所示。注意这里的蒸馏温度T’和分类器中的蒸馏温度T是两个相互独立的温度。

Page 12:

对SKDSAM模型的直观理解如论文23页的图3.4所示，它在每个浅层块和最深层块之间增加一个自注意力连接，经特征提取后得到自注意力机制的输入和；然后通过和计算出第个浅层块对应的注意力权重。计算得出的注意力权重即代表第个浅层块对于最深层块影响力的量化，从而解决了本部分一开始提出的问题。

Page 13:

在第三部分通过实验比较了SKDSAM模型和其他自知识蒸馏模型的分类准确率。

Page 14:

由于时间所限，我们只展示这一组实验结果，它是论文41页的表4.8。第一列代表SKDSAM模型和作为对比的各种自知识蒸馏模型，第二列和第三列分别代表各模型在CIFAR-100数据集和Tiny ImageNet数据集上的分类准确率。实验结果表明，SKDSAM模型具有比其他自知识蒸馏模型（尤其是作为对比的BYOT模型）更优异的分类准确率，达到了目前SOTA结果，这证实了使用注意力机制区分不同浅层块对最深层块影响是有效的。

Page 15:

Page 16:

总结前面的内容，本文将自注意力机制和自知识蒸馏模型结合起来，使用自注意力机制量化自知识蒸馏模型中各个浅层块对最深层块的不同影响，达到了SOTA的分类准确率。

Page 17:

以上就是讲述的全部内容，谢谢大家，请各位老师提问。