创新设计结题报告

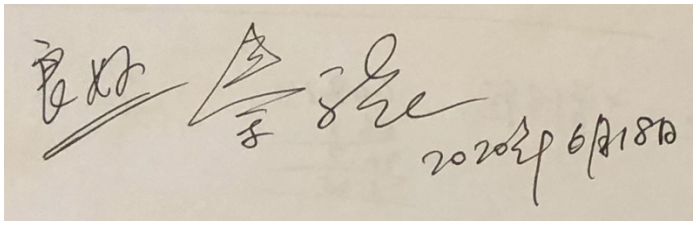
题目：多实例学习

姓名：张峰玮

班级：大数据 17

学号：201700301004

导师：秦学英 教授



目录

介绍 MIL 和 MIML

在典型的机器学习问题中，如图像分类，每张图像清楚地表示一个类别。然而，在许多实际应用中，人们会同时观察到多个实例，并且只了解该类别的一般说明。这种情况称为多实例学习。多实例学习（MIL）最初被提出用于药物活性预测[1]。目前它已被广泛应用于许多领域，成为机器学习中的一个重要问题[2]。MIL 处理一组具有单一类别的实例。因此，MIL 的主要目标是学习一个模型，运用这个模型预测一组实例的类别，如医学诊断。

MIL 是一种弱监督学习。每个样本是一个被标记的包，包括许多与输入特征相关的实例。在二分类任务中，MIL 的目标是训练一个分类器来预测被测试的包的标签。这基于一个假设：若一个包中存在正实例，则这个包是正的；否则，这个包是负的。因此，MIL 的关键点是多个实例的标签的不确定性，特别是当正包的组成可能有多种情况时。

MIL 算法可分为三个层次[3]：实例空间范式、包空间范式和嵌入空间范式。实例空间范式通过聚合实例级分类器的响应来学习实例分类器并执行包分类。包空间范式利用包之间的关系，将包作为一个整体来处理，特别是计算包到爆的距离/相似性，然后用最近邻或贝叶斯分类器进行包分类。嵌入空间范式将一个包嵌入到基于词汇的特征空间中，得到一个包的紧凑表示，例如一个向量表示，然后应用经典分类器来解决包的分类问题。

深层神经网络已经被应用于解决许多机器学习问题。对于监督学习，有几种神经网络：深信念网络（DBN）[4]采用无监督预训练，以固定长度的矢量作为输入进行特征学习和分类；卷积神经网络（CNN）[5]，[6]以二维图像为输入，主

导了图像识别；循环神经网络（RNN）和长短期记忆（LSTM）网络[7]以文本和语音等序列数据为输入，擅长处理序列预测。通常，训练这些深层网络需要大量完全标记的数据，即每个实例都需要一个标记。然而，在 MIL 中，只能得到袋子标签。同时，MI 数据具有更复杂的结构，即一组实例。对于不同的包，实例的数量是不同的。这些问题使得传统的神经网络很难处理 MIL 问题。

为了解决包分类问题，有许多方法被提出，如利用不同包之间的相似性[8]、将实例嵌入到低维空间后再进行包分类[9]或者是混合多个实例级分类器的输出得到包分类器的输出[10]。

上述三种方法中，只有最后一种方法才能提供可解释的结果。然而，研究表明，此类方法的实例级准确度较低[11]，并且通常在实例级 MIL 方法之间存在分歧[12]。这些问题使人们对解释最终决策的现行 MIL 模型的可用性产生疑问。

在许多真实世界应用中，感兴趣的对象具有固定的结构，并且它可以表示了一包实例（a bag of instances），多个标签和这个 bag 级别相关。例如，在文本分类中，每个文档可能有一些句子作为实例（Multi-Instance），并且有许多标签（Multi-Label）指派给文档级别。Multi-Instance Multi-Label 为解决这种问题提供了一个框架。以 MIML 的角度，训练数据为$\{(X_1,Y_1),\ldots,(X_m,Y_m)\}$，其中，每个包 X_i 可以表示为 Z_i 实例。输出 Yi 是一个所有可能性标签$\{y_1,y_2,\ldots y_L\}$的子集,其中 L 为可能的单个标签的数量。

过去几年提出了很多 MIML 的算法，并且应用到了不同域的任务中，例如图片分类，文本分类，视频标注，基因功能预测（gene function prediction），生态系统保护（ecosystem protection）等等。大多数 MIML 的研究假设实例（instances）已经提前给出，或者通过一些手工的实例产生器（instance generators）产生，实

例产生器直接从原数据集中提取实例。最近，在图片任务上的经验研究表明，无手工设计的实例产生器占据主导。考虑到特征学习技术已经在许多领域都击败了手工特征工程，这促使工作者尝试去用自动表示学习来解决 MLML 问题，尽管这需要大量的数据。

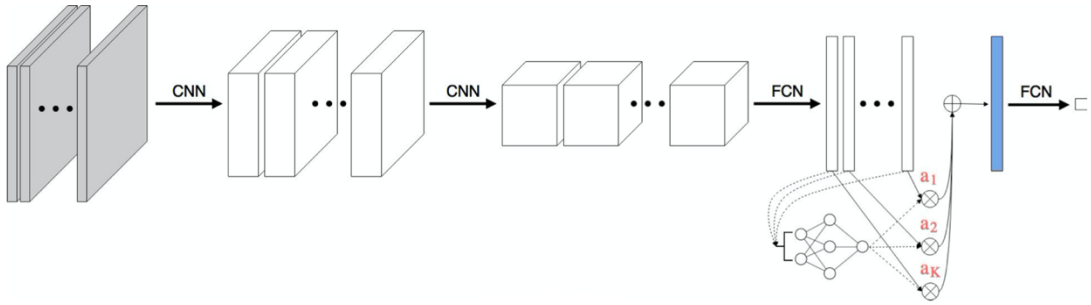
阅读笔记

我阅读了文章 Attention-based Deep Multiple Instance Learning[13]。作者提出了一种将可解释性融入 MIL 的方法。作者利用伯努利分布建立了包标签的 MIL 模型，并通过优化对数似然函数对其进行训练。对称函数基本定理的应用提供了一个建模包标签概率（包得分函数）的一般过程，该过程包括三个步骤：将实例嵌入到低维的转换，置换不变的（对称的）聚合函数，以及到包概率的变换。文章建议使用神经网络（即卷积层和完全连接层的组合）参数化所有转换，这增加了方法的灵活性，并允许通过优化无约束目标函数以端到端的方式训练模型。最后，作者提出用一个可训练的加权平均来代替广泛使用的置换不变算子，如最大算子和平均算子，其中权值由两层神经网络给出。这两层的神经网络类似于 attention 机制[14]。

我阅读了文章 Deep MIML Network[15]。在这篇文章中，作者提出了 Deep MIML Network，一个深度神经网络模型。Deep MIML 天生具有深度模型的表示学习能力，因此，不需要使用另外的实例产生器来产生实例描述。相反，模型本身就可以完成实例表示产生和后继的学习过程。另外，作者仔细地设计了 sub-concept 层。这个层可以插曲其它类型的网络结构中，例如 CNN，使他们具有发现 pattern-label 关系发现能力。Deep MIML 的有效性在作者的实验中被验证了。

相关概念

在 MIL 问题中，一个包 X = {x_1, x_2, …, x_k}，X 中的各项无序且没有依赖关系。假设不同的包的 K 值可以不同。单一的标签 Y 与这个包相关。更进一步的，假设包内的每个实例都存在一个标签 y_k，$y_k \in \{0, 1\}$。然而，这些标签是无法获取的，在训练期间它们仍然是未知的。



对 MIL 问题，可用以下公式简要描述：

$$Y = \max_k \{y_k\}.$$

定理 1 一组实例的评分函数 S(X)∈R 是一个对称函数，当且仅当它可以按以下形式分解：

$$S(X) = g\Big(\sum_{\mathbf{x} \in X} f(\mathbf{x})\Big)$$

f 和 g 是适当的变换。

定理 2 对于任意 ε>0，Hausdorff 连续对称函数 S(X)∈R 可被近似为一个形式为 g($\max_{x \in X}$f(**x**))的函数，

$$|S(X) - g\Big(\max_{\mathbf{x} \in X} f(\mathbf{x})\Big)| < \varepsilon.$$

max 是元素级向量的最大算子，f 和 g 是连续函数。

定理 1 和定理 2 的区别在于前者是一个普遍分解，而后者提供了一个任意

逼近。尽管如此，他们都提出了一种通用的三步分类方法：（i）使用函数 f 的实例转换，（ii）使用对称（置换不变）函数 σ 的转换实例组合，（iii）f 使用函数 g 转换的组合实例的转换。最后，分数函数的表达依赖于 f 和 g 函数类的选择。

在 MIL 问题中，两个定理中的得分函数都是概率 θ(X)，置换不变函数 σ 称为 MIL 池。函数 f，g 和 σ 的选择决定了一种特定的标签概率建模方法。对于给定的 MIL 算子，有两种主要的 MIL 方法：

i. 实例级方法：变换 f 是一个实例级分类器，它返回每个实例的分数。然后通过 MIL 池对每个实例的得分进行汇总，得到 θ(X)。函数 g 是恒等函数。

ii. 嵌入级方法：函数 f 将实例映射到低维嵌入。MIL 池用于获取一个包表示，它独立于包中的实例数。包表示由包级分类器进一步处理以提供 θ(X)。

由于单独的标签是未知的，实例级分类器可能会有训练不足的问题，这会给最终的预测带来额外的错误。嵌入级方法确定了一个包的联合表示，因此它不会给包层分类器带来额外的偏差。另一方面，实例级方法提供了可用于查找关键实例的分数。这里将展示如何通过使用一个新的 MIL 池来修改嵌入级方法，使之可以解释。

MIL 问题要求 MIL 池 σ 是置换不变的。如定理 1 和定理 2 所示，有两个 MIL 池运算符确保得分函数（即包概率）是对称函数，即最大算子和平均算子。但是，它们是预定义的、不可训练的。例如，在实例级的方法中，max 算子可能是一个不错的选择，但它可能不适合嵌入级的方法。类似地，聚合实例

得分时，mean 运算符绝对是一个糟糕的 MIL 池，尽管它可以成功地计算包表示。因此，灵活和自适应的 MIL 池可以通过适应任务和数据来获得更好的结果。理想情况下，这种 MIL 池也应该是可解释的。

Attention 机制 建议使用实例的加权平均（低维嵌入），其中权重由神经网络确定。此外，权重总和必须为 1，以不受包大小的影响。加权平均满足定理 1 的要求，其中权重和嵌入是 f 函数的一部分。假设 $H=\{h_1, \ldots, H_K\}$是 K 个嵌入的包，那么使用以下 MIL 池：

$$\mathbf{z} = \sum_{k=1}^{K} a_k \mathbf{h}_k,$$

where:

$$a_k = \frac{\exp\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_k^\top\right)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top \tanh\left(\mathbf{V}\mathbf{h}_j^\top\right)\}}$$

$\mathbf{w} \in \mathbb{R}^{L\times 1}$ and $\mathbf{V} \in \mathbb{R}^{L\times M}$都是参数

Gated Attention 机制[16] 此外，注意到 tanh（·）非线性对于学习复杂关系可能是低效的。tanh（x）对于 x∈[-1，1]是近似线性的，这可能会限制实例间学习关系的最终表达。因此，建议将 Gating 机制与 tanh（·）非线性结合使用，即：

$$a_k = \frac{\exp\{\mathbf{w}^\top\left(\tanh\left(\mathbf{V}\mathbf{h}_k^\top\right) \odot \mathrm{sigm}\left(\mathbf{U}\mathbf{h}_k^\top\right)\right)\}}{\sum_{j=1}^{K} \exp\{\mathbf{w}^\top\left(\tanh\left(\mathbf{V}\mathbf{h}_j^\top\right) \odot \mathrm{sigm}\left(\mathbf{U}\mathbf{h}_j^\top\right)\right)\}}$$

$\mathbf{U} \in \mathbb{R}^{L\times M}$是参数，⊙是元素相乘，sigm(·)是 sigmoid 非线性。Gating 机制引入了一种可学习的非线性，这可能消除 tanh(·)中的线性。

在 Deep MIML Network 中，作者提出了一个新的二维（2D）神经网络层，叫做 sub-concept layer。这可以使得模型能为每个类别标签在 instance 和 sub-

concepts 之间匹配分数。对于一个给定的实例向量 x，2D Sub-concept 层的第(i, j)个结点表示为实例 x 和第 i 个 sub-concept 对于第 j 个类别标签的匹配分数。

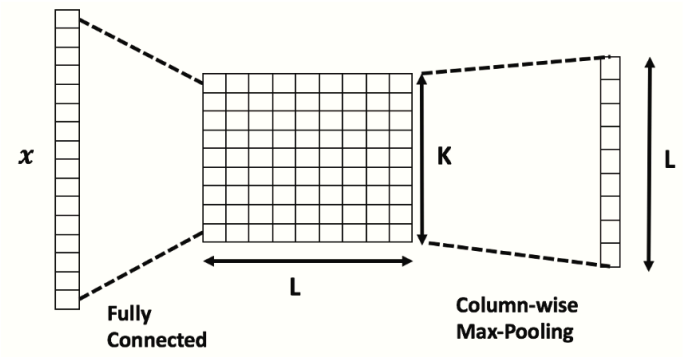


Figure 1: Illustration of a 2D Sub-concept layer

当输入以一袋实例来表示时（这里假设每袋有相同数量的示例，对于那些有不同数量示例的袋，用 0 来补齐），可以一般化 2D Sub-concept 层的想法，融入到 MIML 视角中。基本的想法是通过堆叠许多 2D 层，把 2D sub-concept 层拓展为 3D Tensor 层，tensor 的每一个部分是每个示例的 2D Sub-concept 层。换句话说，给定一袋示例 X_i，为每个示例 X_{ki} 构建 2D Sub-concept 层，然后把这些 2D 层堆叠正一个 3D 的张量。张量的深度和输入袋的示例数目相等。

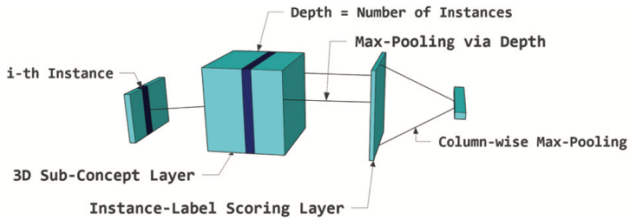


Figure 2: 3D Sub-concept layer and Instance-Label scoring layer. Each instance is connected with its corresponding sub-concept layer only. The resulting three-dimensional tensor has depth equals the number of input instances.

现在，引入 Deep MIML 网络，它从原始输入中产生示例包，学习 instance 级别下每个标签的 sub-concept 的得分函数，最终得到包级别的预测。具体地，原始输入被送入一个 Instance Generator，这个设备在 interest 域内是独立的。对

于图片任务，在 Full Connection 层前面加一个深度卷积网络结构会使得效果变好，随后，一个 3D sub-concept 层接着两个池化层，这直接应用到 Instance Generator，最后的部分是一个全连接层，层的大小和 labels 的数量是独立的。

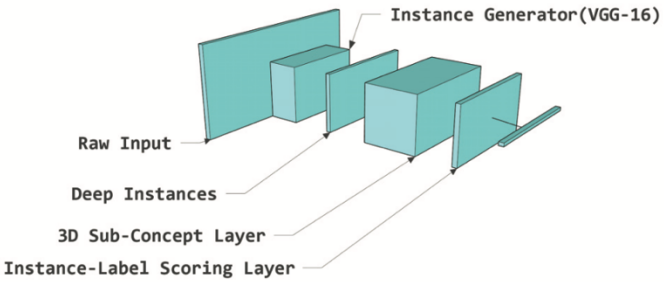
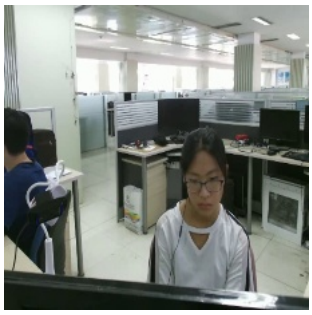


Figure 3: DeepMIML Network

设计新的深度神经网络处理视频多实例问题

1. 问题分析

我们想通过分析一段由电脑前置摄像头录制的视频，获得用户的专注程度。视频的中心区域显示用户的头部。程序会检索视频，根据图像获得用户的头部姿态和用户情绪。然后，以头部姿态和情感为输入，程序计算出用户的专注度。根据获得的专注度信息，我们可以开展进一步的研究。

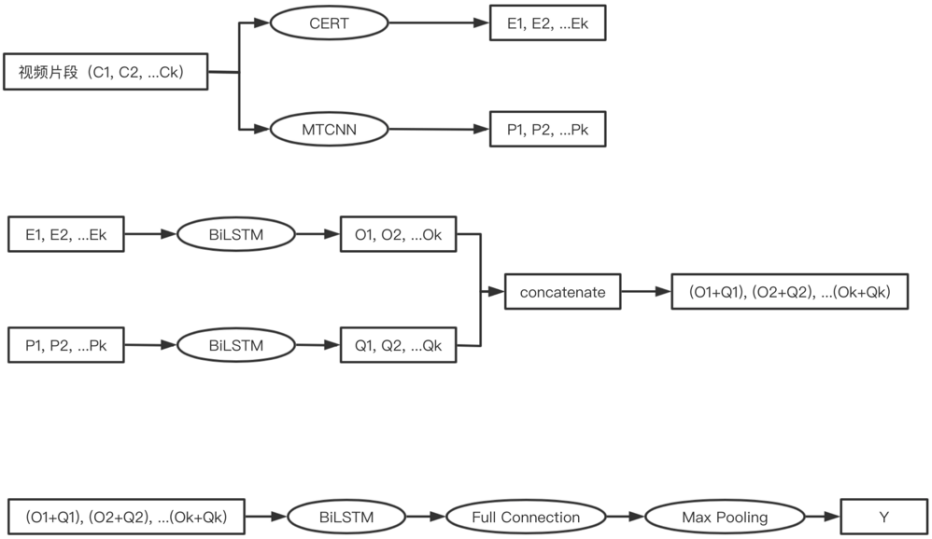


视频中的某一帧图像

2. 解决思路

我们把视频切分为多个片段，把每个片段看作是一个多实例的包。我们希望以包为单位，学习整个视频所包含的信息。

首先，提取头部姿态和情感。参考 CERT[17]的思想，设计情绪提取器；利用 MTCNN[18]设计头部姿态提取器。然后，将视频切分为多个片段（$C_1, C_2, \ldots C_k$）,对每个片段分别提取情绪（$E_1, E_2, \ldots E_k$）和头部姿态（$P_1, P_2, \ldots P_k$）。将这两组序列式的特征分别带入两个 BiLSTM 得到两组新的输出（$[O_1, O_2, \ldots O_k]$，$[Q_1, Q_2, \ldots Q_k]$）。直接拼接（concatenate）这两组输出，构成一组特征向量。将这组特征向量送入一个 BiLSTM，得到一组输出。这组输出经全连接层、平均池化层得到最后的结果 Y。



在这里，我们参考了 Deep MIML Network 的设计思想，并使用 BiLSTM 来提取序列的特征。在 Deep MIML Network 中，作者在最后选择了 Max Pooling。根据任务需求，我们选择使用 Mean Pooling，即计算一个视频片段内的平均结果。

总结

文章 Attention-based Deep Multiple Instance Learning 作者的实验结果说明，该方法在 MINIST 数据集上表现良好。但是，该方法不能给出包内每个实例单独的标签预测。给出每个实例的标签预测会极大地扩展 MIL 的应用范围，但原作者并未实现这一目标。

阅读文章 Deep MIML Network 后，我设计了一个新的多实例识别器，希望高效地处理视频片段。以后我会逐步实现这个识别器的全部功能（网络结构的具体搭建），用已经获得的头部姿态图像（共 23600 张）和人工标注进行训练，并测试它的表现。

参考文献

[1] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, “Solving the multiple instance problem with axis-parallel rectangles,” Artificial Intelligence, vol. 89, no. 1, pp. 31–71, 1997.

[2] Wang, Xinggang , et al. "Revisiting Multiple Instance Neural Networks." Pattern Recognition 74(2018):15-24.

[3] J. Amores, “Multiple instance classification: Review, taxonomy and comparative study,” Artificial Intelligence, vol. 201, pp. 81–105, 2013.

[4] G. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets,” Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in NIPS, 2012, pp. 1097–1105.

[7] R. J. Williams and D. Zipser, “A learning algorithm for continually running fully recurrent neural networks,” Neural computation, vol. 1, no. 2, pp. 270–280, 1989.

[8] Cheplygina, Veronika, Tax, David MJ, and Loog, Marco. Multiple instance learning with bag dissimilarities. Pat- tern Recognition, 48(1):264–275, 2015b.

[9] Andrews, Stuart, Tsochantaridis, Ioannis, and Hofmann, Thomas. Support vector machines for multiple-instance learning. In NIPS, pp. 577–584, 2003.

[10] Ramon, Jan and De Raedt, Luc. Multi instance neural networks. In ICML Workshop on Attribute-value and Relational Learning, pp. 53–60, 2000.

[11] Kandemir, Melih and Hamprecht, Fred A. Computer-aided diagnosis from weak supervision: a benchmarking study. Computerized Medical Imaging and Graphics, 42:44–50, 2015.

[12] Cheplygina, Veronika, Sørensen, Lauge, Tax, David MJ, de Bruijne, Marleen, and Loog, Marco. Label stability in multiple instance learning. In MICCAI, pp. 539–546, 2015a.

[13] Ilse, Maximilian , J. M. Tomczak , and M. Welling . "Attention-based Deep Multiple Instance Learning." (2018).

[14] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[15] Feng J . Deep MIML Network[C]// AAAI-17. 2017.

[16] Dauphin, Yann N, Fan, Angela, Auli, Michael, and Grangier, David. Language modeling with gated convolutional networks. arXiv preprint arXiv:1612.08083, 2016.

[17] Bartlett M S , Littlewort G , Wu T , et al. Computer expression recognition toolbox[C]// 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2008), Amsterdam, The Netherlands, 17-19 September 2008. IEEE, 2008.

[18] Zhang K , Zhang Z , Li Z , et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.