# 人工知能にも笑いを作り出せるか？

## ―ボケるための正確な画像理解―

1852130019 206 木寺隼也

### 1.目的

近年、人工知能は AI 技術の発達と共にあらゆる分野で活躍を見せてきた。特に画像処理や音声分析等統計的分析の分野で社会に貢献してきた。一方でお笑いや美術、音楽等の人が感覚的に表現している分野ではまだまだ人が絵を描くように絵は描けず、人が音楽を聴いて楽しむように音楽を聴いて楽しむことはできない。そこで今回の研究では人工知能がまだ踏み込めてない”お笑い”の分野で活躍するために人工知能にも画像から人を笑わせられるボケを作れるかを研究した。正しい画像理解をして、その後正しい画像理解からわざと少し外した文章構成をするという 2 ステップが画像からボケるには必要と考え、今回の研究では 1 ステップ目の正しい画像理解を行えるようなプログラムを研究した。

### 2.原理

2.1 画像に含まれる物体認識(YOLO)

まず、YOLO の大まかな流れを知るために YOLO に使われている信頼度スコアについて説明する。

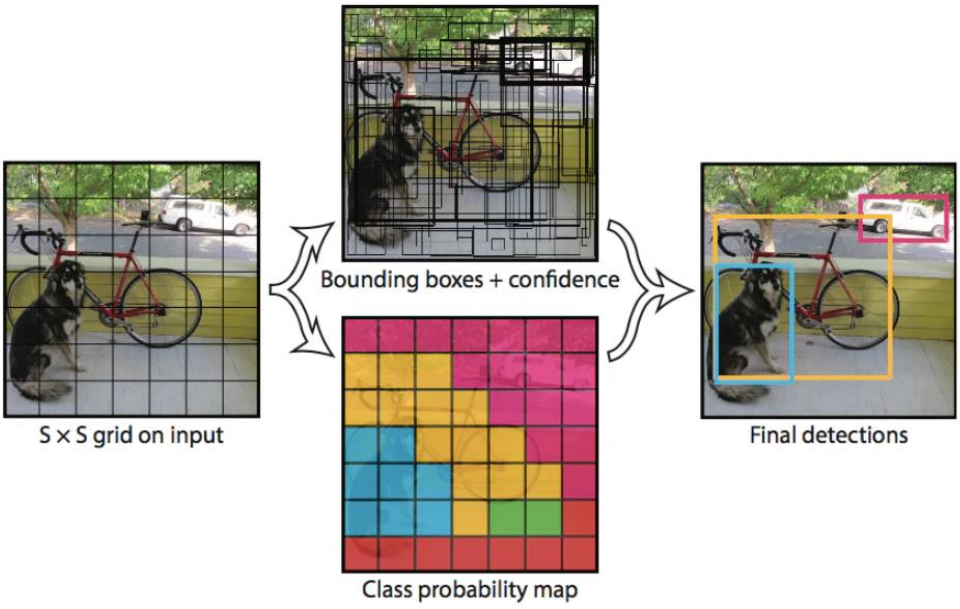


図 1 信頼度スコアの算出方法（1）

図 1 に信頼度スコア算出の流れを表した図を示す。まず初めに入力画像を S×S のグリッドセルと呼ばれる領域に分割し、それぞれのグリッドセルについて B 個の Bounding box と Bounding box ごとの Confidence を推測する。それと同時にそれぞれのグリッドセルは C 個の物体クラスそれぞれの条件付確率を表す。これは各クラスの予測確率を表しているのですべて足し合わせると 1 になる。その後、各 Bounding box の Confidence 値と各クラスの予測確率を掛け合わせて各 Bounding box、各クラスの信頼度スコアを得ることができる。信頼度スコアは各 Bounding box の Confidence(Bounding box 内に物体が入っていて正確に囲えているかの正確さ)と各クラスの予測確率を表している。つまり、この信頼度スコアを用いるとどの Bounding box が対象とするクラスの物体を検出しているかがわかる。

次に学習時、推論時の流れを紹介する。

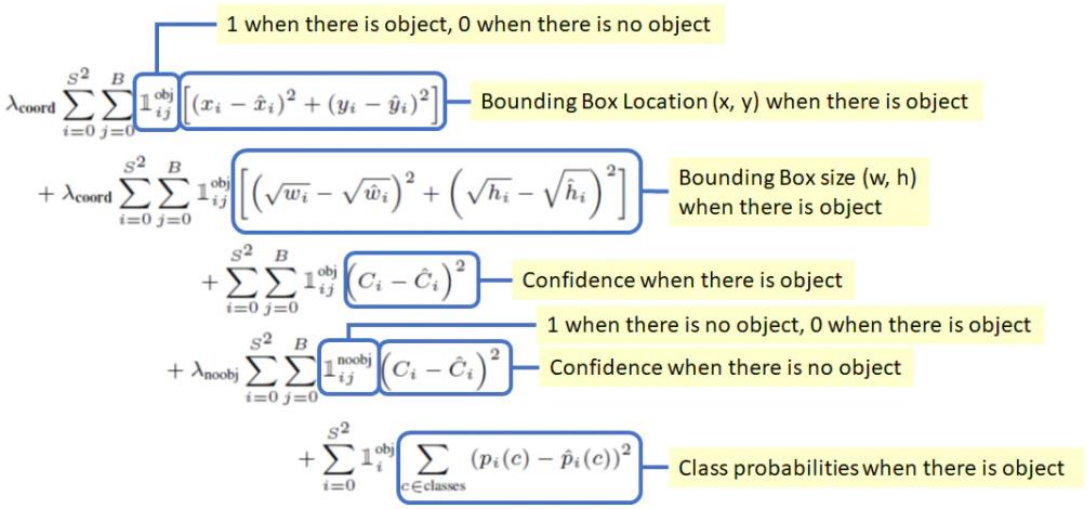


図 2 YOLO の学習時の流れ（2）

図 2 に YOLO の学習の流れを示す。学習時には信頼度スコアを算出するうえで必要になっているクラス予測確率、Bounding box の位置情報、Bounding box の Confidence の 3 要素を図 2 の式のように Loss の項として学習している。その後、推論時に Loss で学習されたモデルに基づいて 3 要素を出力して信頼度スコアを算出することで推論している。これにより検出と識別を同時に行っていることが YOLO の特徴である。

## 2.2 転移学習

転移学習とはある領域で学習したこと(学習済みモデル)を別の領域で活用し、効率的に機械学習を行う手法である。転移学習ではまず、すでに学習済みのモデルの最終層以外の部分を利用して、データ内の特徴量を抽出する。この時すでに学習済みのモデルは特徴抽出機としてのみ使用される。次に抽出した特徴量を用いて新しく分類したいク

ラスを新しく追加した層で学習する。

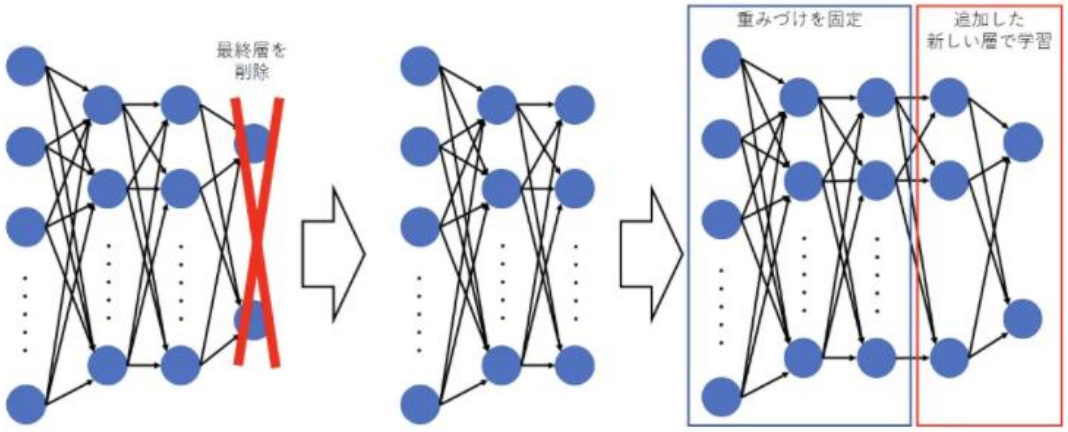


図 3 転移学習の流れ（3）

図 3 にクラス分類のために学習済みモデルの最終層を削除して分類したいクラスを最後の層に追加して学習モデルを作成している流れを示す。これにより、転移学習では短い時間で少ないデータ量で機械学習を行うことができる。

## 3.実験方法

初めに YOLO で入力画像に含まれる物体を検出し、その物体が何であるかと物体の座標を表示した後に人が検出されたら人だけを切り抜いて別の画像として保存した。次に Real-world Affective Faces Database (4)から入手した 7267 枚の表情のデータベースを転移学習の最終層に追加し、機械学習を行った。転移学習のもととなるモデルには VGG16 を使用した。この 7267 枚の表情のデータベースは怒り、嫌悪、恐怖、幸福、平常、悲しみ、驚きの 7 つの感情の時の表情の画像データであり、それぞれの感情ごとに画像が分けられている学習用データである。これと同じように感情ごとに分けられている 2978 枚の画像データを学習の評価データとして用いて、200 回学習を行った。学習の評価データは 1 回の学習ごとに評価して表示した。そして学習を行ったモデルを評価するために 5096 枚の感情ごとに分けられた画像データを使ってモデルの正答率を確かめ、さらにこの 5096 枚の画像の中から感情ごとにランダムに 25 枚の画像を選び、予測することで感情ごとの混同行列を作成することで学習モデルを評価した。最後に、物体検出で切り抜かれた人の画像に対して感情の予測を行った。

## 4.実験結果

図 4 YOLO の入力画像

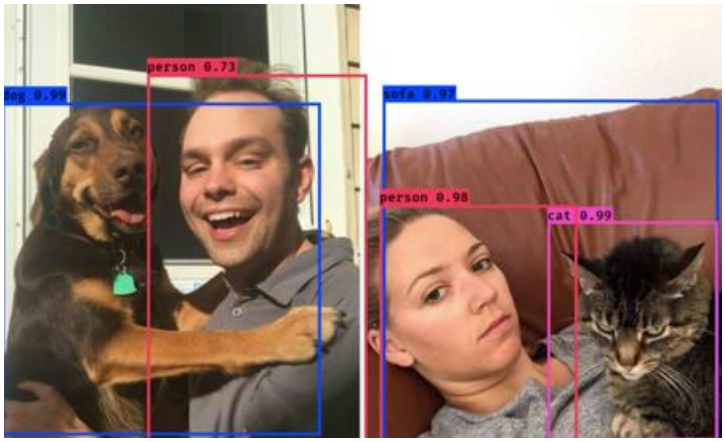


図 5 YOLO の出力画像

図 6 YOLO の結果から人のみを切り取った画像

図 4 に YOLO に入れた入力画像、図 5 に YOLO が返した出力画像、図 6 に YOLO が返した物体の画像の座標から人を切り取った画像を示す。このように入力した画像から物体を検知し、さらに人の画像を切り取って保存することができた。

```
Epoch 200/200
228/228 [==============================] - 97s 424ms/step - loss: 0.2076 - accuracy: 0.9380 - val_loss: 0.8655 - val_accuracy: 0.7492
```

図 7 転移学習により 200 回学習したときの学習精度

```
Found 5096 images belonging to 7 classes.
/usr/local/lib/python3.6/dist-packages/te
  warnings.warn('`Model.evaluate_generato

 test loss: 1.1387447118759155

 test_acc: 0.6807299852371216
```

図 8 学習したモデルを評価用の画像データで評価したときの精度

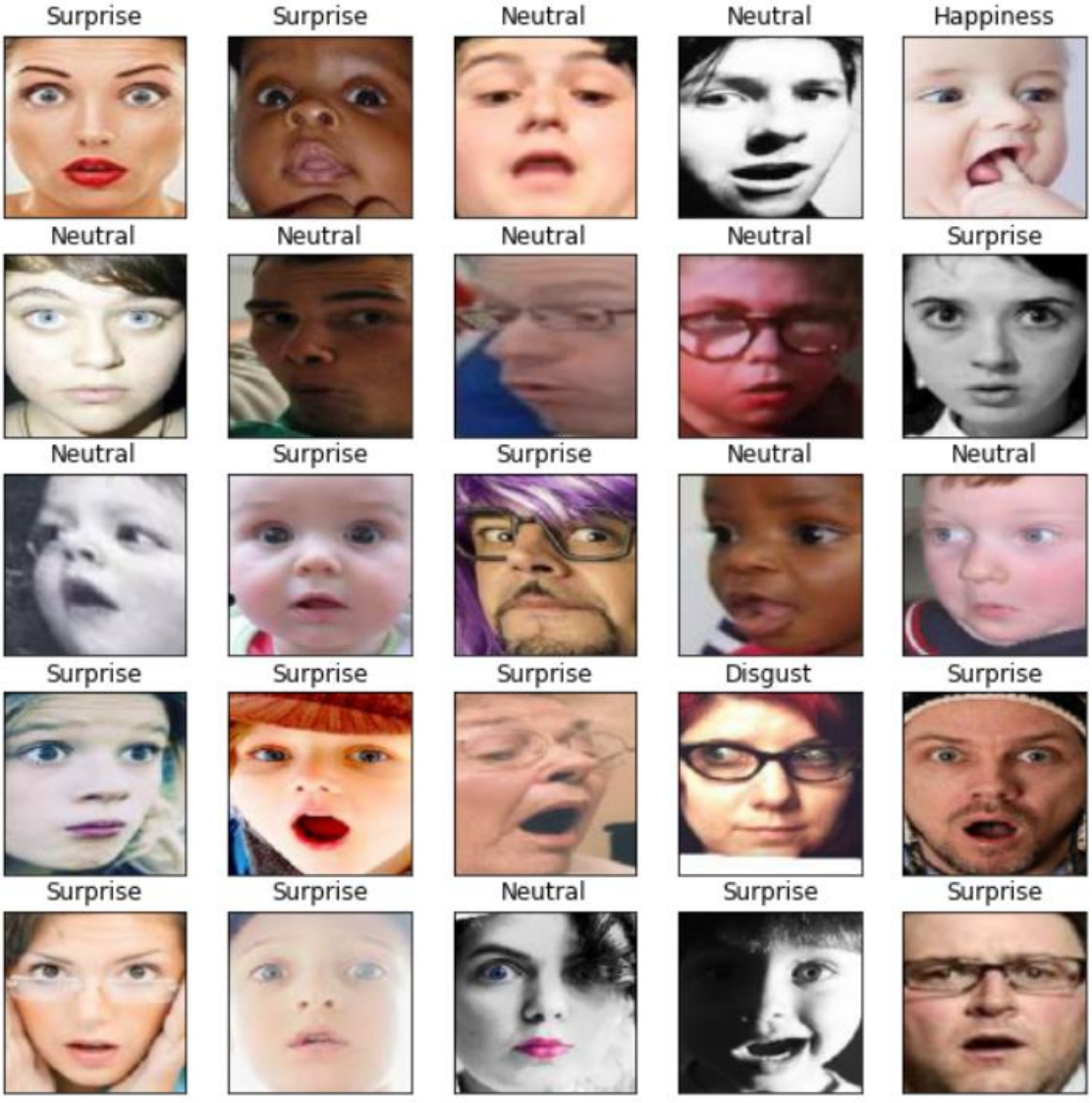


図 9 驚きの感情の 25 枚の画像データで感情予測した結果の出力

| | Anger | Disgust | Fear | Happiness | Neutral | Sadness | Surprise |
|---|---|---|---|---|---|---|---|
| Anger | 10 | 1 | 1 | 4 | 4 | 2 | 3 |
| Disgust | 1 | 8 | 0 | 1 | 11 | 4 | 0 |
| Fear | 9 | 0 | 3 | 0 | 1 | 3 | 9 |
| Happiness | 0 | 1 | 0 | 24 | 0 | 0 | 0 |
| Neutral | 0 | 2 | 0 | 2 | 20 | 1 | 0 |
| Sadness | 0 | 0 | 0 | 2 | 9 | 14 | 0 |
| Surprise | 0 | 1 | 0 | 1 | 10 | 0 | 13 |

図 10 作成した混同行列

図 7 では転移学習により 200 回学習したときの学習データに対する損失関数の値と精度、さらに学習の評価用データに対する損失関数の値と精度を示している。図 8 では学習した学習モデルを評価するためのデータに対する損失関数の値と精度を示している。図 9 では実際に驚きの感情だけをランダムに学習モデルを評価するためのデータから選び、感情予測を行った結果を表示している。図 10 では図 9 のような感情に対する評価を 7 回行い感情ごとの予測結果をまとめたものである。

## 5.考察

### 5.1 物体検出から人の切り抜きについての考察

・今回用いた物体検出では斜めの物体や小さい物体の検出があまりできなかった。YOLO のデメリットとしてあげられるこれらの問題には、M2Det などの最先端の物体検出アルゴリズムを用いることで解決できると考えた。今回用いた YOLO には扱いやすい点や扱いやすいものの中で検知と識別の両方を瞬時にできるというメリットがあった。

・人の切り抜きを行う時に YOLO で人と判別された座標のを全て切り抜いていたが、体全体を切り取ったときに、顔の部分が小さくなり次に続く表情による感情識別の精度を落とす結果になっていたことから、切り抜いた人の画像からさらに顔がどこにあるかを判別するプログラムを別に作り、そこでさらに顔の切り抜きをすることでうまく顔の拡大画像が作れると考えた。顔の切り抜きにはカスケードファイルを用いた顔認識で行うのが良いと思った。

### 5.2 転移学習による感情識別についての考察

・表情による感情の識別は感情によって精度が大きく異なるものとなったことから、人の感情を表情のみから識別するのは難しいことが分かった。人間ですら画像だけから人の

感情を察するのは容易でないので、解決策として周りの物体の状況や位置関係からある程度の人の感情予測をしておいて、そこからさらに表情による感情識別を組み合わせて行うことでより精度の高い感情予測ができるようになると考えた。
・今回の学習では 200 回学習を行ったが、140 回あたりから学習の評価用データの精度はあまり変わらず、訓練データに対する精度ばかりが大きくなっていっていたので過学習が起こり、汎化性能が落ちていた可能性が考えられたので学習用の評価データが 10 回変化しなかったら学習停止などのコードを書き加えることで過学習を阻止すべきだと分かった。
・今回の研究では時間、パソコンのスペックやデータ量から転移学習を行ったが、最終的にはデータ量は海外の大きなデータベースを使うことで大量のデータを学習することができたので、転移学習を行わず、1 から学習を行うことで精度が上がった可能性がある。転移学習のデメリットとして転移元と転移先の関連が低いと負の転移が起こり、精度が悪化することもあるので、転移学習をした場合としなかった場合での比較をして精度の良かった方を採用する方法がいいと考えた。

### 5.3 人工知能が画像でボケることについての考察

・今回の研究ではボケを作り出すまで触れることができなかったが、今回の研究のように画像から正確に情報を読み取ることができるようになれば、あえて正しく読み取った単語と別の単語から文章を作成すればボケになると考えられる。そのために単語同士を意味ごとや発音ごとに意味分けしたグループをあらかじめ作成しておき、文章構成をする際に読み取った情報の属するグループから選んだ単語と全く関係ないグループから選んだ単語を組み合わせて構成することで人工知能にも画像からボケを生み出すことは可能だと考える。単語のグループ構成する際に単語との関連度を数字で表し、関連度の高いものと低いものから文章構成する方法も考えられる。また、単語のグループ化をする際に実際のお笑い芸人のボケ方などから単語のグループを作っていくのが効率のいい方法になると考えられる。

## 6.まとめ

今回の研究により正確な画像認識を行う人工知能の基礎を作成することができた。考察での反省を活かせば正確な画像認識を行う人工知能は作れると想定されることから、人工知能が画像でボケるための一歩を踏み出せたと考えている。さらに、ボケるまでの過程の 2 ステップ目を行うことにより画像でボケる人工知能ができて、人工知能が発展していなかった'お笑い'の分野でも発展していくと考えられる。そしてお笑いだけにとどまらず人の感覚的な'芸術'の分野でも人工知能が活躍していくことにつながっていくと考えている。

## 7.参考文献

(1),(2)…【物体検出手法の歴史 : YOLO の紹介】 - Qiita

(3)…転移学習とは？ディープラーニングで期待の「転移学…｜Udemy メディア (benesse.co.jp)

(4)…Real-world Affective Faces (RAF) Database (whdeng.cn)