E-008

グループディスカッション可視化のための話者クラスタリング

Song Yu(法政大学大学院情報科学研究科)　伊藤　克亘 (法政大学情報科学部)

1　はじめに

近年、ディープラーニングは音声認識技術に大きな影響を与えている。2012 年に、論文 [3] は、ニューラルネットワーク隠れマルコフモデルに基づくハイブリッド音響モデルを提案した声認識システムの性能を大幅に改善し、従来の GMM-HMM フレームワークを迅速に置き換え、主流の音声認識システムの標準となる。その後、簡単なフィードフォワードネットワークより、もっと複雑なニューラルネットワークモデルが提出されました。例えば、Deep Belief Network (DBN) は、認識性能をさらに向上させる。

最近の研究の結果、ディープラーニングは機械の学習と人工知能の分野で良好な結果を得て、特に画像と音声処理の分野であることが証明されます。2011 年以降、マイクロソフト研究院と Google の音声認識研究者は、DNN 技術を採用して、音声認識エラー率 20%～30%を低減し、音声認識分野で 10 年以上の最大の突破性の進展である。

今、ディープラーニング方法は多くのアプリケーションの中で良い効果を得て、ディープラーニングを構築する重要な方法として、教師なし学習技術の役割を無視することはできない。教師なし学習技術は、ラベルがない場合に、自主的にデータを学ぶことができる抽象的な形式であり、学習の範囲を広げるだけでなく、ニューラルネットワークに優れた初期化パラメータを提供することができ、そのためには、教師なし学習技術の機理と方法を理解し分析し、理解と拡大の深さに対して非常に重要な意味を持っている。

2　関連研究

2.1　BM と RBM

単層フィードバックネットワークとしては、ネットワーク中のノードは他のノードにもつながっていて、もしシステムの中に N 個のノードがあるとすれば、各ノードの状態が ON または OFF となる可能性がある。その標識を $(s_1, s_2, ..., s_N)$ とする。[3]

確率モデルとして、Boltzmann Machine(BM) の各ノードの状態はランダムです。 確率はネットワークのエネルギー変化に依存し、ニューロン i と j の間の接続重みは W_{ij}、ニューロン i のしきい値は Q_i とし、ネットワークのエネルギー関数の定義は

$$E = -\sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}s_i s_j + \sum_{i=1}^{N} Q_i s_i \tag{1}$$

ニューロン s_i の状態が変化すると、その状態が 0 から 1 に変化すると仮定すると、結果として得られるエネルギー差は

$$\Delta E_i = E_{i=0} - E_{i=1} = \sum_{i=1}^{N} W_{ij} - Q_i \tag{2}$$

BM の変化過程はマルコフ過程で、ある瞬間の温度を T とすると、他のニューロンが変化しないときにニューロン s_i が 1 になる確率は

$$P_{k(s_k=1)} = \frac{1}{1 + e^{-\frac{\Delta E_i}{T}}} \tag{3}$$

$\Delta E_i > 0$ のとき、$P_{k(s_k=1)} > 1/2$、つまり、システムは常に低いエネルギーから低い確率で状態に変化することがわかる。ボルツマンマシンは確率的に移動し、低エネルギー状態に移行する可能性がより高いが、高エネルギー状態に移行する可能性もあるが、この可能性は温度が下がるにつれて低下する。 一般に高温から始まり、温度が高いとネットワークは小さなエネルギー差を無視し、大域的な状態の構造を大まかに求め、近似的な最小解を求め、そして温度が下がるにつれて急速に最小解に収束する。

最終的に熱平衡に達すると仮定すると容易に検証できるが、このとき BM ネットワークの 2 回の状態はそれぞれ αとβであり、αとβの発生確率は

$$\frac{P_\alpha}{P_\beta} = e^{-\frac{(E_\alpha - E_\beta)}{T}} \tag{4}$$

つまり、ボルツマン分布が成り立ち、数学的性質が良く、情報理論と密接に関係している。Restrict Boltzmann Machine(RBM) モデルは、2 層のニューラルネットワークで、可視層と隠れ層に分かれていますが、BM の一部のノードを可視層として使用すると、その一部が隠れ層になる。 可視層の間、隠れ層の間、可視層と隠れ層の間は、トレーニングおよび学習中の計算をより複雑である。 RBM では、可視層と隠れ層の間に接続はないが、可視層と隠れ層のみが接続されているため、トレーニングと学習が簡単になる。

2.2 DBN

2006 年に Geoffrey Hinton によって提案された Deep Belief Network(DBN) は、RBM に基づく信頼ネットワークを構築することである。 ディープネットワークの難易度トレーニングの欠点はディープラーニングの流行を切り開いています。

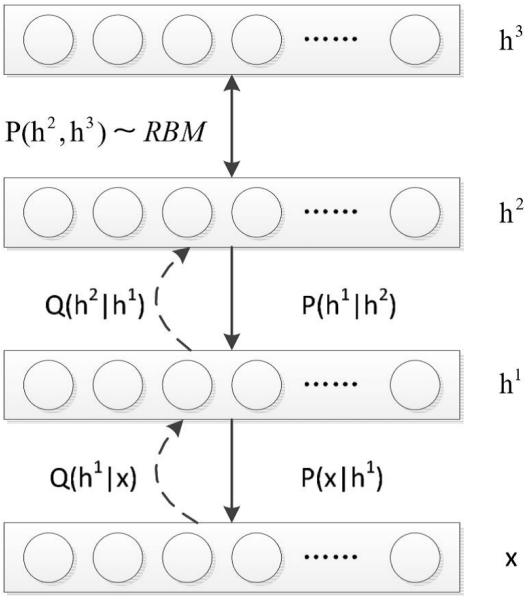


図 1: DBN のアーキテクチャ

上の 2 つの層は無向接続を使用して連想記憶を形成し、残りの層は有向接続を使用する。 ボトムアップを使用して認識を生成する場合、下方向の重みは上から下にデータを生成するための重みを生成し、下の層はトレーニングデータによって決定される可視層で、各ニューロンは可視層ベクトルを表す。 一次元 DBN 事前トレーニングプロセスは層ごとに行われ、各層において、隠れ層を推論するために可視層が使用され、この隠れ層は次の層（上位層）の可視層として扱われる。DBN のトレーニングプロセスは以下のとおりです。

1. 入力として訓練データを用いて基礎となる RBM を訓練する。

2. 上記のステップで生成された隠れ層状態は、その層の入力トレーニングＲＢＭである。

3. モデルに必要な隠れ層の数が生成されるまで (上の 2 層を除く) ステップ 2 を繰り返します。

4. 最上位の 2 層 RBM トレーニング方法ではデータのラベル付けを考慮する必要がありますトレーニングセット内のデータにラベルがある場合は、最上位の RBM トレーニングで分類ラベルを表すニューロンをまとめてトレーニングする必要がある。

DBN のチューニングプロセスはスタックベースのセルフエンコーディングと似ており、パラメータを最小化するためにクロスエントロピーを犠牲にして最適化することができます Geoffrey Hinton はそれをウェイクフェーズとスリープフェーズの 2 つのフェーズに分ける。

1. ウェイクフェーズ：外部特徴と上向きの重み (コグニティブ重み) を介して各レイヤ (ノード状態) の抽象表現を生成し、レイヤ間の下りリンクの重み (重みの生成) を変更するために勾配降下法を使用するコグニティブプロセス。

2. スリープフェーズ：生成プロセスでは、最上位レベルの表現 (Wake アップ時に習得した概念) と下位ウェイトを使用して、レイヤ間の上位ウェイトを変更しながら、基礎となる状態を生成する。

3 相関実験

3.1 IBM の音声認識

実験で使用した録音では、それぞれの録音に 5.5 メートルと 0.1 メートルの距離で 2 つのバージョンが録音され、合計 20 録音が収録されている。音声認識に用いられる音声モデルは、Japanese broadband model（16KHz）です。各記録を順に試験し、以下の結果を得た：

表 1: Recording at 0.1M

No.	Quantity	Drop	Substitution	Accuracy
1	8	0	1	0.87
2	7	0	0	1.00
3	12	0	0	1.00
4	10	0	2	0.80
5	8	0	0	1.00
6	9	1	0	0.89
7	10	0	0	1.00
8	6	0	0	1.00
9	4	0	0	1.00
10	12	1	0	0.92

平均正解率:0.948

表 2: Recording at 5.5M

No.	Quantity	Drop	Substitution	Accuracy
1	8	0	0	1.00
2	7	1	3	0.43
3	12	0	1	0.92
4	10	0	1	0.90
5	8	0	0	1.00
6	9	1	0	0.89
7	10	0	2	0.80
8	6	0	0	1.00
9	4	0	1	0.75
10	12	1	2	0.75

平均正解率:0.844

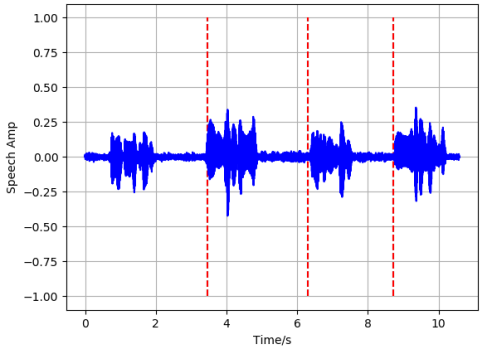
一般に、録音に含まれる単語が多いほどミスが多く、遠距離録音は近距離録音よりも正解率が低い。録音の第 4 セグメントの正解率は、遠距離録音は近距離録音の正解率よりも高く、同様の発音による不明瞭さが原因である可能性がある。第 6 セグメント録音の 2 つのバージョンの識別結果はすべて同じで、「は」が欠けている。録音を聞い、はっきりと聞こえなかった。 理解できるが、識別が困難な場合がある。

3.2 セグメンテーションポイントのスクリーニング

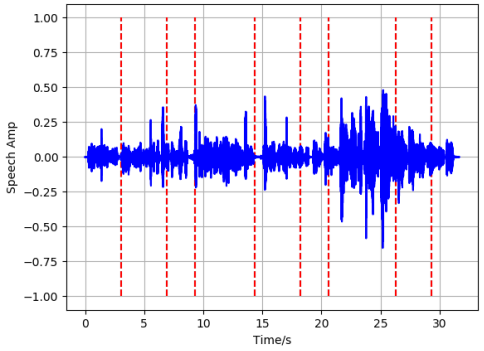
ベイズによる音声セグメンテーションプログラムを試す。音声セグメンテーションを行う前に、音声から MFCC 特徴を抽出する必要がある。簡単に使用できる Python ライブラリ Librosa がある。これはオーディオ信号解析専用のライブラリで、MFCC コンピューティングインタフェースを提供する。

マルチセグメンテーションポイントの問題を解決するには、VAD を使用してセグメンテーションポイントをフィルタリングする。Multi segmentation が終了するセグメンテーションポイントに従って音声セグメンテーションを実行する。VAD 検出は、音声の各セグメントについて実行され、VAD の検出が音声のエンドポイントを有する場合、処理は実行されず、VAD の検出が音声のエンドポイントを有さない場合、セグメンテーションポイントは除去される。

選択したオーディオファイルのサンプリングレートは 16KHz で、オーディオには 4 つの会話がある。結果から、対話は 4 つのセグメントに分けられ、セグメンテーションポイントも表示されることが分かる。無音部分があるの音声ファイルに対して、このプログラムに良い結



果がもらえる。しかし、無音部分がない音声ファイルの場合、プログラムの結果にいくつかの偏差がある。この



結果に基づいてオリジナルの録音ファイルを処理する。

3.3 特徴量の処理

音声のセグメントから MFCC 特徴量を抽出して学習材料としてニューラルネットワークを入力する。オリジナル録音ファイルを分割して得たセグメントの長さは 0 - 4 秒の間にある。Matlab でセグメントの MFCC 特徴量を抽出します。メルフィルタの次数を 24、fft 変換の長さを 256、サンプリング周波数を 16000 Hz に設定する。音声のセグメントの長さが異なるため、抽出された MFCC 特徴量の次元も異なる。入力ニューラルネットワークの特徴量は同じ次元数である必要があり、MFCC 特徴量の次元数を一致させるためには特徴量を処理する必要がある。フィルタには 24 次があるため、MFCC の特徴量あたりの行数は異なるが、24 列あります。列数が一致しているので、すべての特徴量の次元が一致するように MFCC 特徴量行列列の共分散行列を計算できる。特徴量マトリックスは一次元に拡張され、全ての一次元マトリックスは一緒にステッチされる。 行列の各行が音声セグメントの MFCC 特徴量を表すとする。これを特徴量としてニューラルネットワークを入力する。

4　話者認識実験

現在，多くの教育機関でグループワークを取り入れた授業がされている。ディスカッションのテーブルごとに話している内容が違うため、アシスタントが各テーブルの内容をすぐに把握してアドバイスことが困難である。そこで私は会話記録システムを作りたいと思う。ディスカッションの内容をビデオ, 音声, 文字などの情報で整理することができる。今、だれが何を話したか判断するための話者認識を中心に研究を進めている。グループワークの時の話者認識に問題がある。マイクを一人一人に用意，装着するのはコストがかかるって大変である。録音するとき，同じテーブルの人の声が入る、他のテーブルのディスカッションの声も入るため、雑音が多いである。特定の話者ではない。テーブル，授業ごとに発話者が変わる。したがって、マイクはテーブルの中央に配置する必要があります。雑音を考慮し、少ないデータで識別できます。

DBN で話者認識を行うと思う。DBN は、ニューラルネットワークの一種である。 DBN はニューロンのいくつかの層から成り、このニューロンは RBM である。前の RBM の隠れ層は次の RBM の表示層であり、前の RBM の出力は次の RBM の入力である。 トレーニングプロセスの間、上位層の RBM は、最後の層まで現在の層の RBM をトレーニングするために完全にトレーニングされる必要がある。多くの場合、DBN は教師なし学習フレームワークとして使用され、音声認識で良好な結果を達成する。この論文 [3] では、ニューロンの数が一定数に達すると、深層ネットワークモデルが通常の BP ネットワークよりも優れており、ネットワーク規模の拡大に伴ってその性能が向上することが実験により証明された。

Matlab のディープラーニングツールボックスを使用すると、DBN を簡単に試すことができる。MFCC 特徴量を matlab に必要な学習材料としてニューラルネットワークに入力する。テスト集合と予測集合で比較して、誤り率が出た。数回の繰り返し通じて、モデルに誤り率が次第に下がっています。50 回目の時、誤り率が 4%になりました。

表 3: 結果評価

epoch	time	mean squared error	train error
1/50	0.0054	0.4945	0.2495
20/50	0.0053	0.0566	0.2411
50/50	0.0076	0.0396	0.0408

5　まとめ

今まで、DBN に理解してくれて、そしてディープラーニングツールボックスを通して勉強した。オリジナル録音ファイルを分割し、MFCC 特徴量を抽出し、DBN ネットワークを入力して話者識別を行う。識別率を得たが、まだ少し足りないものがある。その後、私はプログラムを修正して、誤り率をさらに低下させる。現在、音声ファイルには雑音が含まれているが、これは識別結果に影響がある。次、音声の雑音を下げてから実験すると思う。

参考文献

[1] HINTON G,DENG L,YU D,et al.Deep neural networks for a coustic modeling in speech recognition:The shared views of four research group [J].IEEE Signal Processing Magazine,2012,29(6):82-97.

[2] YIN Rui-Gang, WEI Shuai, LI Han, YU Hong.Introduction of Unsupervised Learning Methods in Deep Learning

[3] ZHAO Yan，Lu Liang ，ZHAO Li.Research on Speaker Identification Based on Improved Deep Neural Network