MixMatch：一种全面的半监督学习方法

**Abstract**：半监督学习被证明是一种利用无标注数据来减少对标注数据集依赖的范例，在我们的工作中，我们整合了最近半监督领域的方法来产生一个新的算法——MixMatch，猜测低熵的标签作为数据增广无标注例子然后用Mixup混合这两种数据。

**Introduction**：

最近模型的损失项分为三项：

①熵最小化——鼓励模型输出对于未标注数据的自信输出

②一致性正则——鼓励模型在受到扰动时输出相同的输出

③常规正则化——鼓励模型更好的生成并且避免训练时过拟合

**Related Work**：

**一致性正则：**

监督学习中常用的正则化技术是数据增广，应用于假定为类语义不受输入变化影响，例如，图像分类中，引入弹性变换或者加入噪声，能够在不改变标注的前提下改变像素内容，通俗来讲，它能够通过产生无限接近的新的模拟数据来拓展数据集大小。一致性正则的思路就是利用一个分类器应该输出相同的类分布在未标注样例中即使他已经被增广了。



计算两个增广的L2损失，MT将其中一个用模型参数的EMA值代替了，缺点就是他们运用了特定域的数据增强策略，VAT通过在输入增加了一个加性扰动最大化改变了输出类分布。

**熵极小化**：

在许多半监督学习方法中，一个常见的基本假设是分类器的决策边界不应通过边缘数据分布的高密度区域。实现这个需求的一个办法是分类器输出低熵预测在无标注数据上。

这通常被称为集群假设：如果两个样本在输入分布中属于同一个集群，那么它们很可能属于同一个类。聚类假设相当于低密度分离假设：决策边界应位于低密度区域。这种等价性很容易推断：一个位于高密度区域的决策边界将一个聚类分成两个不同的类，要求不同类的样本位于同一个聚类中，这违反了聚类假设。半监督学习的一致性正则化方法通过鼓励对未标记点u+δ的扰动f（u）=f（u+δ）的不变预测来强化低密度分离假设，当且仅当决策边界穿过低密度路径时，这种一致性和小预测误差可同时满足。

通过最小化无标注数据的条件概率实现(Semi-supervised learning by entropy minimization. In

Advances in Neural Information Processing Systems, 2005)。这种熵最小化通过与VAT结合得到更好的结果，Pseudo-Label通过构建硬标签从对无标签数据的高置信度预测和用他们来训练target在标准的交叉熵损失中隐式的实现了熵极小化，MixMatch是通过对无标注数据target分布的“sharpening”函数隐式的实现熵极小化的。

**传统正则化**：

传统正则化就是对模型施加一个约束使得他难以记住训练数据希望他能够更好的推广没见过的数据，我们用权重衰减来惩罚模型参数中的L2范式，我们也用Minxup鼓励曲线表现在样例之间，我们将Mixup用在正则(应用于有标注数据)和半监督方法(应用于无标注数据)中。

**MixMatch**：

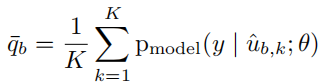
给定相同batch的有标注数据X和独热编码的target，和相同batch的无标注数据U，MixMatch产生一个batch的增广有标注数据X’和增广的带”guessed label”的无标注数据U’。

**数据增广**：

是对每个有标注数据我们对它做一次变换，对每个无标注数据，我们对它做K次变换，我们用这些对每个无标注数据产生一个”guessed label”。

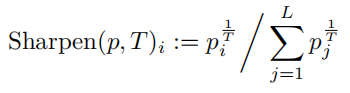
**Label guessing**:

我们用刚刚K次增广的无标注数据来计算模型类预测的平均分布，这个guessing等会会在无监督损失项用到。



**Sharpening**:

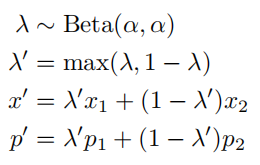
在产生label guess中，我们参考SSL中的熵极小化引入了而外的步骤。给定无标签增强数据的平均，我们通过sharpening函数减少数据分布的熵。



T越接近0，输出分布越接近狄拉克分布，降低T鼓励模型产生低熵预测。

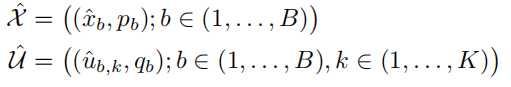
**MixUp**：

对于一对预测标签，我们计算

作为他的权重。

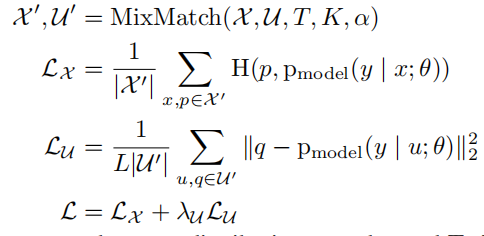
保证了x’是更接近x1而不是x2的，

我们收集了有标注数据的增广，和无标注数据的K组增广。



将他们混合在一起形成，将与组合形成，将剩下的与形成。

将得到的example用于计算损失：



**损失函数**：

我们在无标注损失部分没有使用交叉熵损失的原因是相较于交叉熵，平方L2损失对错分样例有着更低的敏感性。我们不通过猜测的标签传播梯度。