The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

LRP 法による畳み込みニューラルネットワークにおける中間層の振る舞いの可視化

Visualizing the Behavior of the Inner Layers of Convolutional Neural Networks by Layer-wise Relevance Propagation

酒井博貴 *1 Hirotaka Sakai　亀谷由隆 *1 Yoshitaka Kameya　曽田尚宏 *2 Takahiro Sota　有江浩明 *2 Hiroaki Arie

*1名城大学 理工学部 情報工学科
Department of Information Engineering, Meijo University

*2株式会社デンソー 先端研究 4 部
Advanced Research and Innovation Div. 4, DENSO CORPORATION

In recent years, complex machine learning models like deep neural networks play a central role in many real applications, due to their high predictive performance. Interpreting machine learning models is then considered to be important since practitioners constantly need clues for improvement on such complex models, whose behavior is not directly visible to human. In this paper, we focus on the inner workings of convolutional neural networks, visualize them by a method called layer-wise relevance propagation, and report several findings from the visualization.

1. はじめに

深層学習技術の進展により，画像分類等のタスクでは人の能力を超える予測・認識結果を得られるようになり，現実場面でも深層学習が取り入れられ始めている．その一方で，予測・認識結果を解釈・説明することが困難であることが問題となってきており，予測・認識結果の説明責任が問われる医療等のミッションクリティカルな分野への応用の妨げとなっている．予測・認識結果の解釈や説明が可能になれば，学習モデルの改善，責任所在の特定，学習モデルからの知見の獲得が可能になり，より多くの分野への応用が容易になると考えられる．

人工知能分野ではこれまでもニューラルネットワーク（neural network, 以下 NN）等の学習モデルや出力結果を解釈・説明する手法が提案されてきたが [Hotho 03, Duch 04, Yuan 11], 上述の背景によりその進展が著しくなっており，特に畳み込み NN (convolutional NN, 以下 CNN）に対する解釈・説明手法は数多く提案されている [Grün 16, Selvaraju 17]．本研究では，画像分類タスクにおける CNN の中間層の振る舞いに着目し，入力画像内で中間層の出力に強く関連する箇所の可視化を試みる．そのため，学習モデルの入出力のみを解釈・説明する手法 [Ribeiro 16, Lundberg 17] ではなく，CNN の構造に沿って出力への関連度を逆伝播する LRP (layer-wise relevance propagation) 法 [Bach 15] を利用する．

2. LRP 法

d 次元入力ベクトル x に対して値 $f(x)$ を出力する学習モデルがあったときに，LRP 法では各入力要素 x_i $(1 \le i \le d)$ が $f(x)$ に関連する度合い（関連度, relevance）R_i を求める．関連度 R_i は入力ベクトル x 中の重要箇所を示しており，特に x の各要素が画像のピクセルに対応する場合は R_i の大きさに基づいてヒートマップを作成することで画像中の重要箇所をハイライトできる．

LRP 法は NN に限らず利用可能な手法であるが [Bach 15], ここでは図 1 の全結合 NN に対して LRP 法が動作する様子を説明する．NN における通常の予測では，図 1 (a) のように，第 l 層の素子 i の活性値 x_i と第 $l+1$ 層の素子 j への重み w_{ij} から $z_{ij} = x_i w_{ij}$ を計算する．そして第 $l+1$ 層の素子 j の活性値は $x_j = \sigma(\sum_i z_{ij} + b_j)$ として計算される．ここで，b_j はバイアス項，σ は（非線形の）活性化関数である．一方，LRP 法では入力 x と出力 $f(x)$ の対が与えられたときに，図 1 (b) のように関連度を出力層から入力層に向かって逆伝播させる．例えば，simple-LRP と呼ばれる簡単な伝播規則では

$$R_i^{(l)} = \sum_j R_{i\leftarrow j}^{(l,l+1)}, \quad R_{i\leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j} R_j^{(l+1)} \tag{1}$$

のように第 l 層の関連度 $R_i^{(l)}$ が第 $l+1$ 層の素子 j 間の関連度 $R_j^{(l+1)}$ に基づき計算される．ただし，$z_j = \sum_i z_{ij}$ である．入力要素 x_i の関連度 R_i は $R_i^{(1)}$ と定義する．更に，LRP 法では ϵ-LRP（式 2）や $\alpha\beta$-LRP（式 3）という伝播規則も導入されている．

$$R_{i\leftarrow j}^{(l,l+1)} = \frac{z_{ij}}{z_j + \epsilon \cdot \mathrm{sign}(z_j)} R_j^{(l+1)} \tag{2}$$

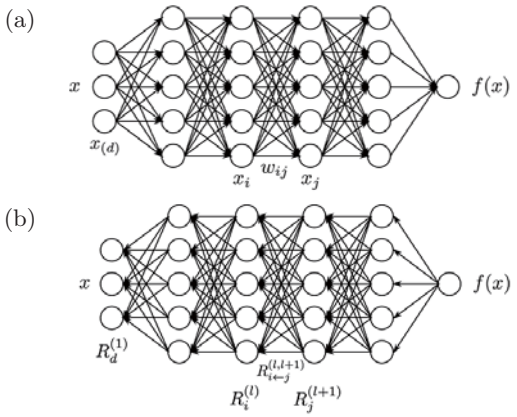


図 1: (a) 通常の予測における活性値の順伝播，(b) LRP 法における関連度の逆伝播（[Bach 15] より引用）

連絡先: 亀谷由隆，名城大学理工学部情報工学科，468-8502 愛知県名古屋市天白区塩釜口 1-501，ykameya@meijo-u.ac.jp

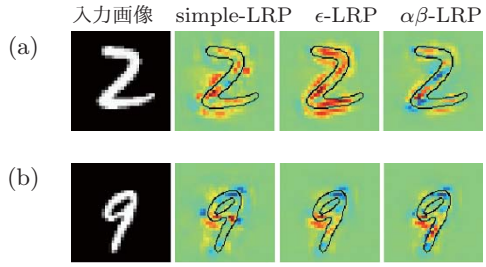


図 2: (a) 入力画像「2」と予測出力 2（正解）を与えた場合，(b) 入力画像「9」と仮想出力 4（誤り）を与えた場合

$$R_{i\leftarrow j}^{(l,l+1)} = \left(\alpha\cdot\frac{z_{ij}^{+}}{z_{j}^{+}} + \beta\cdot\frac{z_{ij}^{-}}{z_{j}^{-}}\right) R_{j}^{(l+1)} \qquad (3)$$

ϵ-LRP は z_j が小さな場合に式 1 が数値的に不安定になるのを避けることができる．また，$\alpha\beta$-LRP において，$z_{ij}^{+} = \max\{z_{ij}, 0\}$, $z_{ij}^{-} = \min\{z_{ij}, 0\}$, $z_{j}^{+} = \sum_i z_{ij}^{+}$, $z_{j}^{-} = \sum_i z_{ij}^{-}$ であり，出力との不整合を意味する負の関連度を多く伝搬するよう α, β を設定できる．ただし，$\alpha+\beta = 1$ という制約を満たさなければならない．なお，CNN においてはプーリング層や入力層について特別な伝播規則を用いるが [Samek 17, Montavon 18]，ここでは詳細を省略する．これらの伝播規則は各層の間で関連度の総和を保存する（すなわち $\sum_i R_i = \cdots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \cdots = f(x)$）ように定義されており，出力 $f(x)$ は入力要素の関連度 R_i の和として近似的に分解される．この意味で LRP 法は，LIME 法 [Ribeiro 16] と同様に，NN には手を加えずに近似モデルで特定入力に対する出力を事後解釈 (post-hoc interpretation) [Lipton 16] する手法であると言え，予測モデルと解釈・説明モデルを同時に学習する手法 [Lei 16, Al-Shedivat 17] とは対照的である．

CNN に対する LRP 法の実行例として，MNIST データセット中の手書き文字画像を入力として与えた場合を図 2 に示す．図 2 (a) では，入力画像「2」と正解の予測出力 2 を LRP 法に与えたときに得られる関連度のヒートマップが伝播規則ごとに示されている ($\epsilon = 1$, $\alpha = 3$, $\beta = 2$)．本論文で示すヒートマップでは赤色が関連度が強い箇所，青色が関連度の低い箇所を表す．図 2 (a) を見ると「2」という数字を認識するのに重要な箇所が関連度が高くなっている．理論的裏付けがあるのは $\alpha = 1, \beta = 0$ の $\alpha\beta$-LRP であるが [*1]，著者らが試した範囲では ϵ-LRP が全般的に見やすいヒートマップを出力したため，以降の実行結果では $\epsilon = 1$ の ϵ-LRP で得られたものを示す．また，LRP 法では実際の予測とは異なる仮想的な出力を正解として与えることもできる．図 2 (b) は入力画像「9」に対して誤りの仮想出力 4 を与えた場合の関連度のヒートマップである．これを見ると，出力 4 が正解として与えられたこと整合しないと思われる箇所が青色で示されている．

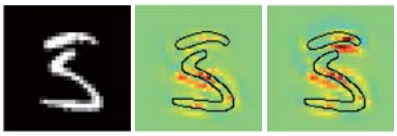


図 3: CNN が正しく判断できなかった場合

*1 [Montavon 17] に一部の伝播規則で得られる関連度が出力 $f(x)$ の一次テイラー展開の結果であることが示され，[Montavon 18] に伝播規則の使い分けのガイドラインが記述されている．

表 1: MNIST 用 CNN の構成

層番号	種類	入力次元 H	入力次元 W	入力次元 D	出力次元 H	出力次元 W	出力次元 D	カーネル	ストライド
0	入力	32	32	1	32	32	1	—	—
1	畳み込み	32	32	1	28	28	10	5×5	1×1
2	プーリング	28	28	10	14	14	10	2×2	2×2
3	畳み込み	14	14	10	10	10	25	5×5	1×1
4	プーリング	10	10	25	5	5	25	2×2	2×2
5	畳み込み	5	5	25	2	2	100	4×4	1×1
6	プーリング	2	2	100	1	1	100	2×2	2×2
7	畳み込み	1	1	100	1	1	10	1×1	1×1
8	Flatten	1	1	10		10		—	—

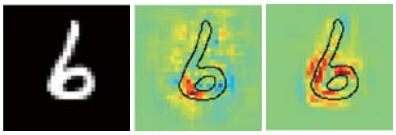


図 4: 入力画像「6」と関連度のヒートマップ

この性質を使うと予測に失敗した理由が分かることがある．例えば図 3 の入力画像は人間にも「3」か「5」の判断が難しいが，CNN は 3 と予測した．その予測出力 3 に対するヒートマップは左のように得られ，確かに青色の部分が少ない．一方，5 を正解出力として与えると右のヒートマップが得られる．上部の横線が下と繋がっていない箇所，上部の横線が丸みを帯びている箇所が青く，ここが問題だと判断されたことが分かる．

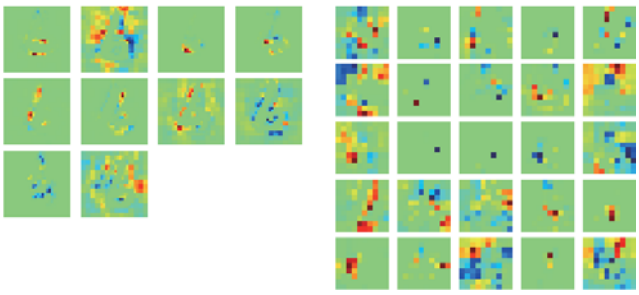
3. 中間層の振る舞いの可視化

前節の LRP 法の実行例では入出力の対を与えて入力と出力の対応関係を可視化した．更に本研究では，CNN における中間層の振る舞いを可視化するために，入力と出力の対応関係に限らない形で LRP 法を利用する．対象としたデータセットは MNIST と CIFAR-10 である．

3.1 MNIST データセットを用いた実験

MNIST データセットに対して LeNet-5 を参考に構成された表 1 の CNN を動かす．畳み込み層では活性化関数に ReLU を用い，プーリング層では平均値プーリングを行っている．Flatten 層では直前の 3 次元情報を 1 次元に落とす．実験では LRP 法の研究グループによる Python 独自実装版 [*2] を用いた．

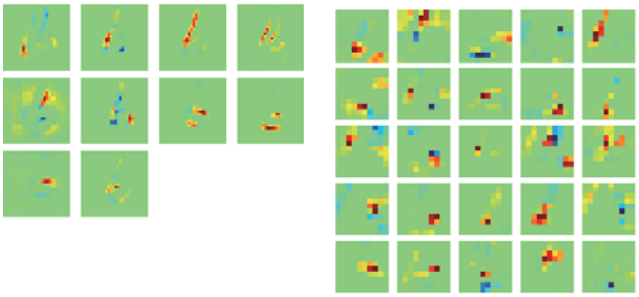
この実験では表 1 の CNN の中間層における特徴マップを関連度でヒートマップ化して入力の特徴がどのように抽出さ



(a) 1–2 層間　(b) 3–4 層間

図 5: 低精度モデルに画像「6」を入力したときの特徴マップ

*2 https://github.com/sebastian-lapuschkin/lrp_toolbox



(a) 1–2 層間　(b) 3–4 層間

図 6: 高精度モデルに画像「6」を入力したときの特徴マップ

れているか観察する．用いる関連度は第 2 層（プーリング層）から第 1 層（畳み込み層）へ渡されるもの，第 4 層（プーリング層）から第 3 層（畳み込み層）へ渡されるものである．また学習の効果を見るために，十分に学習を行った高い予測精度（正答率）99.23 %の CNN と学習回数を減少させた低い予測精度 41.94 %の CNN の 2 種類で比較する．

図 4 は入力画像「6」と出力 6 を与えたときの入力層でのヒートマップ（左：低精度，右：高精度）である．これを見ると高精度モデルの方がより数字「6」の特徴を捉えていると言える．また，同じ入力画像を与えたときの低精度モデルと高精度モデルの特徴マップは各々図 5, 6 のようにヒートマップ化される．まず低精度モデルの特徴マップでは，正解出力を与えたにも関わらず青色の箇所が多く，数字「6」の特徴がうまく捉えられていないように見える（図 5）．一方，高精度モデルの入力側に近い特徴マップ（図 6 (a)）では入力のエッジを捉えており，入力から離れた特徴マップ（図 6 (b)）では数字「6」の特徴を更に細かく捉えている．このように特徴マップを観察することで，学習がうまく行われたかどうかをある程度判断できるのではないかと考えられる．

3.2 CIFAR-10 データセットを用いた実験

CIFAR-10 データセットに対して表 1 の構成の CNN を動かす．予測精度は 69.92 %である．畳み込み層では活性化関数に ReLU を用い，プーリング層では最大値プーリングを行っている．第 9 層と第 10 層の全結合層では活性化関数として ReLU を用いるが，第 11 層では恒等関数を用いる．また，学習時には第 3 層と第 5 層の畳み込み層でドロップアウトを各々 0.9, 0.8 と設定した．実験では LRP 法の研究グループによる TensorFlow 実装版 [*3] を用いた．

この実験では，CNN の全結合層・出力層は使わずに，畳み込み層およびプーリング層によって特徴がうまく抽出できているかどうかを確認する．具体的には，第 8 層（プーリング層）と第 9 層（全結合層）の間で伝播される関連度を人為的に変更してそこから入力層に向かって関連度を逆伝播する．関連度の変更においては第 8 層中の 300 素子から無作為に素子 i $(0 \leq i \leq 299)$ を 1 つ選択し，素子 i の関連度を 1 に，他の 299 素子の関連度を 0 にする．この操作を行うと入力層で得られるヒートマップは素子 i と強く関連する入力要素をハイライトすることになる．また入力画像は，テスト用画像 10,000 枚の中から無作為に選び，元画像そのまま，左右反転した画像，色相をランダムに変化させた画像の 3 種類を与える．

図 7 は馬の画像に対して，関連度を 1 に変更した素子と入力画像の加工方法を変えたときの入力層におけるヒートマップ

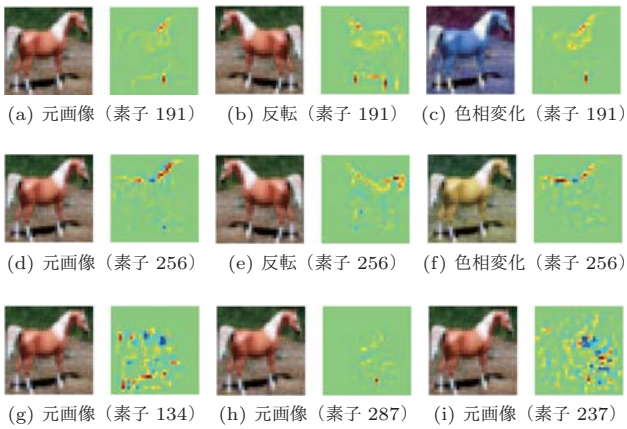
*3 https://github.com/VigneshSrinivasan10/interprettensor

表 2: CIFAR-10 用 CNN の構成

層番号	種類	入力次元 H	入力次元 W	入力次元 D	出力次元 H	出力次元 W	出力次元 D	カーネル	ストライド
0	入力	32	32	3	32	32	3	—	—
1	畳み込み	32	32	3	30	30	300	3×3	1×1
2	プーリング	30	30	300	15	15	300	2×2	2×2
3	畳み込み	15	15	300	14	14	300	2×2	1×1
4	プーリング	14	14	300	7	7	300	2×2	2×2
5	畳み込み	7	7	300	5	5	300	3×3	1×1
6	プーリング	5	5	300	3	3	300	2×2	2×2
7	畳み込み	3	3	300	2	2	300	2×2	1×1
8	プーリング	2	2	300	1	1	300	2×2	2×2
9	全結合	1	1	300		300		—	—
10	全結合		300			100		—	—
11	全結合		100			10		—	—
12	Softmax		10			10		—	—

である．191 番目の素子の関連度を 1 に変更した場合，(b)(c) のように入力画像を加工しても (a) と変わらず馬の胴体・首と足の関連度が高いことから，この素子は胴体と足を特徴として捉えている素子であると考えられる．左右反転しても入力画像に対応して胴体や足の部分が捉えられており，使用している CNN は左右反転に対応できるよう学習できていることが分かる．また，256 番目の素子の関連度を 1 に変更した場合，(d)~(f) のように馬の背中の関連度が高くなっていることから，背中を特徴として捉えている素子であると分かる．一方で，(g)~(i) のようにヒートマップが全体的に乱れた素子や，ヒートマップ中に高い関連度を示す箇所が見られずに入力画像のどこを特徴として捉えているか不明な素子も存在した．

図 8 は，同じクラスの異なる入力画像に対して，特定の素子の関連度を 1 に変更したときのヒートマップである．まず自動車クラスの画像に対するヒートマップ (a)~(c) において，(a) 素子 101 では自動車のタイヤに対して高い関連度は見られないが，(c) 素子 252 ではタイヤに該当する箇所の関連度が高いことが分かる．また，素子 252 と素子 295 はどちらもタイヤ部分を捉えていることが多い．同様に馬クラスの画像に対するヒートマップ (d)~(f) において，素子 191 と 256 は図 7 でも注目した素子である．また，図 8 の入力画像の中には図 7 と異なり，足や胴体が画像中にない画像も含めている点に注意する．図 7 の説明において素子 256 は馬の背中を特徴として捉えていると解釈していたが，他の画像では足部分の関連度が



(a) 元画像（素子 191）　(b) 反転（素子 191）　(c) 色相変化（素子 191）

(d) 元画像（素子 256）　(e) 反転（素子 256）　(f) 色相変化（素子 256）

(g) 元画像（素子 134）　(h) 元画像（素子 287）　(i) 元画像（素子 237）

図 7: 画像加工時の特定素子から逆伝播した関連度の違い

(a) 自動車，素子 101 (b) 自動車，素子 252 (c) 自動車，素子 295

(d) 馬，素子 84 (e) 馬，素子 191 (f) 馬，素子 256

図 8: 同じクラスにおける同じ素子の振舞い

高い．一方，素子 191 はやはり馬の胴体と足を特徴として捉えている．興味深いのは素子 191 が胴体や足がない馬の画像においては顔の輪郭を捉えている点である．NN の素子と入力における特徴の間には多対多の（複雑な）関係が成り立っているものと予想される．

最後に，図 9 は，素子 10 の関連度の値を 1 に変更したときの猫と自動車のクラスから選択した入力画像とヒートマップである．この場合，対象（猫や自動車）が存在する箇所への関連度は低く，対象以外の部分に関連度が見られる．これらのことから，素子 10 は猫や自動車の特徴を捉えていない素子なのではないかと考えられる．他のクラスに興味がない場合，このような素子は削除して計算コストを削減することも考えられる．

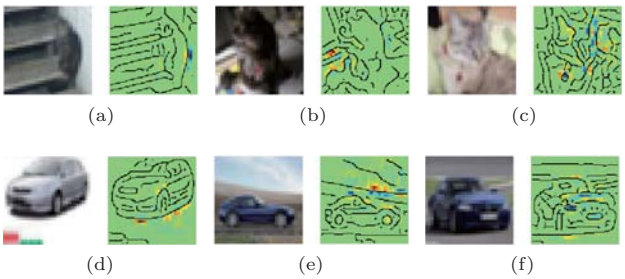
4. おわりに

LRP 法による CNN の中間層の振舞いをヒートマップとして可視化して分析した．中間層の素子が入力のどの部分を特徴として捉えているのか一定の傾向を把握できた．その一方で，CIFAR-10 については実装上の理由により 70 %弱の予測精度を持つ CNN でしか実験を行うことができなかった．現時点では予測精度 90 %を大きく超える CNN が存在しており，改めて検証実験を行う必要がある．また，学習モデルの解釈・説明ができたとしても，そこから学習モデルの改善へはまだギャップがあり [*4]，引き続き他のデータセットでも実験を行い，知見を増やしたい．また，LRP 法では得られたヒートマップの定量的な評価手法が提案されており [Samek 17, Montavon 18]，本実験で行った主観的評価とこれらの評価手法の対応関係を確認する必要がある．

参考文献

[Al-Shedivat 17] Al-Shedivat, M., Dubey, A., and Xing, E. P.: The Intriguing Properties of Model Explanations, in *Proc. of NIPS-17 Symp. on Interpretable Machine Learning* (2017)

*4 [Zeiler 14] では CNN の可視化の結果からストライド幅等を調整して予測精度向上につなげられたとの報告がある．



(a) (b) (c)

(d) (e) (f)

図 9: 素子 10 の関連度を 1 としたときのヒートマップ

[Bach 15] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W.: On Pixel-wise Explanations for Non-linear Classifier Decisions by Layer-wise Relevance Propagation, *PLOS ONE*, Vol. 10, No. 7 (2015)

[Duch 04] Duch, W., Setiono, R., and Zurada, J. M.: Computational Intelligence Methods for Rule-based Data Understanding, *Proc. of the IEEE*, Vol. 92, No. 5, pp. 771–805 (2004)

[Grün 16] Grün, F., Rupprech, C., Navab, N., and Tombari, F.: A Taxonomy and Library for Visualizing Learned Features in Convolutional Neural Networks, in *Proc. of ICML-16 Workshop on Visualization for Deep Learning* (2016)

[Hotho 03] Hotho, A., Staab, S., and Stumme, G.: Explaining Text Clustering Results using Semantic Structures, in *Proc. of PKDD-03*, pp. 217–228 (2003)

[Lei 16] Lei, T., Barzilay, R., and Jaakkola, T.: Rationalizing Neural Predictions, in *Proc. of EMNLP-16*, pp. 107–117 (2016)

[Lipton 16] Lipton, Z. C.: The Mythos of Model Interpretability, in *Proc. of ICML-16 Workshop on Human Interpretability in Machine Learning* (2016)

[Lundberg 17] Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in *Proc. of NIPS-17* (2017)

[Montavon 17] Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R.: Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition, *Pattern Recognition*, Vol. 65, pp. 211–222 (2017)

[Montavon 18] Montavon, G., Samek, W., and Müller, K.-R.: Methods for Interpreting and Understanding Deep Neural Networks, *Digital Signal Processing*, Vol. 73, pp. 1–15 (2018)

[Ribeiro 16] Ribeiro, M. T., Singh, S., and Guestrin, C.: Why Should I Trust You? — Explaining the Predictions of Any Classifier, in *Proc. of KDD-16*, pp. 1135–1144 (2016)

[Samek 17] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R.: Evaluating the Visualization of What a Deep Neural Network Has Learned, *IEEE Trans. on Neural Networks and Learning Systems*, Vol. 28, No. 11, pp. 2660–2673 (2017)

[Selvaraju 17] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, in *Proc. of ICCV-17*, pp. 618–626 (2017)

[Yuan 11] Yuan, C., Lim, H., and Lu, T.-C.: Most Relevant Explanation in Bayesian Networks, *J. of Artificial Intelligence Research*, Vol. 42, pp. 309–352 (2011)

[Zeiler 14] Zeiler, M. D. and Fergus, R.: Visualizing and Understanding Convolutional Networks, in *Proc. of ECCV-14*, pp. 818–833 (2014)