Training very deep network

著者名 : Rupesh Kumar Srivastava, Klaus Greff, Jürgen Schmidhuber

年 : 2015

会議 : Neural and Evolutionary Computing

氏名 : ZHAO XIN

学籍番号 : 1811336

1. Motivation

深層学習は教師ありタスクの正確率を著しく向上することができます。その前の研究により、Imagenetの画像分類の問題に対して、より狭い受容野を利用しても、ディープネットワークの正確率は浅いネットワークより高いです。正確率を84%から95%まで改善しました。

そして、ディープネットワークは浅いネットワークより強い表現力を持つ原因を考えましょう。画像にしても、自然言語にしても、人間がこいったものを認識し、理解するためには、階層的 (hierarchical order) な知識構造と時系列的な入力が必要です。浅いネットワークはこのような知識を表すことができますが、効率は低いので、ディープネットワークで知識を獲得するのは重要です。

ですが、Vanishing Gradient Problem と Cliffs and Exploding Gradientsなどの問題が存在するため、ディープネットワークをトレーニングすることは簡単ではありません。本論文は、ディープネットワークの深い階層で、勾配値の poor propagation の問題を着目しています。LSTM構造に触発されて、本論文はの「adaptive gate」を活用し、情報が深いネットワークで流しても弱くならない構造「highway network」を提案しました。

2. Method

この構造はタスクに限定していなく、deep modelを利用できる任務は全部利用できます。本論文では、MNISTとCIFAR-100のアスクで評価を行っています。

以下は普通のdeep network構造と「highway network」を比較しながら、紹介します。

ここでは普通の deep network の layer 構造を「plain layer」と呼びます。「highway network」の layer 構造を 「block layer」と呼びます。

ネットワークのある層の入力を x とし、出力を y とします。「·」はアダマールを意味します。

- **「plain layer」の場合は**

  $y = H(x, W_h)$

- **「block layer」の場合は**

  $y = H(x, W_h) \cdot T(x, W_t) + x \cdot C(x, W_c)$

  $C(x, W_c) = 1 - T(x, W_t)$

  ここで、$T(x, W_t) = \sigma(x * W_t + B_t)$ 。すなわち、$T(x, W_t)$ の値域は $(0, 1)$。

  $T(x, W_t)$ が大きくなれば、この層状態をより多く下の層に伝搬できます。

  $T(x, W_t) = 0$ の場合で、$y = H(x, W_h) \cdot 0 + x \cdot 1 = x$. すなわち、この層の入力を直ちに次の層に伝搬します。

> この論文で,$H(x, W_h) \cdot T(x, W_t)$の値をtransform gate activityと呼んでいます。

論文の実験によって、この構造を利用することで、1000層以上のディープネットワークに対しても、SGDによるトレーニングが可能です。

3. Utility

3.1 論文の知見

論文の「analysis」の部分で、各層の出力ベクトルと各層の transform gate activity のベクトルを次元ごとに比較することで、highway network の内部状態を可視化しました。

興味深い点は、簡単なタスク(MNIST)に対して、層が浅いうちは層の出力ベクトルの値は層ごとに変化していきますが、ある層から顕著な変化は見られなくなり、受け取った情報を反映させることなく、後ろに流していくという傾向が見られます。すなわち、情報を highway に載せて、深い層で処理することなく、結果に出力します。これによって、多くの層での勾配降下処理を省略したため、Vanishing Gradient Problem などの問題を解決できました。これにより、SGDを利用できます。

そして、ネットワークの階層を増加することで、パラメータの量を減少することもできます。これにより、モデルの性能を上げてきました。

3.2 応用できる場面

深いネットワークを利用する場面で有効です。

たとえば、2018年の論文「Deep contextualized word representations」では、文字に基づくCNNを利用しています。このCNNは2層の highway 付きの layer を使っています。