ResNeXT and Res2Net structures for Speaker verification

背景

ResNet ベースの手法は話者うめ込みの抽出に広く用いられてきた.
(テキスト非依存の話者照合システムにおける.)

ResNet は
- CNN への残差接続を用いる.
- 残差ブロックを正規化する

ことで,高い照合精度を得る.

が,
入力特徴量空間が複雑になると,
ResNetネットワークの深さと幅が増えるだけで,潜在的なパフォーマンスを発揮できていなかった.

→ResNetの拡張をていあん.

この論文の時点(2020)では,バックエンドはCNN(ResNet-based)が主流だった.

論文の主旨

話者照合タスクに向け,ResNetの拡張である"ResNeXt"と"Res2Net"をていあんする.
- ResNeXtとRes2Netはもともとimage recognitionでていあんされていたもの.
- モデルの表現力を上げるために深さと幅に加え,cardinary と scale という2つの次元を導入.
- Res2Netは,Scaleの次元を上げることで,さまざまな粒度で特徴を表現でき,短い発話での照合は良くなるらしく.

3つのタスクで検証.
- VoxCeleb テストセット:ResNeXt,Res2Netは,従来のResNetよりかなり良い性能を.
  - Res2NetはEERを18.5%減.
- 他2つ(内製セット.不一致条件の)
  - ノイズとセグメント(発話?)長の変化に対して汎化Up.

ResNetについて.

width,depthの2つを学習によって調整し,モデルの表現力を上げる.

入力特徴量空間がフクザツになると,
学習時とテスト時の条件に多くのミスマッチが生じ,
depthとwidthを増やすだけではoverfittingする傾向がある.
(汎化能力down)

ResNeXtについて.

ResNeXtブロックは1つの層に多くの方向組み
グループを導入し,ResNetの残差ブロックを
マルチブランチ変換にreplace.

入力チャンネルを複数のグループに分け別々に変換がほどこされる.

実験結果

データセットについて

VoxCelebと、MS内の2つのテストセット。

- VoxCeleb
  - テキスト非依存、いろんな環境下の会話。
  - 多くの音響環境、短時間会話を含む。
  - 検証に用いたモデルは、VoxCeleb2-devのデータを用いている。
    - データ拡張(babble, music, noise, reverb)で学習データを拡張。
  - 評価は
    - VoxCeleb1 test.
    - VoxCeleb1-E (extended)
    - VoxCeleb1-H (hard, includes 同じ性別, 同じ国籍)

    の3つ。
- MS-SV test set
  - MS内製。
  - 近接マイクで、短文の発話をとる。
  - MSの会議室での会話をマイクでとり、処理したもの。
  - 約150人。
  - 2-15 sec 発話(avg. 4sec)
- Cortana テストセット
  - MS.
  - for text-dependent SV.
  - 183 speakers
  - start with "Cortana" or "Hey Cortana"
  - いろいろなノイズ環境下、いろんなキョリでろく音。
  - 4人のとうろく会話は、きれいな環境下で、マイクから1mのキョリ。
  - CTCベースのキーワード検出を用いて、発言をきり出く。

実験

1. on VoxCeleb1

   1.1. Regular test set
   → table2.

   - ベースラインのモデルは、test, E, Hについて、EERは 1.78, 1.76, 3.07である。
   - ResNeXtは testセットで 1.69
     Res2Net 〃 1.60 で、ベースラインより良い。

   - - -

   - モデルの複雑さをUpすると、
     - ベースライン(deeper, wider)はそんなに性能上がらん。(+50% complexity)
     - Res2Net-26w6sは伸びた。(+50% complexity)
     - +50% complexity、Res2Net-26w8sがベスト。

     → Res2Netを用い、cardinalityとscaleを導入すると効果的。
     特に、スケールを上げるのが良い。

1.2. テストセットの発話を短く切って検証

VoxCeleb1 test 内の発話は平均8s.

→ 2, 3, 4 sec に切って検証.

→ table 3.

・発話長が短いと性能が低下する.

が, ていあん手法の方がEERが低い.(特にRes2Net)

→ ベースラインより良い.

2. NS-SV.

- モデルの汎化性能を見る.

・発話は 20sec 長なので, 1-2s, 2-4s, >4s のセグメントに分割した.

→ table 4.

ていあん手法はベースラインより EER 低い.

3. Cortana set.

- VoxCeleb でトレーニングしたモデルを使う.

- トレーニング ⟷ テスト では発話内容不一致

どうろく ⟷ テスト 〃 一致する.

→ table 5.

全体として, ノイズがあるほど, マイクが遠いほど, 精度はさがる.

ちかい場合はRes2Net はResNetより 4.5%向上.

4.20 ← 4.40

quiet and TV でも向上.

→ Res2Net はResNetより精度良い.

が, ResNext はベースラインより精度低い.

→ Res2Net ほどロバストではない

結論

ResNeXt とRes2Net の話者照合タスクにおける有用性を示した.

・VoxCeleb テストセットによるデモでは, どちらもベースライン手法を上回り.

特にRes2Net が優れていた.

- 短い発話や, ミスマッチな条件下においても good.