アンサンブル推論の効果の分析と Attention 相違度を用いたアンサンブル学習の提案

ER16013 岡本直樹　　指導教授：藤吉弘亘

1. はじめに

複数のニューラルネットワークの出力を統合して推論を行うアンサンブル推論は，単一のネットワークによる推論に比べ精度が向上することが確認されている．これは，各ネットワークが学習により獲得した知識に差が生じ，各ネットワークの出力分布に違いが表れるためと考えられる．本研究では，ネットワーク間の知識の差に着目し，アンサンブル推論による精度向上の要因解明を目指す．本研究では，知識の差としてネットワーク間における事後確率の差に注目して分析する．また，分析結果に基づいて，Attention 相違度を用いたアンサンブル学習を提案する．

2. アンサンブル推論

深層学習におけるアンサンブルでは，式 (1) のように複数のネットワークを利用した推論を行う．

$$\boldsymbol{y} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{p}_m(\boldsymbol{x}) \tag{1}$$

ここで，M はネットワーク数，$\boldsymbol{p}_m$ はネットワークの事後確率，$\boldsymbol{x}$ は入力データを表す．複数のネットワークの事後確率を平均した結果から推論を行うことで，単一のネットワークによる推論に比べ精度が向上する．

3. アンサンブルによる精度向上の分析

ニューラルネットワークにおけるアンサンブルによる精度向上の要因は十分に解明されていない．本研究ではアンサンブルに使用するネットワーク間の差に着目し，事後確率の差を表す KL-divergence を指標としてアンサンブルの分析を行う．

3.1. KL-divergence

2 つの確率分布の相違度を測る尺度として KL-divergence $D_{KL}(\boldsymbol{p}_1(\boldsymbol{x}) \parallel \boldsymbol{p}_2(\boldsymbol{x}))$ がある．ここで，$\boldsymbol{p}_1$ と $\boldsymbol{p}_2$ はネットワーク 1 と 2 の事後確率，$\boldsymbol{x}$ は入力データを表す．KL-divergence は距離の公理を満たす尺度ではないため，分析では式 (2) のような双方向の KL-divergence を利用する．

$$(D_{KL}(\boldsymbol{p}_1(\boldsymbol{x}) \parallel \boldsymbol{p}_2(\boldsymbol{x})) + D_{KL}(\boldsymbol{p}_2(\boldsymbol{x}) \parallel \boldsymbol{p}_1(\boldsymbol{x})))/2 \tag{2}$$

3.2. 分析結果

分析には，CIFAR100 データセットを用いて学習した 100 個の学習済みネットワーク (ResNet32) を対象とする．100 個のネットワークの中から 2 つを選択し，アンサンブルによる推論を行う．このとき，2 つのネットワークの平均精度とアンサンブル精度の差をアンサンブルによる効果とする．アンサンブルの全組み合わせ 4950 組において，テストデータに対する KL-divergence とアンサンブル効果の傾向を調査する．

KL-divergence による評価を図 1 に示す．相関係数から事後確率の相違度とアンサンブル効果に正の相関があることがわかる．以上により，事後確率の差がアンサンブル効果を高める要因であるといえる．

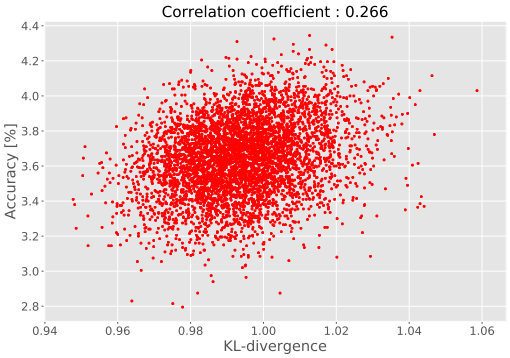

図 1：KL-divergence による評価

4. Attention 相違度を用いたアンサンブル学習

分析の傾向から，アンサンブル効果を高めるために，ネットワーク間の事後確率を離すように学習を行う．事後確率を直接離した場合，入力画像とは関係のない確率値が大きくなることが懸念される．そこで，事後確率を離すために，全結合層の前層の特徴マップから得られる Attention map が一致しないように学習する．

4.1. Attention 相違度

中間層の特徴マップを用いて学習を行う手法として Attention Transfer[1] が提案されている．Attention Transfer は，特徴マップから Attention map を作成し，ネットワーク間における Attention map の差を学習に利用する．Attention map は，式 (3) のようにチャネル数分ある特徴マップから各特徴マップの同位置の要素を合成することで生成される．

$$F^{p}_{\mathrm{sum}}(\boldsymbol{A}) = \sum_{i=1}^{C} \mid \boldsymbol{A}_i \mid^{p} \tag{3}$$

ここで，C はチャネル数，$\boldsymbol{A}_i$ は特徴マップ，p はノルム数を表す．Attention map の差は，式 (4) のようにネットワーク 1 とネットワーク 2 の正規化した Attention map のベクトルを用いて算出する．本研究では，この差を Attention 相違度と呼ぶ．

$$L_{AT} = \parallel \frac{\boldsymbol{Q}^j_1}{\parallel \boldsymbol{Q}^j_1 \parallel_2} - \frac{\boldsymbol{Q}^j_2}{\parallel \boldsymbol{Q}^j_2 \parallel_2} \parallel_p \tag{4}$$

ここで，j は中間層の位置，$\boldsymbol{Q}^j_1$ はベクトル化したネットワーク 1 の Attention map，$\boldsymbol{Q}^j_2$ はベクトル化したネットワーク 2 の Attention map を表す．

4.2. 実験概要

Attention 相違度を用いた学習である Attention と通常の学習である Vanilla を行い，2 つのネットワークを用いたアンサンブルにおけるアンサンブル効果を比較する．ネットワークは ResNet32，データセットは CIFAR100 を使用する．Attention は，負の Attention 相違度を損失関数へ追加することで，31 層目の特徴マップから生成した Attention map をお互いに離すように学習を行う．Vanilla は，2 つのネットワークを個別に学習を行う．同じ初期値を持つネットワークにおいて Attention と Vanilla による学習を行い，アンサンブル効果を比較することで，負の Attention 相違度の導入によるアンサンブル効果への影響を確認する．

4.3. 実験結果

評価結果を表 1 に示す．各値は 130 組のアンサンブルにおけるアンサンブル効果の大きい組み合わせの数を表す．同じ初期値を持つネットワークにおいて，Attention map を離すことで Vanilla より高いアンサンブル効果を発揮したアンサンブルが 76 組あることがわかる．以上のことから Attention map を離すことで，アンサンブル効果が高くなる傾向にあると考える．

表 1：アンサンブル効果の比較

	Attention>Vanilla	Attention<Vanilla
組み合わせ数	76	54

5. おわりに

本稿では，ネットワーク間の差異に着目した分析から得た傾向に基づき，アンサンブル学習を行った．今後は事後確率の相違度に着目した分析を行う．

参考文献

[1] S. Zagoruyko, *et al.*, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer”, ICLR, 2017.