VOIP流中使用CNN-LSTM下对QIM的隐写分析方法

翻译

nooobme 于 2020-01-03 15:37:43 发布 752 收藏 1
分类专栏： # 语音隐写分析
原文链接：https://doi.org/10.1145/3335203.33357
版权

语音隐写分析 专栏收录该内容
6 篇文章 2 订阅
订阅专栏

一、介绍

CNN能够从时间或空间数据中学习局部响应，但缺乏学习序列相关性的能力，而RNN能够处理任意长度的序列并捕获长期上下文依赖性[5，15]，本文指出了利用这两种结构的一种适当方法，并提出了一种新的CNN-LSTM VoIP流的QIM隐写检测模型。在该模型中，采用双向长短期记忆递归神经网络从语音中提取长期上下文信息，采用不同核尺寸的CNN层提取每个语音帧的局部特征。最后，利用全连通层和软最大层作为分类器来计算类概率。此外，我们的模型以声码器后的量化指标序列作为输入。

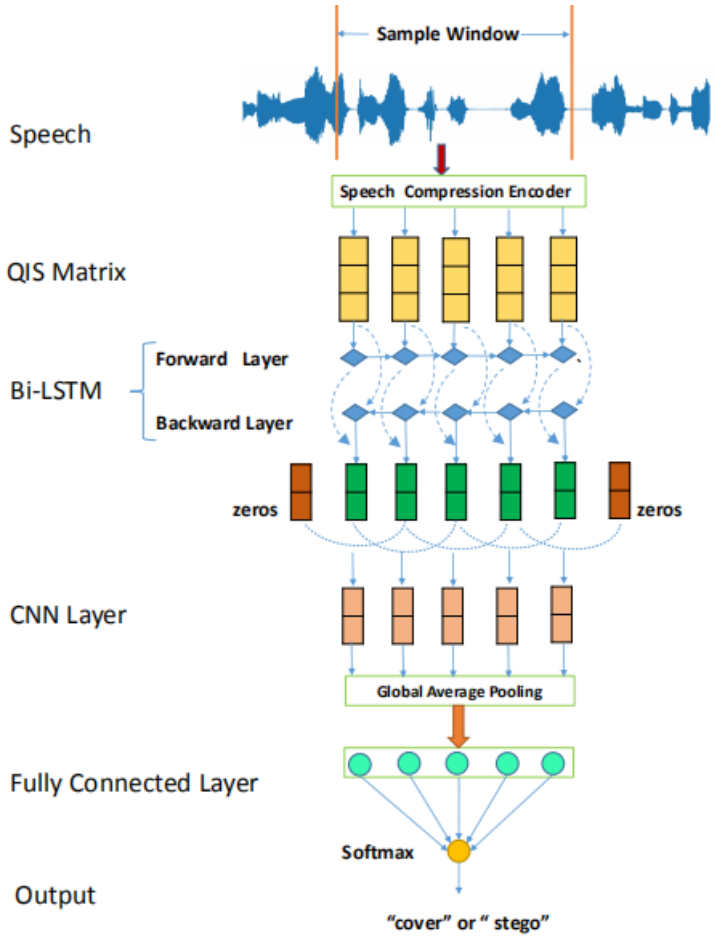

Figure 1: Structure of our proposed CNN-LSTM model

二、架构

2.1 量化指标序列矩阵QIM

量化指标序列（QIS）矩阵是我们网络的输入。模拟语音信号经滑动窗口采样后，用语音压缩编码器进行压缩。在压缩过程中，在预处理阶段对输入信号进行高通滤波。一个n阶线性预测分析得到一组LP滤波器系数，这些系数被转换成线谱对（LSP）并使用VQ量化。在矢量量化过程中，通过综合搜索法分析选择激励信号，根据感知加权失真最小化原始语音和重构语音之间的误差。开环基音搜索和闭环基音搜索构成了重叠候选向量的自适应码本搜索步骤。

经过代数码本搜索，生成量化索引序列。QIM隐写技术根据嵌入的数据将VQ量化码本分成若干部分，这将改变QIS的特性。在大多数情况下，QIS矩阵可以表示为 $QIS = \begin{bmatrix} s_{1,1} & s_{2,1} & s_{3,1} & & s_{T,1} \\ s_{1,2} & s_{2,2} & s_{3,2} & \cdots\cdots & s_{T,2} \\ s_{1,3} & s_{2,3} & s_{3,3} & & s_{T,3} \end{bmatrix}$。其中T是语音样本窗口中的总帧数。si1、si2和si3分别是第i帧语音的码字索引。以G.729A编码器为例，si1、si2和si3分别为7bit、5bit和5bit。由于基于qim的隐写术只改变了每帧中si1，si2和si3的范围，因此QIS矩阵包含了完整的信息.

2.2 双向LSTM层 bidi LSTM layer

我们的网络的第一层是双向LSTM层，并将QIS矩阵输入其中。给定输入序列x=（x1，…，x t），标准RNN通过从t=1到t的迭代，计算隐藏向量序列h=（h1，…，hT）和输出向量序列y=（y1，‥，yT）：

$$h_t = H(W_h \cdot [h_{t-1}, x_t] + b_h),$$

$$y_t = W_{hy} \cdot h_t + b_y,$$

。其中W项表示权重矩阵，b项表示偏差向量。H是隐层函数，通常是一个sigmoid函数的元素级应用。

RNN的一个吸引人之处在于，它们可能能够将以前的信息连接到当前任务，例如使用以前的帧可以提高对当前帧的理解。在语音处理中，所有的话语都是同时被转录的，没有理由不好好利用未来的语境。因此，在[29]中提出的双向RNNs（BiRNNs）通过两个独立的隐藏层处理双向数据，然后将其转发到同一输出层来实现。BiRNNs通过迭代后向层从t=t到1，前向层从t=1到t，然后更新输出层来计算前向隐藏序列→h，后向隐藏序列←h和输出序列y。

$$\overrightarrow{h_t} = H(W_{\overrightarrow{h}} \cdot [\overrightarrow{h_{t-1}}, x_t] + b_{\overrightarrow{h}}),$$

$$\overleftarrow{h_t} = H(W_{\overleftarrow{h}} \cdot [\overleftarrow{h_{t-1}}, x_t] + b_{\overleftarrow{h}}),$$

$$y_t = W_{\overrightarrow{h}y} \cdot \overrightarrow{h_t} + W_{\overleftarrow{h}y} \cdot \overleftarrow{h_t} + b_y.$$

应用RNNs解决各种问题取得了令人难以置信的成功。然而，RNN很难处理长期依赖性[2]。值得庆幸的是，在[7]中使用专门构建的存储单元来存储信息的长-短期内存（LSTM）体系结构在解决这个问题方面更为出色。对于[6]中使用的LSTM版本，H由以下复合函数实现

$$\begin{aligned}
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i),\\
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f),\\
q_t &= \tanh(W_q \cdot [h_{t-1}, x_t] + b_q),\\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o),\\
c_t &= f_t \odot c_{t-1} + i_t \odot q_t,\\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}$$

。其中σ是logistic-sigmoid函数。LSTM有三个门，包括输入门it、遗忘门ft和输出门ot。另一个标记是ct作为细胞激活载体。所有的门和激活向量的大小都与隐藏向量h相同。从单元到门向量的权重矩阵是对角的，因此每个门向量中的m元素只接收来自单元向量m元素的输入。

BiRNNs与LSTM相结合，给出了双向LSTM[29]，它不仅可以发现和利用远程上下文，而且可以充分利用语音载体的双向信息。在该模型中，双向LSTM层也可以作为一个转换层，逐步建立更高层次的声学数据表示。在该层之后，为下一步生成新的代表帧序列向量。

2.3卷积层

卷积层是我们模型的第二部分。由于CNNs能够捕获空间或时间结构的局部相关性。在语音建模方面，CNNs不仅可以模拟时域和频谱的局部相关性，而且可以获得平移不变性，在许多语音任务中也得到了应用，并在文献[27,31,41]中进行了成功的尝试。在该模型中，我们选择使用不同窗口大小的一维卷积来捕获不同尺度的特征。一维卷积包括在序列上滑动的滤波器向量和在不同位置检测特征。设h∈R^d为双向LSTM层生成的第i帧的d维帧向量。设h∈R^(L×d)表示输入片段，其中L是语音帧的个数。设k为滤波器的长度，向量m∈r^(k×d)为卷积运算的滤波器。对于语音帧中的每个位置j，我们有一个具有k个连续帧向量的窗口向量wj，表示为 $w_j = [h_j, h_{j+1}, \cdots, h_{j+k-1}]$. 这里，逗号表示帧向量连接。滤波器m以有效的方式与窗口向量在每个位置卷积以生成特征映射g∈ R^(L−k+1)×d，其中窗口向量gj的每个元素gj的特征映射如下: $g_j = f(w_j \circ m + b)$。式中，◦是按元素进行的乘法运算，b∈R是偏项，f是一个非线性变换函数，可以是sigmoid，双曲正切，ReLU等。本研究选择ReLU作为非线性函数。

在我们的模型中，我们使用多个过滤器来生成多个特征映射。对于长度相同的n个过滤器，生成的n个特征映射可以重新排列为每个窗口wj的特征表示 $W = [g_1; g_2, \cdots; g_n]$ 这里，分号表示列向量连接，gi是使用第i个过滤器生成的特征映射。W∈R^（L-k+1）×n的每一行Wj是由n个滤波器为位置j处的窗口向量生成的新特征表示。

为了降低卷积后的维数或选择最重要的特征，通常对卷积后的特征映射采用一个池层，卷积后主要采用平均池和最大池。在我们的模型中，我们采用全局平均池来减少特征维数。

模型的最后一部分是全连通层，利用软最大激活函数作为分类器来判断样本是否属于“覆盖层”。此外，为了加速收敛和克服过拟合问题，我们的模型使用了, Batch Normalization [14]和 Dropout [30]。

三、实验

数据集使用：RNN-SM的公开数据集，该数据集包含41小时的中文语音和72小时的英文语音，采用PCM格式，每个样本16位来自互联网。不同的样本包含不同类型的母语人士。这些语音样本构成了封面语音数据集。
对于覆盖语音数据集中的每个样本，应用G.729A得到QIS矩阵。使用CNV-QIM隐写方法嵌入秘密数据[36]。嵌入的样本构成了隐写语音数据集。嵌入率定义为嵌入比特数与整个嵌入容量的比率。当在隐写术中进行a%嵌入率时，我们以a%的概率嵌入每个帧。在我们的实验中，我们将a设为20，40，60，80来产生不同嵌入率的样本。概率表示帧选择的随机性。有嵌入秘密数据的样本由类别标签“stego”分配，没有嵌入秘密数据的样本由类别标签“cover”分配。
此外，修剪长度是影响检测精度的另一个因素。实验中，将覆盖语音数据集和stego语音数据集中的样本分为0.1s、0.3s、0.5s、2s和6s，测试不同持续时间下的模型性能。相同长度的节段是连续不重叠的。对于0.1s剪辑的训练集，有2486708个1:1比例的封cover剪辑和stego剪辑的样本。310810个修剪用于测试和验证。

模型使用：我们的模型中的超参数是通过对试验集的交叉验证来选择的。更具体地说，BiLSTM隐藏态的维数为64，CNN滤波器的窗口大小分别为3、4、5。每个CNN过滤器的数量是128个。全连通层的维数为64，全连通层的漏失率为0.6。批量大小为256，最大训练时间设置为100。我们使用Adam[16]作为网络训练的优化器。我们的模型是由Kera实现的。模型性能由分类精度来评估，分类精度定义为正确分类的样本数与样本总数的比率

模型对比：与IDC[21]、QCCN[20]、RNN-SM[23]等隐写分析方法作对比。

通过对不同模型的比较，可以得出IDC和QCCN将手工特征与传统的机器学习算法性能相结合的结论。同时，深度学习方法RNN-SM和我们的模型在检测精度上有了显著的提高。我们还注意到，在大多数情况下，当嵌入率和嵌入时间相等时，所有模型在英语语音样本中的性能都优于汉语语音样本。这一现象可以用两种语言的字母表、语法、语音等不同特征来解释。尤其是音系学可能是解释这一结果的最重要因素。因为汉语有412种音节，而英语有20个元音和28个辅音。这种多样性使得汉语的关联关系更加复杂。

样本长度的影响：在VoIP流中检测基于QIM的隐写术时，语音的持续时间是一个重要因素，因此，我们将嵌入率固定在20%，并研究了片段长度的影响。精度随着样本长度的增加而增加。这种现象很容易解释，较长的序列提供了更多的码字相关性的观察，从而可以更准确地建模。因此，stego语音的码字相关模式和覆盖语音的码字相关模式之间的差异更加明显，从而使分类更加容易。此外，当样本长度较小时，增加样本长度可显著提高精确度。随着样本长度的增加，增加样本长度的好处减小。最重要的是，我们可以得出结论，我们的模型比以前的所有方法都好。

嵌入率的影响：嵌入率是影响检测精度的重要因素。当嵌入率较低时，随着嵌入率的增加，精度显著提高。当嵌入率大于40%时，检测准确率达到95%以上。同时，该模型在嵌入率较低的情况下，显著提高了检测精度。一般情况下，为了避免被检测到，隐写算法通常采用低嵌入率策略，这给隐写分析带来了挑战。我们的模型在低嵌入率下的优异性能使得它在现实场景中更加实用。

四、总结

本文提出了一种将CNN和LSTM相结合进行隐写分析的方法，特别是CNN-LSTM网络用于VoIP流上基于QIM的隐写检测。该模型充分利用了LSTM和CNN两种主流结构，利用双向LSTM捕获语音的长时间上下文信息，在语音载体中生成更好的帧向量表示。而CNN随后被用来捕捉局部特征以及全局和时间语音特征。实验证明，与以往在voip流上检测基于QIM的隐写术的方法相比，该模型能够达到目前的效果。此外，我们的模型是一个实用的有效模型，可以进一步推广。

五、参考文献

[36] Bo Xiao, Yongfeng Huang, and Shanyu Tang. 2008. An Approach to InformationHiding in Low Bit-Rate Speech Stream. In Global Telecommunications Conference,2008. IEEE GLOBECOM. 1
[20] Songbin Li, Yizhen Jia, and C. C. Jay Kuo. 2017. Steganalysis of QIM Steganography in Low-Bit-Rate Speech Signals. IEEE/ACM Transactions on Audio Speech &Language Processing 25, 99 (2017), 1–1.
[21] Song Bin Li, Huai Zhou Tao, and Yong Feng Huang. 2012. Detection of quantization index modulation steganography in G.723.1 bit stream based on quantizationindex sequence analysis. Journal of Zhejiang University-Science C(Computers &Electronics) 13, 8 (2012), 624–6.
[23] Zinan Lin, Yongfeng Huang, and Jilong Wang. 2018. RNN-SM: Fast Steganalysis of VoIP Streams Using Recurrent Neural Network. IEEE Transactions on InformationForensics & Security PP, 99 (2018), 1–1.