论文标题

Deep clustering: Discriminative embeddings for segmentation and separation

要点

“信息**分割**问题”（segmentation problem）主要是将一个对象中的各个元素分配到不同的组或者分区，并为每个元素分配一个组/分区标签。这里的对象，例如，可以是图像中的像素，音频中的 T-F 单元，因此，可以注意到聚类方法可以被用到信息**分割**问题当中，不过传统上信息分割问题会被描述为一个域独立的问题---用一个基于点对之间关系的目标函数来描述，而**分区**（partitioning）则是一个应用在整个输入上的更加复杂的处理过程。

本文提出一种基于分区（partition-based）的分割问题，它可以为所有的输入元素学习一种嵌入表示（embeddings），之后可以方便地使用聚类算法打上标签。在音频应用领域，单声道的分离方法如将频谱图中的每个 T-F 单元（被目标说话人占据）分割到不同的区域，可以使用分类器或者生成模型（class-based 方法，其实就是监督学习模型）不过监督学习的方法需要已知 class 有多少，这制约了模型的泛化能力因为现实世界里 class 有很多很多。谱聚类在机器学习的研究领域（图像分割，音频分割）非常活跃。这种方法在信号的元素的特征之间使用局部近似度量，然后使用对归一化的近似矩阵进行谱分解的方法来优化大量的目标函数。谱聚类比 K-means 的有点事它不要求点必须紧紧聚集在某个中心上，能够在混乱的拓扑结构中寻找到簇，聚集到一起的点会构成一个不规则的子图。

定义如下的一个原始混合信号：

$$X_n = g_n(x), n \in \{1, \ldots, N\}$$

如果是图像，n 代表（超）像素的位置（图片已经被向量化），如果是音频，n 代表每个 T-F 单元的位置：

$$X_n = X_{t,f}$$

我们假设存在一个合理的分区将元素 n 划分进去。在信号源分离领域，这些划分出来的不同分区可以被定义为 T-F 单元的集合，每个分区里的 T-F 单元都被一个

特定的信号源（说话人）占据。通过估计分区的方法能够构建出 T-F 掩码矩阵，应用到混合信号X_n上就可以将其中的各个信号源隔离出来。

为了估计出分区，我们选择了一个 K 维的嵌入表示$V = f_\theta(x) \in \mathbb{R}^{N\times K}$，参数是$\theta$，在嵌入空间里使用简单的聚类方法为元素$\{1,\ldots,N\}$划定不同的分区。研究中，$V = f_\theta(x)$基于深度的神经网络都整个输入信号$x$进行计算（允许先进行特征提取），嵌入向量可以简单理解为对 T-F 单元的一种编码，在这里我们考虑使用一种经过单位向量化的嵌入表示：

$$|v_n|^2 = \sum_k v_{n,k}^2 = 1, \forall n$$

$v_n = \{v_{n,k}\}$，$v_{n,k}$ 是第 n 个元素的嵌入向量的第 k 个维度。定义一个标签指示器：

$$Y = \{y_{n,c}\}$$

它起到将每个元素 n 映射到类别分区 c 的作用，因此如果$y_{n,c} = 1$，则代表元素 n 位于分区 c。训练用的目标函数（K-means）如下：

$$C(\theta) = |VV^T - YY^T|_W^2 = \sum_{i,j:y_i=y_j} \frac{(\langle v_i, v_j\rangle - 1)^2}{d_i} + \sum_{i,j:y_i\neq y_j} \frac{(\langle v_i, v_j\rangle - 0)^2}{\sqrt{d_i d_j}}$$

$$= \sum_{i,j:y_i=y_j} \frac{|v_i - v_j|^2}{d_i} + \sum_{i,j} \frac{\left(|v_i - v_j|^2 - 2\right)^2}{4\sqrt{d_i d_j}} - N,$$

$|A|_W^2 = \sum_{i,j} w_{i,j} a_{i,j}^2$ 是一个 weighted Frobenius norm（类似 L2），其中 $W = d^{-\frac{1}{2}} d^{-\frac{T}{2}}$，$d_i = |\{j : y_i = y_j\}|$=$YY^T 1$。D 是一个关于分区大小的 N*1 向量，的如果元素 i 和元素 j 在同一个分区，这个目标函数会促使$\langle v_i, v_j\rangle$的值为 1，同时促使平方距离$|v_i - v_j|^2$的值为 0.注意到公式的第一部分是 K-means 所要最小化的目标函数，第二部分可以看成一个约束，在训练过程中，目标函数的分数值会不断降低。（简单理解该公式，缩小两个相似矩阵之间的差距）。

同理，如果是谱聚类，我们可以定义一个理想的相似矩阵$A^* = YY^T$，而可以真正计算出来的近似矩阵为$A = VV^T$，我们的目标就变成了：

$$C = |A - A^*|_{\mathrm{F}}^2$$

它通过计算模型的相似矩阵和理想相似矩阵之间的绝对偏差来优化整个模型

定义 $D = \mathrm{diag}(YY^T1)$，则：

$$C = |VV^T - YY^T|_{\mathrm{W}}^2 = |V^TD^{-\frac{1}{2}}V|_{\mathrm{F}}^2 - 2|V^TD^{-\frac{1}{2}}Y|_{\mathrm{F}}^2 + |Y^TD^{-\frac{1}{2}}Y|_{\mathrm{F}}^2$$

这样计算的好处是可以避免建立 N*N 的相似矩阵（N>K），目标函数对 V 的梯度计算如下：

$$\frac{\partial C}{\partial V^T} = 4D^{-\frac{1}{2}}VV^TD^{-\frac{1}{2}}V - 4D^{-\frac{1}{2}}YY^TD^{-\frac{1}{2}}V.$$

上述方法可以认为是在直接优化一个 low-rank 相似矩阵因此处理更加高效，参数也被调整到 low-rank 结构。

在测试的时候，我们计算了测试信号的嵌入表示 V，然后为每一行 $v_i \in \mathbb{R}^K$ 进行聚类，比如使用 k-means。在进行聚类之前，我们还有选择性地使用了谱聚类形式降维方法，用 SVD 进行分解，得到 $\tilde{V} = USR^T$，其中 $\tilde{V} = D^{-\frac{1}{2}}V$，$D = VV^T1_N$，然后按照特征值从大到小的顺序排好序（注意降维后的矩阵是 U，U 的主成分数目是 m），对 U 进行归一化后进行聚类：

$$\tilde{u}_{i,r} = u_{i,r} / \sqrt{\sum_{r'=1}^{m} u_{i,r'}} \ : r \in [1, m]$$

实验数据集

我们的实验主要考虑分离 2 说话人和 3 说话人（说话人的数量不能太多，因为人说话的语音相似性很大，很有可能被划分到同一个类别下，不过源的数量可以不限制，即混合同一个说话人的多段语音，因为模型会找到同一说话者的发声的特点），数据集采用的是 Wall Street Journal (WSJ0) corpus，其它数据集限制比较大。

训练集主要包含了 30h 的双说话人混合音频，它们是从 WSJ0 的训练集 si_tr_s 中的不同说话人中随机选择一些发言片段合成的，并且这些合成的音频段被混入到 0dB-5dB 之间的多个 SNR 等级中。我们从上面的整个训练集（假设取名为 whole）设计了两个训练子集，一个考虑了混合音频在性别上的平衡性（男女各

自 22.5h，取名为 balanced），另外一个仅仅使用了女性说话人的混合音频（7.5h，取名为 female），10h 的交叉验证集（closed speaker set）以相同的方式从 WSJ0 训练集 si_tr_s 中生成，主要用来优化和调整某些参数，并用来评估 closed speaker set（CC，依然是来自训练集的数据，模型已经见过了）上的语音分离的表现，另外有 5h 的评估数据从 WSJ0 验证集 si_dt_05 和评估集 si_et_05 的 16 位说话人中使用他们的发言片段以类似方式生成，注意到 si_dt_05 和 si_et_05 中的说话人和训练集和 closed speaker set 中的是不一样的，评估数据（测试集）被称为 open speaker set（OC），很多已有的语音分离方法是不能处理 OC 数据集的，因为这些方法所使用的模型缺少对未知说话人的认知。对于评估数据，我们还分别为 OC 和 CC 数据集创建了各自 100 段发言，它们都是 3 说话人的混合音频。所有的数据在处理前被降采样到 8kHz，用于减少计算和内存使用的代价。

输入特征 X 是混合音频的 log short-time Fourier spectral magnitudes，使用的窗长为 32ms，窗位移为 8ms，并使用 hann 窗的平方根，为了保持局部一致性，混合音频使用 100 帧的长度进行分割，这个长度大约是一段语音中一个单词的长度，并在对分割后的输出嵌入表示分开处理。IBM 在训练我们的网络的时候被用于构建目标 Y。IBM 给出了一个 T-F 单元的所有者（某个信号源），该所有者的幅度在所有位于该 T-F 单元的信号源的幅度中是最强的，掩码的值为 1 代表属于该所有者，否则属于其它信号源，YY^T 是混合音频的理想近似矩阵。

为了避免安静区域在分离过程中引发的问题，在训练的时候一个二值化的权重会添加到每个 T-F 单元上，只保留每个信号源中幅度超过某个 ratio 的 T-F 单元（例如可以设置阈值为-40dB，可以得到一个布尔矩阵作为二值权重矩阵），这些二值化的权重可以指引神经网络忽略那些对所有的信号源来说不太重要的 T-F 单元。

实验结果分析

网络在训练的过程中主要基于输入 X 和理想相似矩阵 YY^T（随机初始化？全 0 初始化？）实验中使用的网络结构为 BLSTM 层，并在上方添加一个全连接层，每个 BLSTM 层有 600 个 hidden cells，全连接层则与 embedding dimension (K). 换言之，全连接层就是 embedding（嵌入）层，在每一步的更新中，一个均值 0

方差 0.6 的高斯噪声被添加到权重中。我们准备了多个神经网络用于语音分离的实验（嵌入层的维度为 5-60）。此外，两个不同的激活函数 logistic 和 tanh 被用于探索嵌入空间 V 中 $v_{n,k}$ 的不同范围。对于每个神经网络中嵌入层，权重初始化采用 0 均值和 0.1 方差并使用 tanh 作为激活函数，训练时使用 whole 训练集。另有一部分实验使用快乐 logistic 作为激活函数，训练数据集使用 balanced 和 female，这部分实验在网络初始化时用 tanh 激活函数和 whole 训练集。实验实现采用 CURRENT（支持 GPU 的深度学习框架）。

测试阶段和训练阶段类似，由于输入的混合音频按 100 帧长度分割为很多片段，因此聚类时要把这些片段的嵌入空间 V 进行拼接，在聚类过程中需要解决序列问题（在掩码过的混合音频和每个信号源的频谱之间最小化 L2 损失），这个模型有趣的地方是训练用双说话人却可以在测试时分离出 3 说话人（简单调整簇的数目 2->3）。实验时使用的基线是 SNMF（256bases，详见论文）。评估标准用 SDR（使用 bss_eval toolbox），在双说话人和 3 说话人的混和音频上，初始 SDR 分别为 0.16dB 和 2.95dB。

Table 1: SDR improvements (in dB) for different clustering methods.

method	closed speaker set	open speaker set
oracle NMF	5.06	-
DC oracle k-means	6.54	6.45
DC oracle spectral	6.35	6.26
DC global k-means	5.87	5.81

表格 1 可以看出，新的模型均比基线的 5.06dB 高（暂时不理解 oracle 的处理思路）

Table 2: SDR improvements (in dB) for different embedding dimensions K and activation functions

	closed speaker set		open speaker set	
model	DC oracle	DC global	DC oracle	DC global
$K = 5$	-0.77	-0.96	-0.74	-1.07
$K = 10$	5.15	4.52	5.29	4.64
$K = 20$	6.25	5.56	6.38	5.69
$K = 40$	6.54	5.87	6.45	5.81
$K = 60$	6.00	5.19	6.08	5.28
$K = 40$ logistic	6.59	5.86	6.61	5.95

表格 2 中 K 是嵌入层的维度，K=5 效果很差，优化失败，K=20，40，60 的结果很接近，使用的激活函数为 tanh，因为理论上说使用 tanh 比 logistic 有更大的嵌入空间，但实际是使用 logistic 的效果会微小提升（看最后一行）。

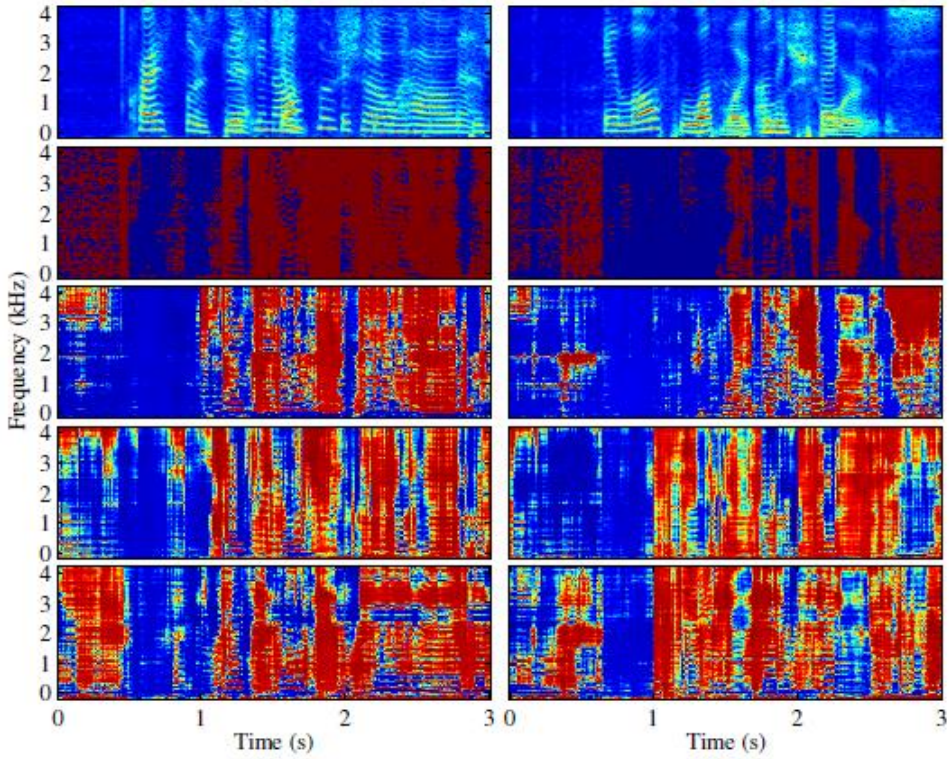

Figure 2: Examples of embeddings for two mixtures: f+f (left) and f+m (right). 1st row: spectrogram; 2nd row: ideal binary mask; 3rd-5th row: embeddings.

图 2 显示了一些不同的混合音频（女-女，男-女）在嵌入向量的某几个维度上的 T-F 单元特征，用于显示它们对每个信号的不同层次的敏感程度。

Table 3: SDR improvement (in dB) for each type of mixture. Scores averaged over male-male (m+m), female-female (f+f), female-male (f+m), or all mixtures.

	training gender	closed speaker set				open speaker set			
method	**distribution**	m+m	f+f	f+m	all	m+m	f+f	f+m	all
oracle NMF	speaker dependent	3.25	3.31	6.53	4.90	-	-	-	-
DC oracle permute	*whole*	3.79	4.29	9.04	6.54	4.49	3.21	8.69	6.45
	balanced	3.89	4.35	8.74	6.42	4.61	3.49	8.27	6.41
	female	-	5.03	-	-	-	4.04	-	-
DC global *k*-means	*whole*	2.54	2.85	9.07	5.87	3.51	1.42	8.57	5.80
	balanced	2.78	2.87	8.63	5.72	3.89	1.74	8.27	5.83
	female	-	3.88	-	-	-	2.56	-	-

Table 4: SDR improvement (in dB) for three speaker mixture

method	**closed speaker set**	**open speaker set**
oracle NMF	4.42	-
DC oracle	3.50	2.81
DC global	2.74	2.22

男女混合音频的分离比同一性别的混合音频分离要简单，从表格 3 中也可以看出，训练数据的性别比例越平衡，表现性能越好，不过如果仅仅关注女性，表现也依旧良好。

表格 4 和图 1 指出的是训练用双人测试用 3 人，且效果不错，文中的模型可以适用于音频源数量不确定的情况（不是说话人数目不确定），因此能更好适应真

实世界。另外，实验也做了 3 人混合音频的训练和测试，SDR 大约为 6.15dB 左右。

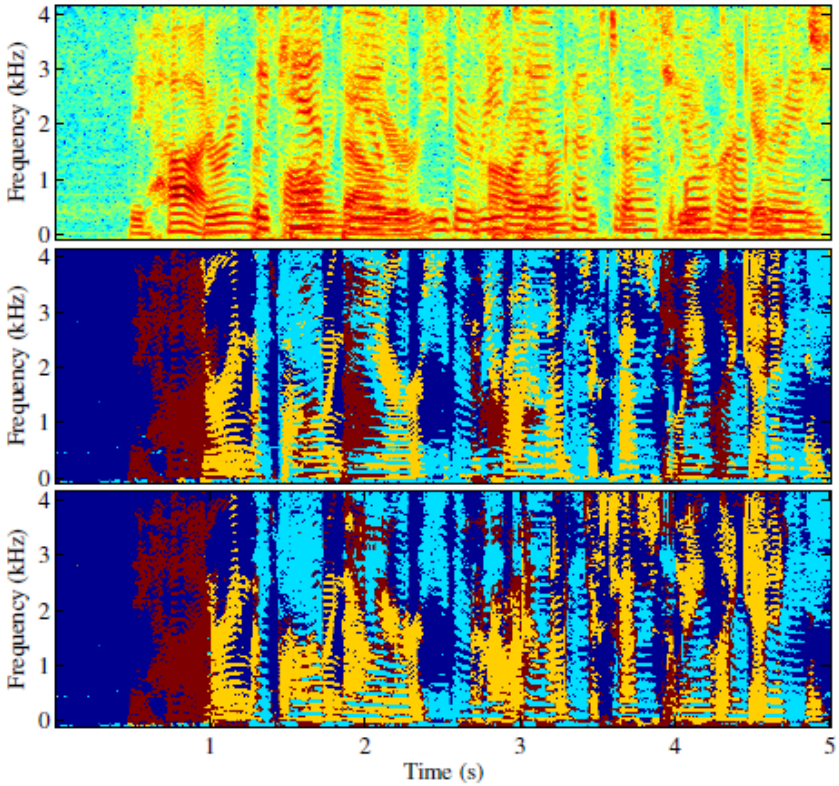

Figure 1: An example of three-speaker separation. Top: log spectrogram of the input mixture. Middle: ideal binary mask for three speakers. The dark blue shows the silence part of the mixture. Bottom: output mask from the proposed system trained on two-speaker mixtures.