

Object Detection における検出位置精度を考慮した Confidence Calibration

田口賢佑^{†1} 林佑介^{†1}

概要: 近年深層学習を用いた Object Detection は様々な手法が提案されており、自動運転技術への応用が期待されている。一般に安全面が重視されるアプリケーションでは、解析性やセンサフュージョンにおける信頼性の観点から正しい Confidence 出力を得ることが重要である。理想的な Confidence 出力を得る手法として、Temperature Scaling による Confidence Calibration が提案されており、Object Detection への適用も考えられる。しかしながら、Object Detection における Confidence 出力は検出性能評価や後処理に用いられるため、カテゴリ判別と位置精度の情報が反映されている必要がある。本稿では、このような課題に対応するために Object Detection における検出位置精度を考慮した Confidence Calibration 手法を提案する。検証として一般交通シーンの歩行者検出タスクにおいて、提案手法の Confidence 性能と認識精度の評価を行った。

キーワード: Confidence Calibration, Object Detection, Convolutional Neural Network

1. はじめに

近年深層学習を用いた画像認識の分野で Convolutional Neural Network (CNN) [1]が非常に高い精度を達成している。CNN は画像を用いた様々なタスクに用いられ、画像中の対象物体の位置とカテゴリを推測する Object Detection タスクへの応用も進んでいる。一般交通シーンにおける Object Detection は車道に存在する歩行者や他車両との衝突を未然に防ぐ予防安全装置として利用が進んでおり、自動運転技術への応用が期待されている。

一般交通シーンのような安全面が重視されるアプリケーションでは、解析性やセンサフュージョンにおける信頼性の観点から、認識精度に対する正しい Confidence 出力を得ることが重要である。認識精度に対する正しい Confidence 出力が得られた状態とは Confidence と Accuracy が等価の状態である。

しかしながら、一般的に深層学習のスキームでは認識精度に対する正しい Confidence 出力を得られない場合がある。図 1 に画像認識タスクにおける深層学習モデルの Confidence と Accuracy の関係を示す。特に図 1 のように、認識精度に対して高い Confidence を示す Over-Confidence 状態に陥りやすいことが報告されている [2]。Over-Confidence 状態では間違った認識結果に対しても高い Confidence を出力しやすいため、Confidence が認識結果に対する説明情報として機能しないという問題がある。同様に、認識精度に対して低い Confidence を示す Under-Confidence 状態も説明情報として機能しないため問題となる。さらに、認識結果を他の情報と組み合わせるようなフュー

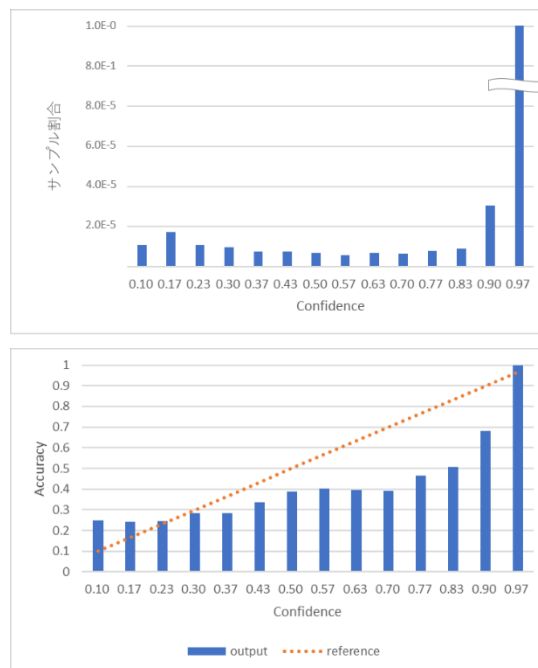


図 1 Over-Confidence 例. (上段) Confidence 分布. (下段) Confidence に対する Accuracy 特性

ジョンの場面においても説明情報が機能しなければ問題が生じる。このような状態を改善する手法として、Confidence Calibration が提案されており、近年では深層学習を対象とした様々な手法が発表されている。

Object Detection においても、認識精度に対する正しい Confidence 出力を得ることは、(Confidence) が検出精度評価や検出結果後処理に利用されることを鑑みると、非常に

^{†1} 京セラ (株)
Kyocera Corp.

重要である。検出精度評価においては、検出結果の再現率と精度はトレードオフの関係であるため、Average Precision (AP) や Average Log Miss Rate が用いられる。その際、認識結果スコアとして Confidence が使用され、高い Confidence の検出結果が優先的に評価される。検出結果後処理においては Non Maximum Suppression (NMS) に代表されるような検出結果の重複を防ぐ手法が用いられる。NMS は重複度の高い検出結果の中で Confidence の低い検出結果を抑制対象とする処理である。このように Object Detection における Confidence は検出結果に影響を及ぼすため、タスクに求められる要素であるカテゴリ判別と位置精度の情報が反映されている必要がある。例えば、Confidence 値の高い検出結果の位置精度が高ければ高位置精度な Object Detection を実現できる。従って、Object Detection における Confidence Calibration はカテゴリ判別と位置精度の観点で実施するべきである。

本研究では Object Detection における Confidence Calibration としてカテゴリ判別だけでなく、位置精度を考慮する手法を提案する。提案手法を用いることで認識精度と位置精度に対する Confidence 特性が改善されることを狙う。実験として歩行者検出タスクで Confidence 特性と正解領域に対する IoU 閾値を変えた検出精度を評価し、提案手法の有効性を確認した。

2. 関連研究

Confidence Calibration はモデル学習時に実施する手法と学習済みモデルに実施する手法がある。モデル学習時に実施する手法として、Confidence 偏りに対してペナルティを掛ける手法[3]やモデル不確実性を損失関数に反映することで Over-Confidence を防ぐ手法[4]が提案されている。[5]は教師無し学習のスキームで Confidence 調整パラメータを学習する手法であり、Object Detection タスクを対象として検証している。しかしながら、Over-Confidence 状態に対するペナルティにより検出精度の悪化が示されている。学習済みモデルに実施する手法として Temperature Scaling[20]を用いた Confidence Calibration[2]が提案されている。さらに[6]は Confidence 領域別に温度パラメータを設定することで、Calibration 性能向上を実現している。これら学習済みモデルに実施する手法はモデルの再学習が不要なため、モデル学習に関与できない場合においても有効な手法である。

CNN ベースの Object Detection として R-CNN[7]が提案されて以降、様々な手法が提案されている。Faster R-CNN[8]は物体候補領域を推定するネットワークと、候補領域に対するクラス分類、詳細位置を推定するネットワークで構成されている 2-stage 型 Detector の代表例である。YOLO[9]や Single Shot Detector(SSD)[10]は候補領域推定からクラス分

類まで一気通貫で学習できる 1-stage 型 Detector であり、計算コストの面で 2-stage 型に対し優位である。SSD は特徴マップ毎にアスペクト比、スケールの異なる Default Box を定義し、各 Default Box 毎にカテゴリ、詳細位置予測を行う。よって、2-stage 型で必要となるクラス分類ネットワークに入力するための特徴マップ再サンプリング処理が不要であり、シンプルな学習が実現できる。DSSD[11]や Retina Net[12]、Refine Det[13]、M2Det[14]など SSD をベースにさらなる改良が加えられた 1-stage 型 Detector が近年多く提案されている。

本研究では最も基本的な 1-stage 型 Detector である SSD に対して、学習済みモデルに対する Confidence Calibration を検証する。

3. 提案手法

本研究では、学習済みの Object Detection モデルに対して Temperature Scaling による位置精度に応じた Confidence Calibration を提案する。温度パラメータを最適化する際に学習データの位置精度に応じた重み付き学習を行うことで位置精度に対する Confidence 特性の改善を狙う。以下に提案手法の詳細を述べる。

3.1 Temperature Scaling

Temperature Scaling は温度パラメータ T によって、尤度 (Confidence) 分布をスケールリングする手法である。 T は Confidence を出力する SoftMax 関数の中で使用される。 T の効果として、 $T > 1$ の場合は低い Confidence が強調されるため、分布をなだらかにする効果、 $T < 1$ の場合は高い Confidence が強調されるため分布が偏る効果がある。 $T = 1$ であれば元の Confidence を出力する。本研究が対象とする 1-stage 型 Detector は、Default Box 毎に Confidence を出力する SoftMax 関数が存在し、Default Box はそれぞれが固有の位置、アスペクト比、スケールを持っている。そこで、本研究では T を Default Box 毎に設定することで、位置、アスペクト比、スケールの特性を考慮した Confidence Calibration を実現する。総数 N の各 Default Box を $D_i, i = 0, 1, 2, \dots, N$ とし、 T_i でスケールされた Confidence 分布 q_i は式(1)のように示される。

$$q_i = \sigma_{SM}(Z_i/T_i) \quad (1)$$

Z_i は最終層ベクトル、 σ_{SM} は SoftMax 関数である。

3.2 位置精度に対する重み付き学習

本研究では位置精度に対する Confidence 特性を改善するために、 T_i を最適化する際に位置精度に対する重み付き計算を行う。これにより、位置精度が高い D_i の Confidence エラーを優先して修正する。 D_i と正解領域 Y_i の IoU (Intersection

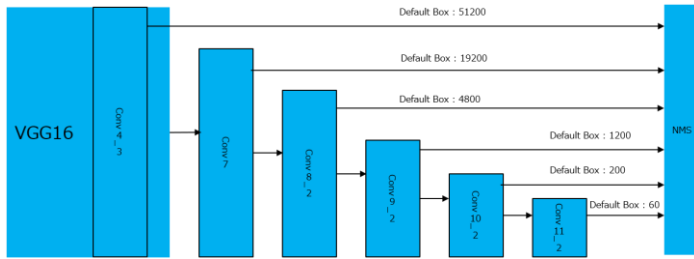


図 2 SSD の Default Box

over Union)を $Area_i$ とすると位置精度に対する重み w_i は式(2)のように示される.

$$w_i = (Area_i + \alpha) \times \beta \quad (2)$$

ここで α, β は調整パラメータである. T_i は学習済みモデルの評価用データに対して Negative Log Likelihood(NLL)を最小化するように、式(3)のような損失関数で最適化される.

$$L_i = NLL(q_i, Y_i) \times w_i \quad (3)$$

4. 評価実験

提案手法の有効性を確認するため、一般交通シーンを対象とした歩行者検出タスクで Confidence 特性と検出精度を評価する.提案手法に対して温度パラメータを全体に一つと設定した場合、位置精度の重み付き学習の有無を比較する.以下にデータセット、Detector、評価手法、評価結果を述べる.

4.1 データセット

評価データには、Citypersons[15]データセットを用いる.Citypersons は Cityscapes[16]データセットの画像が使われており、歩行者領域に対してバウンディングボックスアノテーションが付けられている.特徴としてオクルージョンしている被写体が多数存在しており、難易度が高いデータセットである.本実験では難易度設定としてオクルージョン、小スケール被写体が比較的少ない、リーズナブルセットアップで実験を行う.学習用データとして 2975 枚、評価用として 500 枚が用意されており、学習用データを Detector の学習に使用し、評価用データは Confidence Calibration の学習と評価に使用する.

4.2 Detector

Detector として SSD を使用する.バックボーンネットワークには VGG16[17]を用いた.本研究では 1280x640 にリサイズした Citypersons の画像を入力し、図 2 に示すような特徴マップ上にカテゴリー判別と詳細位置回帰を行う Default Box を設定した.学習の最適化手法には Adam[18]を用いた.

表 1 Confidence Calibration 性能比較

手法	ECE	ACE	NLL
Calibration 無し	1.E-02%	15%	8.E-04
Single	8.E-03%	13%	6.E-04
Single(w)	8.E-03%	13%	6.E-04
Box	3.E-03%	11%	4.E-04
Box(w) (Ours)	4.E-03%	11%	4.E-04

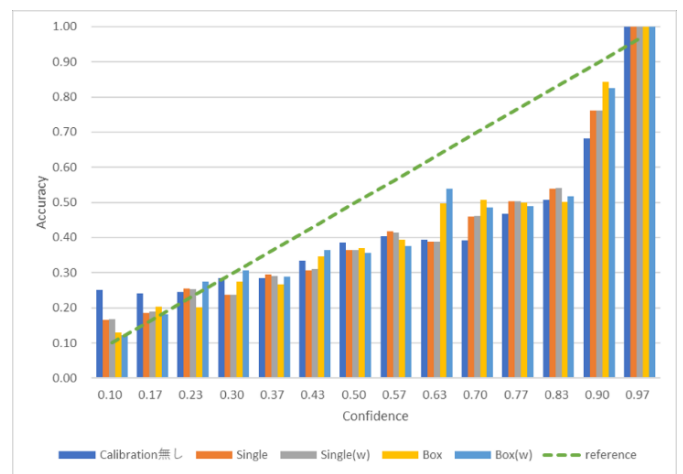


図 3 Confidence と Accuracy 特性

4.3 評価手法

認識精度に対する Confidence 特性の評価は Expected Calibration Error(ECE)[19]、Average Calibration Error(ACE)[5]を用いる.ECE は任意のビン数で分割された Confidence と認識精度の差分を評価する手法である.テストサンプル数を J 、ビン数を M 、ビン内の認識精度を Acc 、ビン内の平均 Confidence を $Conf$ とすると ECE は式(4)のように示される.

$$ECE = \sum_{i=1}^M \frac{|B_m|}{J} |Acc(i) - Conf(i)| \quad (4)$$

ECE はビン内の出現頻度重み付きで算出されるため、評価データの全体傾向を示す手法である.一方、ACE は出現頻度重み付き無しでビン内を平等に評価する手法で式(5)のように示される.

$$ACE = \sum_{i=1}^M |Acc(i) - Conf(i)| \quad (5)$$

表 2 IoU 閾値別 AP

手法	AP@0.5	AP@0.6	AP@0.7	AP@0.8	AP@0.9
Calibration 無し	0.811	0.778	0.659	0.412	0.064
Single	0.833	0.786	0.670	0.419	0.066
Single(w)	0.833	0.793	0.671	0.423	0.067
Box	0.848	0.806	0.672	0.421	0.066
Box(w) (Ours)	0.848	0.805	0.677	0.430	0.068

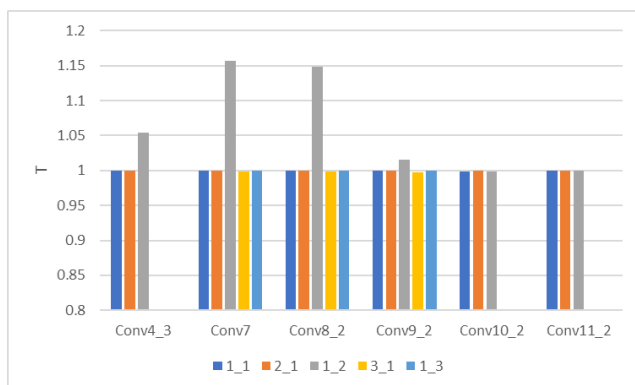
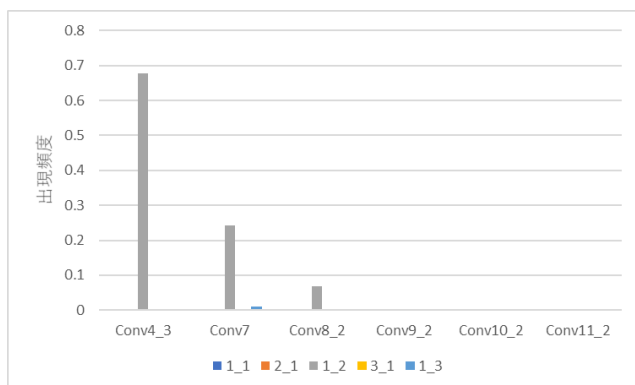


図 4 Default Box 別オブジェクト出現頻度と T.
(アスペクト比は横、縦で記載)

(上) オブジェクト出現頻度, (下) 温度パラメータ T

ACE は頻度に関係なく認識精度と Confidence が乖離している領域を示すため、安全が重視されるアプリケーションの評価に用いられる。本研究では ECE、ACE とともにビン数 $M=12$ で検証する。

4.4 Confidence Calibrations 性能評価結果

提案手法の有効性を検証するために温度パラメータ T の設定方法と位置精度に対する重み付き学習(w)の有無で Confidence Calibrations 性能を比較する。温度パラメータ T の設定方法は T をグローバルパラメータとして共通設定とした場合(Single)、提案手法のように Default Box 毎に設定す

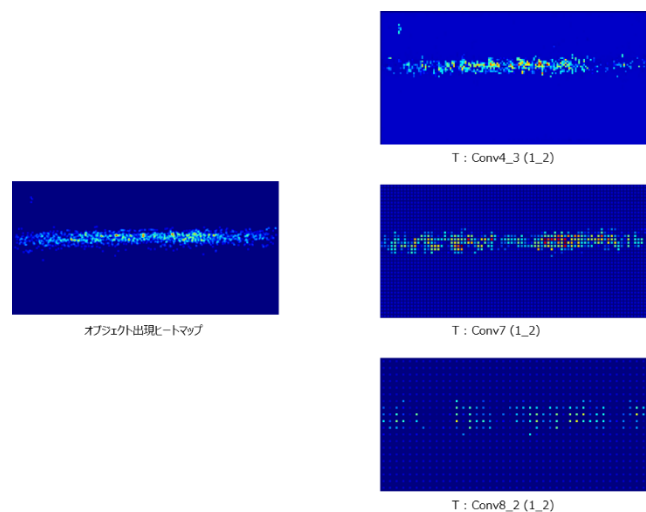


図 5 オブジェクト出現ヒートマップと T 分布
(左) オブジェクト出現ヒートマップ。(右) T 分布

る場合(Box)する。

上記条件での評価結果を表 1、図 3 に各ビン別の Confidence と Accuracy の関係を示す。位置精度の重み付き学習有無による Calibration 性能差は少ないが、T の設定方法によって Calibration 性能に差が生じており、T を Box 毎に設定することが Calibration 性能向上に寄与していることが分かる。図 4 に Default Box のスケールとアスペクト比別に集計した Citypersons のオブジェクト出現頻度と Box 毎 Calibration での T 平均値を示す。Citypersons は対象クラスが歩行者のみであり、固定カメラで撮影されていることからスケール、アスペクト比に偏りがあることが分かる。T 平均値はオブジェクト出現頻度の高い部分が大きくなっており、オブジェクト出現部分に対して Calibration されていることが分かる。図 5 にオブジェクト出現ヒートマップと T 分布を示す。位置性に関してもオブジェクト出現部分と T の関連が見られる。T が Single の場合、Default Box 全体に共通な作用してしまうため、最適な Calibration できないと考えられる。よって、T を Default Box 毎に配置することでオブジェクトの出現傾向を考慮することができ、高精度な

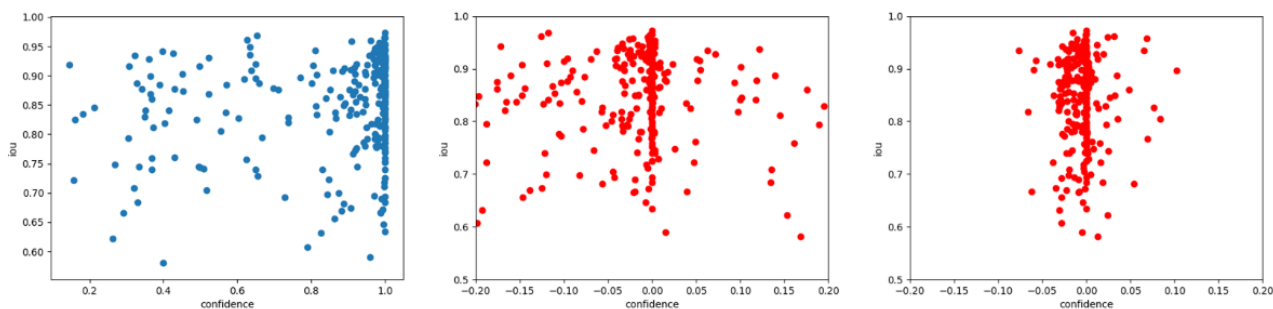


図 6 IoU と Confidence

(左) Calibration 前. (中央) Box Calibration 時の Confidence 差分. (右) 重み付き有無の Confidence 差分 (重み有 - 重み無)

Confidence Calibration が実現できる.

メータ T の導入などが考えられる.

4.5 検出精度評価

位置精度に対する Confidence 特性を評価するため、検出領域と正解領域の対応付けするための IoU 閾値を変えながら Average Precision(AP) を算出する. 位置精度に対する Confidence が改善されていれば、IoU 閾値が高い場合において AP が改善されていることが期待できる. 表 2 に IoU 閾値別の AP 結果を示す. 位置精度の重み付き学習を実施することにより、IoU 閾値が高い場合において AP が改善されていることが分かる. また、 T が Single の場合重み付き学習の有無で改善が見られない. これは候補領域のスケール、アスペクト比に対する T の表現力が不足しているため、重み付き学習が機能しづらいと考えられる. 図 6 に Confidence Calibration 前後での IoU に対する Confidence 変化と重み付き学習有無での Confidence 変化を示す. 通常の Confidence Calibration を実施すると IoU の高い Box も Confidence 低下するに対して、重み付き学習を実施することで IoU が高い Box の Confidence 低下が抑制されていることが示されている. よって、温度パラメータを Default Box 毎に設定し位置精度による重み付き学習を実施することで、位置精度に対する Confidence 特性を改善することができる.

5. まとめ

本研究は、Object Detection における検出位置精度を考慮した Confidence Calibration 手法として、1-stage 型 Detector の Default Box 毎に温度パラメータ T を設定し、位置精度に応じた重み付き学習で最適化する手法を提案した. 一般交通シーンにおける歩行者検出タスクにて、Confidence Calibration 性能と位置精度に対する Confidence 特性改善による検出精度向上を評価し、提案手法の有効性を検証した.

今後の検証として、2-stage 型 Detector での検討や更なる Confidence Calibration 性能の向上として領域別の温度パラ

参考文献

- [1] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.
- [2] Guo, Chuan, et al. "On calibration of modern neural networks." In *Proceedings of the 34th International Conference on Machine Learning* 2017.
- [3] Pereyra, Gabriel, et al. "Regularizing neural networks by penalizing confident output distributions." *arXiv preprint arXiv:1701.06548* (2017).
- [4] Seo, Seonguk, et al. "Learning for single-shot confidence calibration in deep neural networks through stochastic inferences." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
- [5] Neumann, Lukas, et al. "Relaxed Softmax: Efficient Confidence Auto-Calibration for Safe Pedestrian Detection." In *Proceedings of the Advances in Neural Information Processing Systems Workshop*. 2018.
- [6] Ji, Byeongmoon, et al. "Bin-wise Temperature Scaling (BTS): Improvement in Confidence Calibration Performance through Simple Scaling Techniques." In *Proceedings of the IEEE International Conference on Computer Vision Workshop*. 2019.
- [7] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.
- [8] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." In *Proceedings of the Advances in neural information processing systems*. 2015.
- [9] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [10] Liu, Wei, et al. "Ssd: Single shot multibox detector." In *Proceedings of the European conference on computer vision*. Springer, Cham, 2016.
- [11] Fu, Cheng-Yang, et al. "Dssd: Deconvolutional single shot detector." *arXiv preprint arXiv:1701.06659* (2017).
- [12] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." In *Proceedings of the IEEE international conference on computer vision*. 2017.
- [13] Zhang, Shifeng, et al. "Single-shot refinement neural network for object detection." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [14] Zhao, Qijie, et al. "M2det: A single-shot object detector based on multi-level feature pyramid network." In *Proceedings of the AAAI*

- Conference on Artificial Intelligence. Vol. 33. 2019.
- [15] Zhang, Shanshan, et al. "Citypersons: A diverse dataset for pedestrian detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
 - [16] Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding." In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
 - [17] Simonyan, Karen, et al. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
 - [18] Kingma, Diederik P. et al. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
 - [19] Naeini, Mahdi Pakdaman, et al. "Obtaining well calibrated probabilities using bayesian binning." In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.
 - [20] Platt, John. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." Proceedings of the Advances in large margin classifiers 10.3 (1999): 61-74.