深層学習によるマルチメディアデータを入力とした動作ラベルの推定

1　はじめに

近年, 深層学習を始めとする機械学習技術の大きな発展を受けて, 人工知能を用いた創作物理解が注目されている. しかし, 創作は高次の知的活動であるため, いまだに実現が困難なタスクである. 人の創作物の理解に関する分野の中でも動画を対象とした研究は, 画像データと音声データに加えて, 場合によってはテキストデータをも用いるマルチメディアデータを扱う分野である. 動画を対象とした生成タスクには, 例えば白黒映像に色情報を付与する自動着色や, 映像を滑らかにするフレーム補間, 動画に対する音声の自動生成などがあり活発な研究分野であるといえる. 本研究では, 最終的な目標を深層学習によるマルチメディアデータを用いた動画に対する音声の自動アノテーションを行うモデルの構築とし, そのための前段階として, 動画に含まれるマルチメディアデータから動画の動作を推定する.

2　関連研究

動画に対する音声の自動生成に関連した研究として, Yipin らの研究 [1] について述べる. Yipin らは, 図 1 に示すように音声の生成部分に RNN を階層状に積み, 上層ほど長い間隔の依存をとらえるのを担当するような構成を構築した SampleRNN を用いて, SampleRNN の最上階層への入力として, 学習済みの VGG19 モデルを用いてフレーム画像から得た特徴量や, 入力映像の内容と動きをより捉えるために, オプティカルフローを使用することで高精度な音声の生成を可能にした.

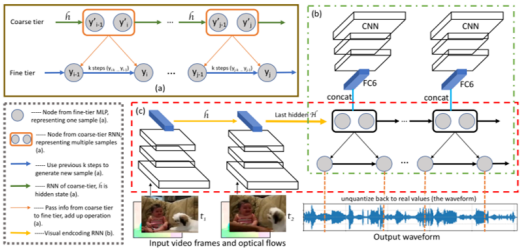

図 1: 文献 [1] における音声生成モデル

3　データセット

動画に関係したデータセットには, Kinetics や Moments in Time などがあるが, 本研究では STAIR Labs らによって作成された家庭内で見られる人間の行動 100 種類を集めた行動認識研究のための大規模な動画データセット, STAIR Actions [2] を用いた. STAIR Actions データセットに含まれる動画の半数は Youtube から取得された動画, 残りの半数はクラウドソーシングによって作成された動画から構成されており, 1 つの動画には 1 つの動作ラベルが付与されている. さらに, クラウドソーシングによって 1 つの動画につき平均 5 つの検品済み日本語キャプションがアノテートされており,「誰が」(Who),「どこで」(place),「何をしているか」(action) の情報が含まれている.

4　要素技術

4.0.1　メル尺度, メルスペクトログラム

メル尺度とは音高の知覚的尺度の一つで, 人間が高い周波数よりも低い周波数の違いを検出するのが得意である特性に合わせてメル尺度の差が同じであれば, 人間が感じる音高の差が同じになるようにスケール変換したものである. メル尺度と通常の周波数測定値との基準点として, 1000Hz の音を 1000 メル ($mels$) の音高と定義されており, 周波数 f をメル尺度 m に変換する一般式は,

$$m = m_0 \ln(\frac{f}{f_0} + 1) \tag{1}$$

で表される. ここで, f_0 は自由な周波数パラメータで, m_0 は 定義から導かれる式,

$$m_0 = \frac{1000}{\ln(\frac{1000}{f_0} + 1)} \tag{2}$$

から算出される従属パラメータである. そして, 信号を短時間フーリエ変換して得られたスペクトログラムに対して, 周波数をこのメル尺度に変換したものをメルスペクトログラムという.

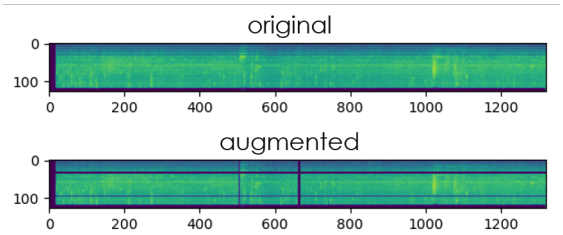

図 2: SpecAugment 例

4.0.2 SpecAugment

SpecAugment [3] とは, 2019 年に Google Brain が提案した音声データの拡張手法であり, スペクトログラムに対してランダムに選んだ点を時間軸方向に動かし, 動かした分だけ点の左右に存在する矩形領域を伸び縮みさせ, 時間方向の変調を行う Time Warping, 数チャネル分の周波数成分をマスキングする Frequency Masking, 数ステップ分の時間軸成分をマスキングする Time Masking の 3 つの変換方式から構成される. 本研究では, Frequency Masking 及び Time Masking のみを扱う. 図 2 に SpecAugment を適用した例を示す.

4.0.3 VGG

VGG [4] とは, ImageNet [5] と呼ばれる大規模画像データセットで学習済みの畳み込みニューラルネットワークであり, 構成する層の数に応じて VGG11 や VGG16 などと呼ばれている.

4.0.4 Residual Network (ResNet)

Residual Network (ResNet) [6] は Microsoft Research が 2015 年に考案した手法である. ショートカット接続機構を導入し, 手前の層の入力を後ろの層に直接足し合わせることで, ある層で求める最適な出力を学習するのではなく, 層の入力を参照した残差関数を学習し, 勾配消失問題を解決した.

4.0.5 Word2Vec, Doc2Vec

Word2Vec [7] は単語の分散表現を獲得する手法の 1 つである. この手法は, 同じ文脈で出現する単語は類似した意味を持つと予想されることに基づいており, 写像されたベクトルは, One-hot 表現のような局所表現と異なり, 単語間の意味を考慮した類似度測定や,「王様」–「男」+「女」=「女王」のような単語間の意味における演算などができるようになる. Word2Vec では, 自己から周りの単語あるいは周りの単語から自己を予測することにより分散表現を獲得する. 前者の手法を Skip-gram といい, 後者の手法を Countinuous Bag-of-Words (CBOW) という.

Doc2Vec [8] は Word2Vec をベースとした, 文書をベクトル空間上に写像して分散表現を得る自然言語処理の手法である. Paragraph ID は各文書と紐づいており, 単語の学習時に一緒にこの Paragraph ID を学習することで文書の分散表現を獲得する. このベクトルを用いると文書間の類似度の算出や文書間での加減算が可能になる. CBOW を拡張したモデルを Distributed Memory モデルといい, Skip-gram を拡張したモデルを Distributed Bag-of-Words という.

5 数値実験

本研究では, STAIR Actions データセットの内, 動作ラベルが “drinking”, “eating meal”, “washing face”, “gardening”, “fighting” の全 5 種類のみを用いて, 入力を動画に含まれる音声データまたはキャプションデータ, 出力を動作ラベルとする 5 クラス分類問題を学習した.

実験 1 および実験 2 では, 使用するデータからランダムで 50%, 30% サンプリングしたものをそれぞれ訓練用データおよび検証用データとし, 残り 20% を評価用データとした. そして, 損失関数に使うクラス重みとして訓練用データの各動作ラベルのデータ数の逆数を正規化したものを用いた. また, 検証用データにおける Acc が最大となるエポックのものを評価用モデルとして採用し, モデルの性能を測った. このとき, ベースラインとしてはすべてをランダムに予測した場合とした.

5.1 実験 1: 音声を入力とする動作ラベル推定

音声データの前処理として, 学習の際には入力される時間長を合わせる必要があるため, 本研究では開始 0 秒から 5 秒間のデータを切り出し, 長さが不足している場合は音声データを繰り返すことでパディングした. 次に, 音声データをメルスペクトログラムに変換し, 最後に次元数を $(2\times128\times128)$ にリサイズしたものをネットワークへの入力とした. ネットワークには VGG16 に Batch Normalization を導入した VGG16-BN と, ResNet34 の事前学習済みモデルを用いて最後の全結合層のみを fine tuning し

表 1: 音声データ数

label	train	validation	test	avg length [sec]
drinking	254	149	107	5.64 ± 0.60
eating meal	231	122	88	5.82 ± 0.78
washing face	490	308	177	5.59 ± 0.68
gardening	64	43	35	5.59 ± 0.38
fighting	130	81	61	5.64 ± 0.32
合計	1169 (50%)	703 (30%)	468 (20%)	

表 2: キャプションデータ数

label	train	validation	test
drinking	646	390	268
eating meal	602	345	256
washing face	252	143	82
gardening	1005	623	416
fighting	1042	628	398
合計	3547 (50%)	2129 (30%)	1420 (20%)

た. このとき, 音声データの入力次元とクラス数に合わせてアーキテクチャを一部変更している. そして, 訓練用データに対して入力の度に SpecAugment を行う場合と行わない場合の 2 つの設定について学習し, 評価用データを用いて精度の結果を比較した.

表 1 に実験 1 で用いたデータ数および動作ラベルごとの平均時間長, 表 5 に学習時のパラメータ, 表 6 に SpecAugment のパラメータを示す. 表 1 において, 実験 1 ではクラウドソーシングによって作成された動画データとその音声データのみを用いた.

表 3 に実験 1 の結果を示す. 表 3 より, どの場合もベースラインを超えることができ, ネットワークに VGG16-BN を用いて SpecAugment をした時, 最も評価指標の値が高かったが, SpecAugment の有意性を示すことはできなかった. 図 3 にネットワークに ResNet34 を用いた場合の学習曲線を示す. 図 3 より, SpecAugment をしたことで訓練用データの Acc が検証用データの Acc を下回るように学習が進んでいる一方で, 検証用データにおける loss の増加が抑えられており, ネットワークに VGG16-BN を用いた場合でも同じ傾向が見られた. この事より, SpecAugment によって過学習が抑えられ汎化性能が向上していると推測できるが, 本実験で用いたハイパーパラメータは恣意的なものであり, 妥当性を考慮していない. よって optuna などを用いたハイパーパラメータのチューニングによって, 最大性能同士の比較が必要であると考えられる. 図 4 に SpecAugment を行わなかった場合の VGG16-BN における評価用データに対する混合行列を示す. 図 4 より, 5 つの動作ラベルの内, 精度が低かったのは “drinking”, “eating meal” であり, どの実験設定でも同じ傾向が見られた. この原因としては, これら 2 つの動作ラベルが飲食に関わるものであることや, 動作そのものに音があまり伴わないデータが多かった事が挙げられる. その他にも動作ラベルに対応する前景よりも別動作に対応する背景の音声が強調されたものや無音であるデータも含まれていたため, データの選び方についても考慮すべき問題である.

表 3: 実験 1 結果

	評価指標	
	Acc	F 値 (macro)
VGG16-BN	0.771	0.752
VGG16-BN (SpecAugment)	0.773	0.754
ResNet34	0.756	0.740
ResNet34 (SpecAugment)	0.741	0.724
baseline	0.200	-

表 4: 実験 2 結果

	評価指標	
	Acc	F 値 (macro)
	0.761	0.769
baseline	0.200	-

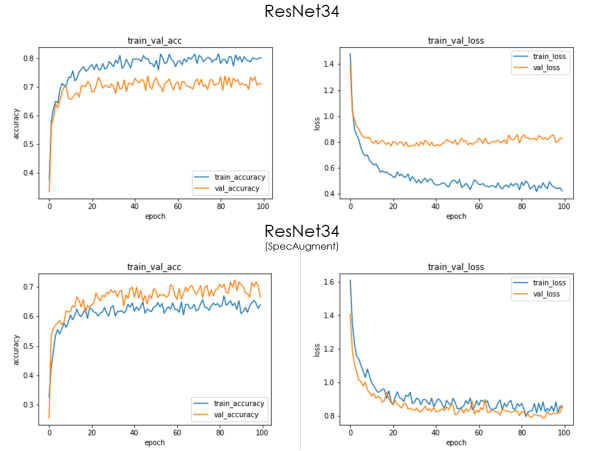

図 3: ResNet34 学習曲線 (実験 1)

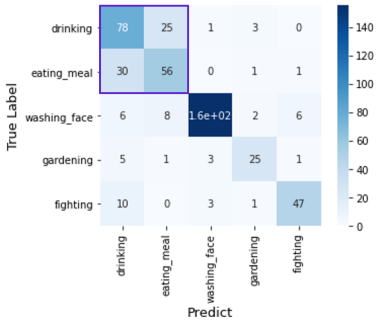

図 4: VGG16-BN 混合行列 (実験 1)

5.2 実験 2: キャプションを入力とする動作ラベル推定

アノテートされたキャプションの内,「何をしているか」(action) を表すキャプションのテキストデータを JUMAN++ [9] によって分かち書きにしたものを, 事前学習済み Doc2Vec モデルを用いて 300 次元のベクトルに変換し, これを識別器である 3 層 MLP

表 5: 学習パラメータ

parameter	value
epoch	100
batch size	64
learning rate	0.001
optimizer	Adam
loss function	Cross Entropy Loss

表 6: SpecAugment パラメータ

parameter	value
time drop width	[0, 4]
time stripes num	[0, 1]
freq drop width	[0, 4]
freq stripes num	[0, 1]

表 7: 3 層 MLP パラメータ

parameter	value
input size	300
hidden size	30
output size	5
activation function	tanh
dropout rate	0.2

表 8: Doc2Vec パラメータ

parameter	value
corpus	日本語 Wikipedia 小説家になろう 青空文庫
vector size	300
epoch	30
window size	8
min count	5
model type	Distributed Memory

への入力として学習した.

表 2 に実験 2 で用いたデータ数, 表 5 に学習時のパラメータ, 表 8 に 事前学習済み Doc2Vec モデルのパラメータ, 表 7 に 3 層 MLP のパラメータを示す. 表 2 において, 実験 2 では作成されたすべてのキャプションデータの内, 同じ動作ラベル内で同じ文章が 2 つ以上含まないように整形したものを用いた. 表 2 より, Acc はベースラインを超えることができた. 図 5 に学習曲線を示す. 図 5 より, 学習曲線の推移に差が見られない事から, データの分布が大きく似ていると推測でき, これはキャプションデータが検品済みであるノイズの少ないデータであることに起因すると考えられる.

6 まとめと今後の課題

本研究では, 最終的な目標を深層学習によるマルチメディアデータを用いた動画に対する音声の自動アノテーションを行うモデルの構築とした上で, 基礎的な実験として動画に含まれるマルチメディアデータから動画の動作を推定した. 実験結果より, データの選び方や前処理, 音声のデータ拡張方法に関して更なる模索が必要であることが分かった. 今後の課題としては以下が挙げられる.

- 動画に含まれる画像データを入力とした動作ラベル推定
- マルチモーダルな動作ラベル推定
- マルチメディアデータを用いた動画に対する音声の自動アノテーションモデルの構築

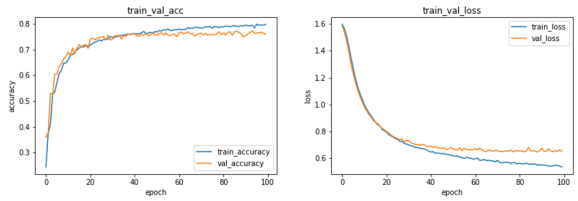

図 5: 学習曲線 (実験 2)

参考文献

[1] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. Visual to sound: Generating natural sound for videos in the wild. *CoRR*, abs/1712.01393, 2017.

[2] Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi. Stair actions: A video dataset of everyday home actions, 2018.

[3] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*, Sep 2019.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[7] Greg Corrado Tomas Mikolov, Kai Chen and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[8] Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014.

[9] 京都大学大学院情報学研究科 黒橋・河原研究室. 日本語形態素解析システム juman++ version 1.0. 2016.