CE-008

# マルチモーダル Transformer エンコーダに基づく音声とテキストに含まれる複数感情の認識の評価

Evaluation of Multiple Emotion Recognition in Speech and Text based on Multimodal Transformer Encoder

仁平正彦 *1　渥美雅保 *1
Masahiko Nihira　Masayasu Atsumi

*1創価大学大学院理工学研究科情報システム工学専攻
Information Systems Science, Graduate School of Science and Engineering, Soka University

## 1. はじめに

感情表現の認識は，商業・教育など様々な分野へ応用されており，感情を伴う発話認識・生成や，感情的心理状態の可視化に重要な要素だと考えられる．近年では，“音声のみ”や“テキストのみ”から感情表現を認識するユニモーダルモデルではなく，複数のモダリティを用いたマルチモーダルな感情表現認識を行う研究が盛んに行われている [Delbrouck 20][Shiwardhana 20]．しかし，マルチモーダル感情表現認識には未だ多くの課題が存在している．本研究では，音声モーダルの特徴量抽出方法の改善，実際の音声対話をする状況を想定した認識精度の評価，音声に含まれる複数感情をセグメントに分けて認識する処理，の 3 つの課題に着目し，音声の wav2vec 2.0[Baevski 20] によるエンコーディング，及び音声認識モデル Whisper[Radford 22] による文字起こしテキストの RoBERTa[Liu 19] によるエンコーディングを入力とする音声・テキスト特徴間のクロスモーダルアテンション Transformer[Vaswani 17] に基づく感情表現認識モデルを提案する．また，本モデルでは，長い音声データは入力可能なサイズに分割し，分割した音声毎に感情表現認識をして合議を取る．そこで，音声のセグメント毎に異なる感情ラベルを含む小規模データセットを作成し，合議において音声データに含まれる複数の感情表現それぞれが現れるセグメントを適切に識別可能か検証する．

## 2. 関連研究

テキストと音声を使ったマルチモーダル感情表現認識の研究として，“BERT-like”自己教師あり学習 (SSL) アーキテクチャを用いた研究 [Siriwardhana 20] がある．ここでは，単純な融合手法と複雑な融合手法について認識性能を評価している．単純な融合手法では，Speech-BERT と RoBERTa を感情表現認識のタスクに合わせてファインチューンすることでネットワークの複雑化やパラメータの増加を最小限に抑え，当時最先端の認識性能を達成している．また複雑な融合手法では，各モダリティの間でソースターゲットアテンションを計算する Co-Attention 層を使用している．Speech-BERT と RoBERTa をファインチューンして実験した場合，単純な融合メカニズムの方が認識性能が高い結果となった．しかし，この 2 つの SSL モデルを特徴抽出器としてのみに使用して 2 つの融合手法の性能を比較したアブレーション実験では，いずれも上記のファインチューニングをした性能には及ばないが，テキストと音声のモダリティ間で相互作用が増えることから，複雑な融合手法の認識性能が単純な融合手法の性能を上回る結果となっている．

また，感情表現認識に関する研究として，Delbrouck らの研究 [Delbrouck 20] がある．音声，テキスト，及びビデオの 3 つのモダリティを用いて，Transformer Encoder をベースとしたユニモーダルとマルチモーダル感情表現認識モデルの認識性能を比較・検証している．テキストと別のモダリティとの間でアテンションを計算する融合手法で，言語と音声の組み合わせと，言語と音声とビデオの組み合わせでそれぞれ認識精度を検証している．その結果，音声とテキストの組み合わせが最も高い認識精度となった．一方で，テキストの単語埋め込みベクトルの獲得には GloVe を用いているため，文脈を考慮した単語ベクトルを使った他の研究のモデルより精度が劣ってしまう課題も述べられている．

これらの研究から，特徴量の抽出方法改善について検討の余地があるといえる．また，実際の音声対話の状況では，感情表現の認識には音声から得られる音声特徴と言語特徴を得ることで感情表現の認識を行うことが考えられるため，音声対話の状況を想定した感情表現認識に対する評価が必要である．更に，複数の感情が存在する 1 つのサンプルに対して，それに含まれる複数の感情の認識と，感情が現れる句・節などを適切に識別するための研究は現状十分にされていない．そこで本研究では，特徴量抽出方法の改善や，実際の音声対話を想定した音声を単一の入力とする音声とその Whisper による書き起こしテキストを用いた感情表現認識タスクの性能を評価する．また，長い音声データは入力可能なサイズに分割し，分割した音声毎に感情表現認識をして合議を取る．

## 3. 感情表現認識モデル

### 3.1 基本モデル

感情表現認識のベースラインとするモデル 1 を図 1 に，モデル 1 を拡張した 4 つの拡張モデルを図 2(a)，図 2(b)，図 3，図 4 に示す．モデル 1 ではテキストと音声の特徴ベクトルをそれぞれ異なる Transformer エンコーダへ入力する．図 1 内の SA(Self-Attention) ブロックは，音声特徴量のセルフアテンションに基づく Tranformer エンコーダである．また CA(Cross-Attention) ブロックは，音声特徴量と言語特徴量とのソースターゲットアテンションに基づくクロスモーダル Transformer エンコーダである．テキスト特徴量としては事前学習言語モデル RoBERTa で得る単語ベクトル，音声特徴量としては音声の周波数や強弱の情報を含むメルスペクトログラムを用いる．モデル 2，3，4，5 は，モデル 1 と同様のアンサンブル手法を持ち，メルスペクトログラムの代わりに事前学習モデル wav2vec 2.0 を使用して音声特徴を取得する．さらに，

連絡先: 仁平正彦，創価大学大学院理工学研究科情報システム工学専攻，東京都八王子丹町 1-236

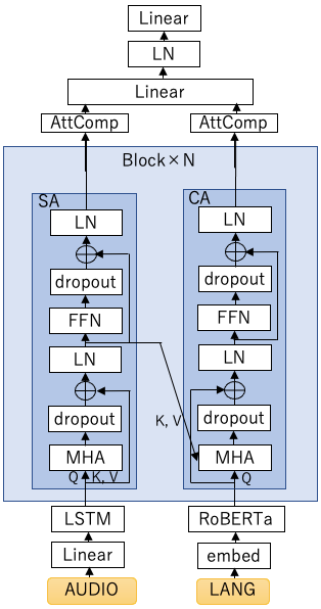


図 1: モデル 1

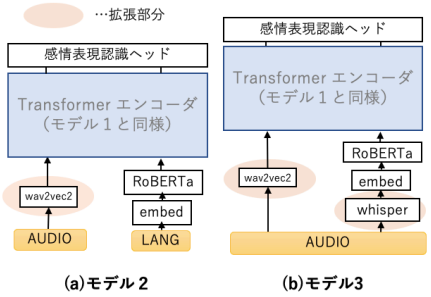


図 2: モデル 2 とモデル 3

音声認識モデル Whisper をモデル 3，4，5 に適用し，モデル 4，5 は長い音声を分割し，その分割した各音声で感情表現認識をしてそれらの合議を取るモデルである．合議の手法はモデル 4 の各感情ラベルの確率の平均値を取る手法をベースとし，モデル 5 では確率の最大値をとる手法を適用した．また，音声のセグメント毎に異なる感情ラベルを含む小規模データセットを新たに作成し，CMU-MOSEI データセットで学習した合議モデル 4 を事前学習モデルとして新たなデータセットでファインチューンしたものをモデル 6 とする．

### 3.2 Wav2vec 2.0 の組み込み

拡張モデル 2 から 5 では，音声を音声処理ライブラリ LibROSA によりサンプリングした後，wav2vec 2.0 によるエンコーディングから音声特徴量を得る．Wav2vec 2.0 は，文脈を考慮した音声特徴量を得ることができる音声認識フレームワークである．この wav2vec 2.0 の事前学習済みモデルを用いることで，メルスペクトログラムより多くの感情表現特徴を得られると考える．

### 3.3 音声認識の組み込み

拡張モデル 3，4，5 は音声認識モデル Whisper による文字起こしテキストを，データセットで用意されたテキストデータセットの代わりに用いるモデルである．音声とテキストを別々に入力する必要がなく，音声を単一の入力とするため，実際の音声対話をする状況を想定した認識精度への評価が可能となる．

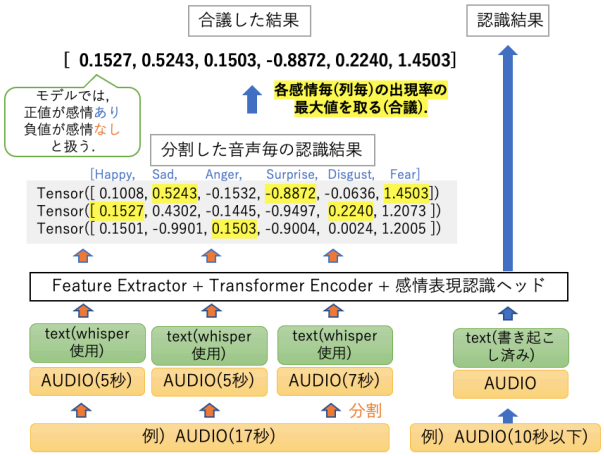


図 3: モデル 4

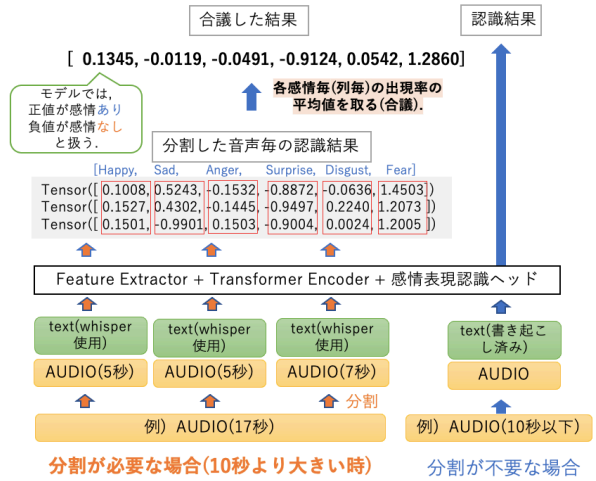


図 4: モデル 5

### 3.4 合議制の採用

拡張モデル 4，5 は，長い音声データを入力可能なサイズに分割し，分割した各音声で感情表現認識をして合議を取る．基本モデルでは，音声は感情表現認識モデルへ入力可能なサイズに前処理で切り落とすため，用意されたテキストデータセットとの整合性が保てないという問題や，複数感情を含む音声の場合，含まれる感情数が多くなるほど感情表現認識の精度が低下するといった問題がある．この問題に対処するためにも，長い音声を分割し，その音声の文字起こしテキストをモデルへ入力することで，音声に含まれる複数感情を分割して処理するアプローチをとった．入力可能な音声サイズは 10 秒までで，10 秒より長い音声は分割し，10 秒以下の音声はそのまま用いる．分割の仕方は先頭から 5 秒ずつ切取り，音声の末尾が 10 秒以下になったらそのまま使用する．分割した各音声の認識結果は，分割した音声毎に Transformer エンコーダに入力をして感情表現認識をし，それらの各感情ラベルの確率の平均値または最大値を取ることで，合議した 1 つの感情表現を取得する．

### 3.5 音声セグメント感情表現データセットでの学習

合議制の採用をした拡張モデル 4 と 5 では，分割した音声毎に感情ラベルが存在しないため，分割した音声毎に感情表現が適切に認識されたかは評価できないという問題がある．人が他人の感情表現を認識して発話を行う場合では，文に含まれる

感情表現と感情が現れる句・節などを識別し，その箇所に対応する発話生成をおこなうケースが考えられる．そのため，感情表現の認識に伴う発話生成等を行うには，文に含まれる複数の感情表現と，各感情が現れる句・節などを適切に識別する必要がある．そこで，本研究では基本モデル 1 及び拡張モデル 2 から 5 で使用する CMU-MOSEI データセットから長い音声サンプル（20 秒程度）を 100 個選択し，5 秒毎に分割した各音声にアノテーションを行った小規模データセットを作成する．そして，このデータセットを用いてモデル 4 を事前学習モデルとしてファインチューンをすることにより，分割した音声毎の認識結果の評価を可能にする．短い秒数かつ含まれるラベルが少ないサンプルの学習を通じて，各感情を識別するための特徴がより的確に学習されることが期待できる．

## 4. データセット

### 4.1 CMU-MOSEI データセット

基本モデル 1 及び拡張モデル 2 から 5 の感情表現認識精度の比較実験では，CMU-MOSEI データセット [Zadeh 18] を用いる．本データセットには，23,500 を超える文の発話ビデオが含まれており，各データに対し 2 センチメント，7 センチメント，6 感情の 3 種類のでラベル付けがされている．実験では，まずデータセットを train，valid，test の 3 つのデータセットに 7:1:2 の割合で分割した．そして，分割したデータセットそれぞれから実験用データファイルを作成した．ここでは，そのうち sentiment と emotion ファイルについて説明する．sentiment ファイルには辞書形式で { ビデオ名：実数 (-3 から 3 の範囲)} が格納されていて，実数は，正の値がポジティブ，負の値がネガティブを表す．emotion ファイルには，辞書形式で { ビデオ名：6 つの実数のリスト } が格納されている．6 つの実数はそれぞれ「Happy，Sad，Anger，Surprise，Disgust，Fear」の 6 感情クラスに対応し，0 から 3 までの値を取り，正値が感情有り，0 が感情無しを示す．本研究では sentiment データを用いた「Positive，Negative」の 2 クラス分類と，「Strong negative, Negative, Weakly negative, Fear, Weakly positive, Positive, Strong positive」の 7 クラス分類，また emotion データを用いた 6 感情クラス分類の精度を検証する．なお，モデル 3，4，5 では CMU-MOSEI データセットのテキストは用いずに，Whisper で音声を書き起こしたテキストを使用する．

### 4.2 音声セグメント感情表現小規模データセット

本研究では，分割した音声セグメント毎の感情表現認識精度の評価と，音声に含まれる複数の感情表現をセグメント毎に識別できる感情表現認識モデルの作成を目的とし，CMU-MOSEI データセットから 100 個の音声サンプルを選定・分割し，アノテーションすることで小規模データセットの作成を行った．CMU-MOSEI データセットから選定したデータ 100 個のラベル数（6 クラスと 2 クラス）を表 1 に，Whisper による書き起こしテキストと CMU-MOSEI のテキストデータセットに対する WER（Word Error Rate）と CER（Character Error Rate）の平均値及び音声の平均秒数を表 2 に，分割後の音声 340 個のアノテーションのラベル数を表 3 に示す．具体的な作成手順は，まず CMU-MOSEI データセットから 100 個の音声データを選定する．選定条件は優先度順に「15 秒から 30 秒程度の音声かつ感情を 3 つ以上含む」,「Whisper による書き起こしテキストと CMU-MOSEI データセットのテキストで WER が 30 %以下，CER は 15 %以下」,「感情ラベル数のバランスに偏りが出ないよう選定」の 3 つで，それぞれ train と valid データセットから 70 個，test データセットから 30 個選定する．選定した 100 個の音声データを全て 5 秒程度に分割し，分割後の音声 340 個を whisper で書き起こしてテキストを用意し，音声と書き起こしテキストの両方をもとに 6 感情クラスと 2 センチメントクラス，7 センチメントクラスの 3 つのクラスでアノテーションを行う．

評価者の人数は 1 つの音声につき 3 人とし，日本人 7 名と留学生 2 名の合計 9 人の同年代の大学生・大学院生の評価者でアノテーションを行った．アノテーションの方法は 6 感情クラスが 0 から 3 までの値，7 センチメントクラスが 0 から 6 までの実数値，2 センチメントクラスが 7 センチメントクラスのラベルが 3（Neutral）以上のものをポジティブ，3 より小さい値のものをネガティブとしてラベル付けする．最終的なラベルは評価者 3 人の素点の平均値により決定する．ただし，6 感情クラスでは各ラベル数の偏りや，1 つの音声サンプルに含まれる感情数が極端に多くなる可能性を考慮し，評価者 3 人の素点の平均に対し，感情の有無を判断する閾値を 0.3 に設定した．実験では，6 感情クラスと 2 センチメントクラスの 2 つのクラス分類で，音声セグメント毎の認識性能の評価を，CMU-MOSEI データセットで学習した事前学習モデル 4 をファインチューンしたモデル 6 で行う．

表 1: 選定した音声データの label 数（6 と 2 クラス）

| class | 6 感情クラス | | | | | | 2 センチメントクラス | |
|---|---|---|---|---|---|---|---|---|
| label | happy | sad | anger | surprise | disgust | fear | Positive | Negative |
| train | 46 | 55 | 49 | 34 | 43 | 40 | 40 | 40 |
| test | 10 | 17 | 12 | 7 | 11 | 10 | 10 | 10 |
| total | 56 | 72 | 61 | 41 | 54 | 50 | 50 | 50 |

表 2: 選定した音声データの WER・CER・秒数の平均値

| WER | CER | 秒数 |
|---|---|---|
| 17 % | 5 % | 19.2 秒 |

表 3: アノテーションを行った音声データの label 数

| class | 6 感情クラス | | | | | | 2 センチメントクラス | |
|---|---|---|---|---|---|---|---|---|
| label | happy | sad | anger | surprise | disgust | fear | Positive | Negative |
| train | 118 | 131 | 78 | 40 | 73 | 65 | 157 | 116 |
| test | 23 | 34 | 35 | 26 | 31 | 14 | 32 | 36 |
| total | 141 | 165 | 113 | 66 | 104 | 79 | 189 | 152 |

## 5. 実験

実験ではまず，感情表現認識タスクに対する音声の事前学習モデルの有効性を評価する．次に，用意された音声とテキストを用いた場合と，音声と音声の書き起こしテキストを用いた場合の認識精度を比較し，Whisper による書き起こしテキストの有用性を検証する．更に，合議制の採用をしたモデル 4 と採用していないモデル 3 を比較し，モデル 4 の音声データを分割して合議を取るアプローチの有用性を評価する．また，モデル 4 と異なる合議の手法を取るモデル 5 との比較により，合議手法の違いによる認識性能の比較を行う．最後に，合議を平均で取るモデル 4 による音声セグメント感情表現小規模データセットの test データを用いた，音声セグメント毎の感情表現認識性能の評価と，モデル 4 を事前学習モデルとしてファインチューンしたモデル 6 のそれら認識性能を評価し，音声に含まれる複数感情をセグメント毎に認識可能かを検証する．モデル 4・5 で用いる書き起こしテキストは分割した音声ファイルから書き起こす．テキストのトークン数が 0 または 1 と

なる場合は学習に悪影響を及ぼす恐れがあるため省いている．

学習パラメータの調整は，グリッドサーチにより計 16 通りの組み合わせから最適なパラメータを設定した. 最適化関数は adam，ロス関数は，6 感情クラス分類にロジット・バイナリクロスエントロピー，2 と 7 センチメントクラス分類ではクロスエントロピーを用いた．また，乱数シードのバラツキによる影響を考慮し，最適なパラメータ設定をした後に異なるシードを設定した状態で計 3 回分の実験を行い，その平均を各モデルの性能比較に使用する．モデル 1 と拡張モデル 2 から 5 の実験結果を表 4 に，モデル 4 とモデル 6 の，音声セグメント毎の認識性能の結果を表 5 に示す.

表 4: 基本モデルと拡張モデル 2，3，4，5 の認識精度

| モデル名 | 6 クラス | 2 クラス | 7 クラス | 平均 |
|---|---|---|---|---|
| モデル 1 | 81.670 | 85.144 | 47.510 | 71.441 |
| モデル 2 | **81.777** | **86.289** | **48.554** | **72.207** |
| モデル 3 | 81.745 | 85.470 | 48.499 | 71.905 |
| モデル 4 | 81.341 | 83.746 | 46.833 | 70.640 |
| モデル 5 | 81.443 | 83.725 | 46.753 | 70.640 |

表 5: モデル 4 と 6 のセグメント毎の認識精度

| モデル名 | 6 クラス | 2 クラス | 平均 |
|---|---|---|---|
| モデル 4 | 77.206 | 61.765 | 69.486 |
| モデル 6 | **78.431** | **77.941** | **78.186** |

### 5.1 事前学習モデルによる音声特徴量抽出の有効性

表 4 の実験結果より，メルスペクトログラムを音声特徴量として使用したモデル 1 と，wav2vec 2.0 を使用したモデル 2 の性能を比較すると全てのクラス分類において wav2vec 2.0 を使うことによる精度の向上が確認できた．これにより，感情表現認識タスクにおいて事前学習モデルで得られる音声特徴量が有効であることが確かめられた.

### 5.2 書き起こしテキストの有用性の検証

表 4 の実験結果より，データセットで用意された音声とテキストを用いるモデル 2 と，音声とその書き起こしテキストを入力とするモデル 3 の性能を比較すると，同程度の精度であることがわかる．これにより，最新の音声認識ライブラリ Whisper を適用した書き起こしテキストが十分に感情表現認識タスクに適用できることが確かめられた.

### 5.3 合議の有用性と合議手法の違いによる性能比較

表 4 の実験結果より，長い音声は入力可能なサイズに切り落として入力するモデル 3 と，長い音声をモデルに入力可能なサイズに分割し，分割した音声間で合議を取るモデル 4 の性能を比較すると，平均で約 1.26 ポイント性能が下がる結果となった．長い音声でも入力可能となる便宜的なモデルではあるが，精度の面で課題が残る結果となった．また，合議において各感情ラベルの確率の平均を取るモデル 4 と，最大値を取るモデル 5 の性能を比較すると，同程度の精度であるとわかる．モデル 4 の平均を取る合議手法は，ある感情が続く期間と強さを考慮する手法であるのに対し，モデル 5 の最大値を取る合議手法は，ある感情の強さのみを考慮するモデルといえるが，どちらの合議手法でも性能に差は現れなかった.

### 5.4 分割した音声毎の認識性能

表 5 の実験結果より，音声セグメント毎の認識性能を音声セグメント感情表現小規模データセットのテストデータで評価したモデル 4 と，モデル 4 をこのデータセットでファインチューンしたモデル 6 の性能を比較すると，6 クラスと 2 クラス分類において精度の向上が確認できた．これにより，音声セグメント毎のラベル付きデータセットを用いた学習が音声セグメントの感情表現認識性能を向上させることが確かめられた．このことから，モデル 6 を用いてセグメント毎の認識結果の合議を取ることで認識精度の向上が期待される.

## 6. むすび

本研究では，マルチモーダル Transformer エンコーダに基づく感情表現認識モデルにおいて，感情表現認識タスクに対する音声の事前学習モデルの有効性と，Whisper による書き起こしテキストの有用性を確かめることができた．また，小規模の音声セグメント感情表現データセットを用いたファインチューンにより，複数感情を含む 1 つの長い音声を分割したセグメント毎の認識性能の向上を確かめることができた．今後の課題として，まず作成した音声セグメント感情表現データセットについて，評価者による点数分布の違いや感情による点数分布の違いが存在するため，評価者毎の解答分布を正規化する等の工夫を取り入れることで，より適切なデータセットへの調整を検討する．また，複数感情を含む音声を分割して認識した結果を合議する，より有効な手法の検討を行う.

## 参考文献

[Delbrouck 20] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, Stéphane Dupont: A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis, arXiv:2006.15955, 2020

[Siriwardhana 20] Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, Suranga Nanayakkara: Jointly Fine-Tuning “BERT-like” Self Supervised Models to Improve Multimodal Speech Emotion Recognition, arXiv:2008.06682，2020

[Baevski 20] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations，arXiv:2006.11477，2020

[Radford 22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever: Robust Speech Recognition via Large-Scale Weak Supervision，arXiv:2212.04356，2022

[Liu 19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov.: RoBERTa: A Ro-bustly Optimized BERT Pretraining Approach，arXiv:1907.11692，2019

[Vaswani 17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need, arXiv:1706.03762，2017

[Zadeh 18] Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, Louis-Philippe Morency. :Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018