人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第20回)
SIG-AM-20-02

深層学習における分類パターンの解釈支援

Interpretation Support System for Classification Patterns in Deep learning

安藤 雅行[1*]　河原 吉伸[2,3]　砂山 渡[4]　畑中 裕司[4]
Masayuki ANDO[1]　Yoshinobu KAWAHARA[2,3]　Wataru SUNAYAMA[4]　Yuji HATANAKA[4]

[1] 滋賀県立大学大学院工学研究科
[1] Graduate School of Engineering, The University of Shiga Prefecture
[2] 大阪大学産業科学研究所
[2] The Institute of Scientific and Industrial Research
[3] 理化学研究所革新知能統合研究センター
[3] RIKEN Center for Advanced Intelligence Project
[4] 滋賀県立大学工学部
[4] School of Engineering, The University of Shiga Prefecture

Abstract: This paper describes an interpretation support system for classification patterns based on the contents of learning results in deep learning with texts, and verified its effectiveness. It is well known that classification patterns by deep learning models are often difficult to interpret the reasons derived. The proposed system extracts the contents of learning results in deep learning with texts and provides seeds for interpretations of the patterns learned. Then, the system displays learned network structures so that anyone can easily understand learning results. In verification experiments to confirm the effectiveness of the system, based on the learning result of deep learning classifying sentences, test subjects were instructed to give meanings of classification patterns peculiar to each output. The results show that test subjects who represent novice data scientists could understand the meanings of the classification patterns of deep learning with texts.

1　はじめに

近年，深層学習を用いたAIシステムが世の中を席巻している．画像認識を始めとし，車の自動運転や株の自動取引，医師を助ける診断アシストなどさまざまな分野において，活用が進められている．

人間が何らかの判断を行う際には根拠を問い，その判断が妥当かどうかを客観的に鑑みることが大切である．しかし現状のAIは，判断は行ってもその根拠は示さないという大きな問題点がある．このAIによる判断の根拠を，人間が解釈可能な形で取り出し、「根拠」を条件部,「判断」を帰結部とする因果関係を知識として得ることができれば，人間がその知識を共有することで，関連する仕事の効率を上げるために役立てられる．そこで本研究では，テキストベースの文章分類を題材として，深層学習により学習された分類パターンを表すネットワークの解釈を支援するシステムを提案する．

本研究の狙いは，ブラックボックスと言われている深層学習による学習結果の解釈支援にある．そのため，文章分類の精度向上を目指すアプローチではなく，人間による学習結果の解釈が可能なモデルを検討する．今回は人間の解釈は言葉によって行われるとし，単語をそのまま利用するBoW(Bag of Words)形式の入力を扱うことで，学習結果への意味づけを目指す．また深層学習の構造としても，最も単純なDNN(Deep Neural Network)による構造を扱い，本構造における解釈を可能にすることを目指す．本研究による知見が，単語の分散表現を入力とする深層学習や，RNN(Recurrent Neural Network)など他の構造の深層学習の学習結果の解釈を支援するシステムの，ベースラインとして役立てられることを期待する．

以下本論文では，2章で関連研究について述べる．3章で深層学習による分類パターンの解釈支援システムの構成と詳細について述べる．4章で提案システムの評価実験について述べ，5章で本論文を締めくくる．

*連絡先：滋賀県立大学大学院工学研究科 電子システム工学専攻 安藤雅行
〒522-8533 滋賀県彦根市八坂町2500
E-mail: oh23mandou@ec.usp.ac.jp

2 関連研究

インターネットの普及に伴い，また，SNS（Social Networking Service）の出現により，世の中の文章データが大規模になり，自分が欲しい文章や要らない文章を見分けるのが難しくなってきている．そこで注目されているのが，深層学習を用いたテキストマイニングシステムである [ダヌシカ 14, Arisoy 12]．深層学習は，一般に多層から構成されるニューラルネットワークを用いた学習を指す．畳み込みニューラルネットワーク [LeCun98, Krizhevsky12] の出現により画像を用いた場合に限らず多くの場面で高い分類性能を実現できることが報告されている．

その一方で，深層学習は，その出力を導いた根拠についての解釈が困難であることも知られている．畳み込みニューラルネットワークを用いた画像認識においては，この問題に対する研究も最近進められており，いくつかの方法が提案されている．例えば，入力画像に対応する畳み込みニューラルネットにおける層間のスコアの勾配を計算することでネットワークの可視化を行う方法や [Zeiler 14, Mahendran 15]，学習済みのネットワーク中間層のノード情報を用いて，対応する画像中の画素への寄与度を計算することにより画像の分類に重要な部位を表示する方法などが提案されている [西銘 15]．また最近は，解釈性の高いリカレントニューラルネットワーク [Wu17] や Generative Adversarial Network[Chen 16] など，他のニューラルネットワークモデルにおいても議論が進んでいる．

しかし自然言語への深層学習の適用においては，上記のような画像認識における方法を直接適用できないため，その分類根拠を抽出する効果的な方法については議論が進んでいないのが現状である．そこで本研究ではこのような問題意識の下，文章（テキスト）の分類問題を例として，深層ニューラルネットを用いた学習によってネットワークの各層に付けられた重みの値を抽出し，各中間層が学習した情報を単語の集合として表現する．そして，単語の組み合わせによって中間層ごとの学習された情報を解釈し，そこから分類のパターン，つまり出力を導くルールの意味付けをする手助けを行うシステムの開発を目指す．

3 テキストベースの深層学習における分類パターンの解釈支援システム

本章では，テキストを分類するための深層学習ネットワークにおいて，その分類パターンの解釈を支援するためのシステムについて述べる．

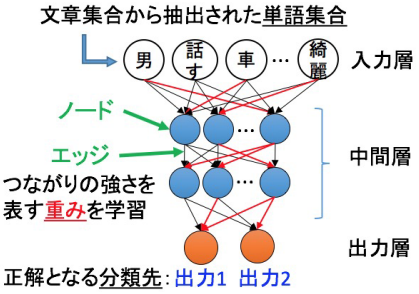

図 1: テキスト分類ネットワーク

3.1 深層学習によるテキスト分類

深層学習とはニューラルネットワークの中間層を 2 層以上に多段化したものを指す．本研究では，最も単純な DNN(Deep Neural Network) による深層学習を考え，テキスト集合を分類する問題に適用する場面を考える（図 1）．すなわち，正解ラベルが与えられたテキスト集合をもとに，各テキストの単語ベクトル（テキスト中の名詞，動詞，形容詞を対象としてその出現の有無を 0,1 で表す）を入力として，中間層が 3 層の全結合型の DNN におけるエッジの重みをその分類精度が高くなるように学習させる．学習された分類パターンを解釈するために，適切な学習（学習精度が 90%以上）が行われていることを前提としている。

本研究においては，ここで学習されたテキストの分類パターンを表す学習ネットワークをもとに，その分類パターンを人間が解釈することの支援を行う．

3.2 テキスト分類パターン解釈支援システムの構成

提案するテキスト分類パターン解釈支援システムの構成を図 2 に示す．まず，深層学習により学習された分類パターンを表すネットワークを入力とする．次に，入力されたネットワークの中から，特定の出力に結びつく部分ネットワークをパスとして抽出し，パス上のノードにテキスト中の単語を用いたラベルづけを行う．その後，これらのパスとラベルを可視化した分類パターンの解釈支援機能を，ユーザに提供する．ユーザは，可視化される重要パスとパス上のラベル，ラベルの単語が出現する原文，などを参照しながら，分類パターンに解釈を与える．

また本研究では,「ある単語を含むと，ある出力に分類される」という分類パターンを解釈に利用することを仮定する．これは，テキスト中で用いられている単語情報と，分類先となる出力との関係を表す最もシンプルな表現として設定する．

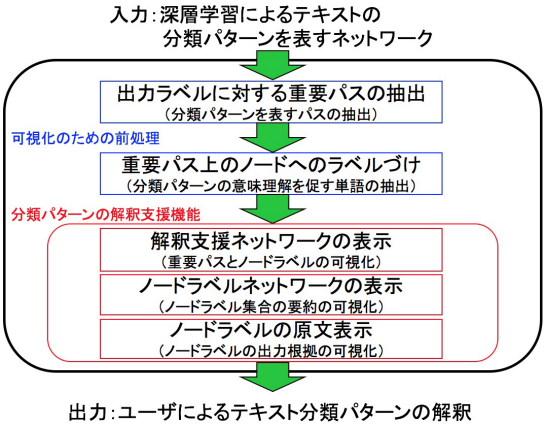

図 2: テキスト分類パターン解釈支援システムの構成

3.3 出力ラベルに対する重要パスの抽出

学習ネットワークにおいて，ある出力を導く部分ネットワークの中で，分類に大きく寄与するパスを抽出する．ここでパスとは，入力層のノードから指定の出力ノードまで各中間層のノードを 1 つずつ辿った一本道のネットワークを表す．入力層のノード数を n_0，第一中間層，第二中間層，第三中間層のノード数をそれぞれ n_1, n_2, n_3 としたとき，ある出力層へのパスの総数は，$n_0 \times n_1 \times n_2 \times n_3$ 本となる．

ノード i からノード j へのエッジ $e(i,j)$ の重みを $W_{i,j}$ で表すとする．このとき，ノード i からノード j を経由してノード k に至るパス $path = \{e(i,j), e(j,k)\}$ の重要度を，パスに含まれるエッジの重みの積として，$W_{i,j} \times W_{j,k}$ で定義する．

すなわち，パス $path$ の重要度 $\mathrm{Imp}(path)$ を，式 (1) で定義する．

$$\mathrm{Imp}(path) = \prod_{e(i,j) \in path} W_{i,j} \qquad (1)$$

エッジの重みには負の値を持つものもあるが，重要度は重みの積として計算されるため，最終的に重要度の正の値の大きい順に，重要なパスとして抽出する．

3.4 重要パス上のノードへのラベルづけ

本節では，前節の方法により抽出された重要パス上の中間層ノードの意味を表すラベルを付与する方法について述べる．

重要パスが含む中間層のノードについて，各ノードと強いつながりのある入力層のノードが表す単語を，つながりの強い順に抽出し，抽出された単語集合を，各中間層ノードのラベルとする．

各ノードと強いつながりのある入力層のノードの抽出には，前節で述べた重要パス抽出の方法を転用する．

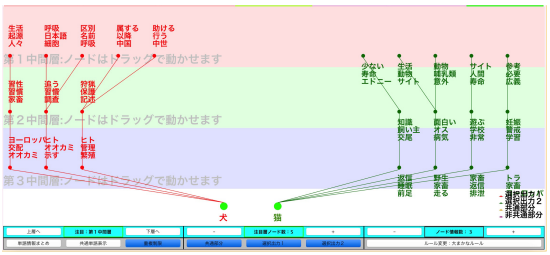

図 3: 解釈支援ネットワークの表示例

すなわち，中間層ノード m における入力層ノード w の重要度 $\mathrm{Label}(w,m)$ を，w から m に至るすべての部分パス $kpath$ の重要度の和として，式 (2) で定義する．ただし式中の $\mathrm{Allkpath}(w,m)$ は，入力層ノード w から中間層ノード m に至るすべての部分パスの集合を表す．

$$\mathrm{Label}(w,m) = \sum_{kpath \in \mathrm{Allkpath}(w,m)} \mathrm{Imp}(kpath) \qquad (2)$$

この重要度の値の高い順に，指定された数だけ入力層ノードの単語を取り出して，中間層ノードのラベルとする．

3.5 解釈支援ネットワークの表示

前節までに抽出した，ある出力を導く重要パスと，パス上のノードのラベルを可視化した解釈支援ネットワークを表示する．図 3 に解釈支援ネットワークの表示例を示す．3 つの中間層と出力層の 4 つの領域ごとに色分けを行い，最下部を出力層としてネットワークを表示する．

ユーザは初めに，インタフェースの下部に並べられた出力ラベルから，1 つまたは 2 つの出力ラベルを選択する．システムは選択された出力ラベルに対して，指定数の重要パスを表示する（図 3 の例では 5 本）．またパス上のノードに対して，ノードラベルを指定した数を最大として表示する（図 3 の例では 3 個）．ノードラベルは，一つのパス上で同じ単語が重複して現れる場合は，出力層に近いノードにのみ表示させる．ユーザは，出力されたネットワークを眺めることで，どのような単語の利用が出力を導いているかのパターンを見つけ，その解釈を行う．

3.6 ノードラベルの原文表示

分類パターンの解釈に向けて，単語情報だけでは，その単語が実際にどのような文脈で使われていたのかを把握することは難しい．そのため，ノードラベルとし

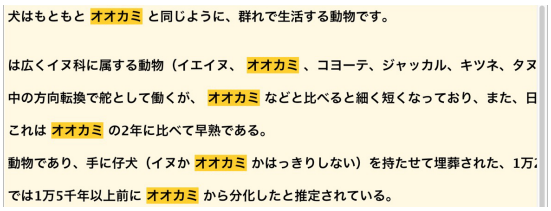

図 4: ノードラベル（オオカミ）の原文表示例

て出力された単語が，学習に用いたテキスト内で実際にどのように使われているかを表示させる．

図 4 にノードラベルの原文表示例を示す．ユーザは，解釈支援ネットワーク上で単語をクリックすることで，その単語を含む原文を表示させて参照することができる．

3.7 分類パターンの比較支援機能

分類パターンの解釈において，特定の出力を導く分類パターンの解釈以外に，複数の出力の分類パターンを比較したい場合も想定される．そのため本システムにおいては，2 つの出力に対するそれぞれの重要パスを比較するための機能を設ける．

解釈支援ネットワークにおいては，2 つの出力ラベルに対する重要パスをノードラベルを，それぞれ赤と緑で可視化すると共に，2 つの出力ラベルに対する共通の重要パスを茶色で表示する．

これらの機能によって，2 つの出力を導く分類パターンについて，それぞれの出力に独自の分類パターンの解釈を支援する．

4 テキスト分類パターン解釈支援システムの有効性の検証実験

本章では，提案システムが，分類パターンの解釈に有効に用いられるかを確認するために行った実験について述べる．本実験では，理系の大学生，大学院生 14 名に，システムの利用方法について理解してもらった後，3 種類のデータ分析課題を与え、分類パターンに意味づけを行ってもらった．

4.1 実験手順

4.1.1 実験手順概要

実験の手順を以下に示す．

1. サンプルデータの分析：システムの操作を一通り試してもらう．
2. 課題 1 「動物」：動物の生態を説明するテキストの分類：2 種類の動物の生態の違いをまとめてもらう．
3. 課題 2 「映画」：映画「シン・ゴジラ」のレビュー記事の分類：映画「シン・ゴジラ」の評価されている箇所と，評価されていない箇所を分析してもらう．
4. 課題 3 「ツイート」：受験にまつわるツイートの分類：受験にまつわるツイートの中で，どのようなツイートが，より「いいね」をもらいやすいかを分析してもらう．

4.1.2 実験手順詳細

3 つの課題のそれぞれについて，以下のステップで分類パターンに意味づけを行ってもらった．

1) 2 種類の出力ラベルを選択する：「動物」課題では 2 種類の動物を被験者に選んでもらった．「映画」課題では評価が 5 のレビューと，評価が 1 または 2 のレビュー，「ツイート」課題では，いいねが 11 以上のツイートと，いいねが 0 のツイートとした．
2) 3 つの中間層のそれぞれに表示されている単語を元に，それぞれの出力につながるルールを推定して答えてもらう．
3) それぞれの出力につながる 5 つのパスとパス上の単語を元に，各出力につながるルールを推定して答えてもらう．
4) ステップ 2) とステップ 3) で答えた内容を整理して，2 種類の出力の相違点についてまとめてもらう．

4.2 実験データ詳細

学習データは，課題「動物」においては，50 種類の動物において，動物名と「生態」というキーワードで検索エンジンによる検索結果の上位の Web ページから，100 件ずつ，合計 5000 件のテキストを利用した．課題「映画」においては，映画レビューサイト[1]において，映画「シン・ゴジラ」の映画レビューの項目から，レビュー記事と評価値を 1000 件（評価 1 と 2 を評価 153 件，評価 3 を 119 件，評価 4 を 330 件，評価 5 を 398 件）利用した．課題「ツイート」においては，キーワード「受験」で検索した Twitter[2]のツイートから，

[1]映画.com (URL) http://eiga.com
[2]Twitter: (URL) http://twitter.com

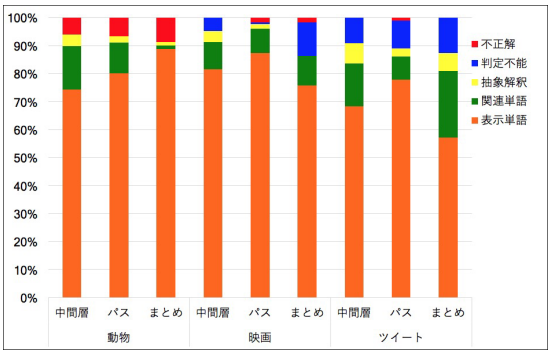

図 5: 被験者に記述された解釈の内訳（被験者平均）

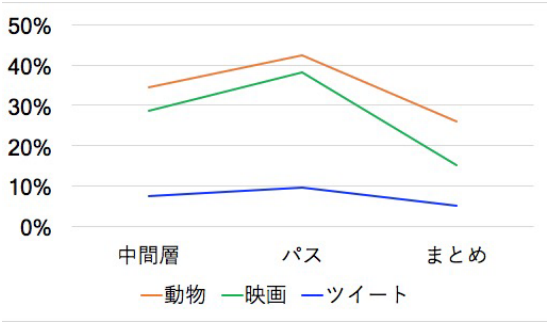

図 6: 被験者の解釈が当てはまる原文の割合（被験者平均）

フォロワー数が 100 人以上のユーザのツイートを 6000 件（いいね 0 個，いいね 1 から 10 個，いいね 11 個以上をそれぞれ 2000 件）を利用した．学習は全結合型のディープニューラルネットワークによって行い，いずれの課題も中間層は 3 層とした．

4.3 実験結果と考察

被験者により記述された解釈の内訳を図 5 に示す．ただし，表に示す解釈の内訳は以下に定義する内容をもとに，著者の 1 名が分類を行った．また，実験手順詳細のステップ 2) から 4) に対応する，中間層，パス，まとめごとに被験者平均を算出している．

- 表示単語：提案システムが表示する単語を使っており，その内容の正しさが原文から確認できる．
- 関連単語：表示単語は使われていないが，表示単語が使われている原文に含まれる単語を使っており，その正しさが原文から確認できる．
- 抽象解釈：表示単語，および表示単語が使われている原文に含まれる単語が使われていないが，その内容の正しさが原文から確認できる．
- 判定不能：解釈の正しさが判定できない．
- 不正解：解釈の内容が確実に誤っている．

図 5 の結果から，多くの解釈は「不正解」と「判定不能」以外に分類され，その正しさが確認できており，実際に当てはまる事例を具体的な知識として捉えることができていた．このことから，中間層とパスの解釈において，本システムが出力する単語情報まとめと意味づけ支援ネットワークが，分類パターンの解釈に役立てられたと考えられる．また表 5 において，「表示単語」だけでなく「関連単語」や「抽象解釈」に分類される解釈も一定の割合で存在しており，システムが提示する単語を活用した解釈が行われていたことがわかる．

課題ごとの比較として，「動物」の課題では不正解が多く，「映画」と「ツイート」の課題では判定不能が多くなった．これは動物の生態はその正しさが確認しやすかったのに対して，レビューやいいねの数などの人間の感覚に依存する評価においては，正しさの絶対的な基準がなかったことによると思われる．そのため，即座にその正しさを確認できない解釈について，その判断を支援する情報を提供できれば，より効果的なシステムに改善できると考えられる．

また課題「映画」と「ツイート」においては，中間層とパスの解釈に比べ，まとめの解釈において表示単語の割合が少なくなり，判定不能に分類される解釈が多くなった．このことから，個別の中間層やパスの解釈をまとめる際に，一定の抽象化が行われ，表示単語そのもの以外の言葉による解釈が行われていたことがわかる．そのため，複数の解釈を集めてまとめることは，より汎用的な知識を得る上で一定の意味があると考えられるが，先の考察と同様に，その正しさを検証できる仕組みが必要と考えられる．

被験者の解釈を著者らがまとめたところ，課題「映画」においては，『音楽や映像などの演出，映画が扱っているテーマ，については高い評価が得られ，ストーリーやセリフ，これまでのシリーズとの違い，については低い評価が得られた』となる．また課題「ツイート」においては，『近況報告，人との関わり，自分の心境，のツイートについては「いいね」をもらいやすく，勉強方法や，広告，のツイートは「いいね」がもらえない』となる．特に直感的には「いいね」がもらえると思われる勉強方法のツイートは，かえっていいねに結びつかないという興味深い結果も得られている．

実際の分析の場面で同様のまとめを得るためには，一人の分析者が主要な複数の解釈を列挙してまとめる道筋を用意することや，分析の初心者であっても複数人が分析を行うことで，一人のメタ分析者が結果をまとめあげる形式をとることで，有効な知見を得られる可能性もあると考えている．

被験者の解釈が当てはまる原文の割合を，図 6 に示す．ただし集計対象は，先の解釈の内訳で「表示単語」「関連単語」に分類されたもののみとし，それらの単語を含む原文の数を，各課題ごとに 2 種類の出力ラベルの学習に用いたテキスト数[3]で除した値を，解釈が当てはまる原文の割合としている[4]．ただし，図 6 の中間層の項目では，3 つの中間層に対する解釈が当てはまる原文の合計を，パスの項目では，5 つのパスに対する解釈が当てはまる原文の合計を，まとめの項目ではまとめとして書かれた 1 つの解釈が当てはまる原文の数を用いている．

まとめの解釈においては，課題「動物」「映画」「ツイート」の順に，26%(52 件)，15%(84 件)，5%(208 件)の原文に当てはまっている。この数値には判定不能となった解釈の数値は含まれていないこと，また先述の複数人の解釈をまとめる方法をとることなどにより，この割合を増やしていける可能性があることから，本システムによって，学習データとなったテキスト集合に対して一定の解釈を与える支援を行うことができたと考えている．

課題「動物」「映画」「ツイート」の順に，まとめの解釈が原文に当てはまる割合が下がったのは，テキストが対象とする内容の範囲の広さによるものと考えられる．すなわち，課題「動物」に用いたテキストは特定の動物の生態についての話であり，内容には一定のまとまりがあると考えられるのに対して,「映画」は特定の映画に対するレビューであるものの，個人の主観に依存する部分があるため，内容に広がりが生じたと考えられる.「ツイート」は，受験をテーマにしている以外には制約がないため，個人の主観に加え，個人の置かれている状況や環境の違いがあるため，さらに内容の幅が広くなり，全体に当てはまる解釈を導くことは難しかったと考えられる．

5 おわりに

本論文では，テキストベースの文章分類を題材として，深層学習により学習された分類パターンを表すネットワークの解釈を支援するシステムを提案した．提案する環境が，深層学習に精通していない，データ分析の初心者でも分類パターンの解釈につなげられることを実験により確認した．

今後は，文章分類におけるリカレントニューラルネットワークやその発展手法における学習結果の解釈や，より大規模なネットワークの解釈に繋げられる環境に改良していきたい．また，ユーザが与えた解釈の正しさを裏付けるデータを，学習データの内外から取得してユーザに提示することで，仮説として得られた知識の検証を行える枠組みを構築していきたいと考えている．

[3]課題「動物」では 200 件,「映画」では 551 件,「ツイート」では 4000 件.

[4]解釈に用いられた単語が使われている原文においても，解釈が当てはまらない可能性も考慮して，一通り原文の内容を確認し，解釈の当てはまりに問題ないことを確認している．

参考文献

[Arisoy 12] Ebru Arisoy, Tare N. Sainath, Brian Kingsbury, Bhuvaba Ramabhadran：Deep Neural Network Language Models, In Proceedings of the NAACLHLT Workshop, Will We Ever Really Replace the N-gram Model?, pp.20–28, 2012.

[Chen 16] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “Info-GAN: Interpretable representation learning by information maximizing generative adversarial nets,” In Proceedings of NIPS, 2016.

[ダヌシカ 14] ボレガラ ダヌシカ：自然言語処理のための深層学習，人工知能学会誌，Vol.29，No.2, pp.195–201, 2014.

[Krizhevsky12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” In Proceedings of NIPS, 2012.

[LeCun98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” In Proceedings of the IEEE, 1998.

[Mahendran 15] Aravindh Mahendran and Andrea Vedaldi, “Understanding deep image representations by inverting them,” In Proceedings of CVPR, 2015.

[西銘 15] 西銘 大喜：ディープニューラルネットワークによる画像からの表情表現の学習，第 29 回人工知能学会全国大会，3L4-3，2015.

[巣篭 16] 巣篭 悠輔：Deep Learning Java プログラミング 深層学習の理論と実装，株式会社インプレス，2016.

[Wu17] T. Wu, X. Li, X. Song, W. Sun, L. Dong, and B. Li, “Interpretable R-CNN,” In Proceedings of arXiv:1711.05226, 2017.

[Zeiler 14] Matthew D. Zeiler and Rob Fergus, “Visualizing and understanding convolutional networks,” In Proceedings of ECCV’14, pp.818-833, 2014.