

部分データ集合に基づくパラメータ初期化による ニューラルネットワークの学習安定化

Progressive Data Increasing as Initialization of Neural Networks

佐藤利鷹理^{*1}
Ryosuke Sato

木村正成^{*2}
Masanari Kimura

^{*1}株式会社 Ridge-i
Ridge-i inc.

^{*2}株式会社 Ridge-i
Ridge-i inc.

Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on a variety of tasks. With the improvement of the hardware performance, the structure of the neural network becomes more and more complicated, and the amounts of data used for training networks are becoming larger. Although these changes greatly contributed to the recognition accuracy of the neural network, it often leads to training instability in some cases. In this paper, we propose to use a subset of training data for initialization of neural networks. We found that it can help us to stabilize the training process.

1. はじめに

近年、ディープニューラルネットワークは様々なタスクにおいて大きな成功を収めている [1, 2]. こうした成功の多くは、学習データの大規模化とネットワークの複雑化によってもたらされている. しかし, こうしたネットワーク構造と学習データの複雑化によって, ニューラルネットワークの学習が不安定になってしまうケースが多く存在する.

我々は, ネットワークの重みの初期値問題に着目することでこうした問題を解決することを目指す. 本論文における提案手法は, 学習データ全体の小規模な部分データ集合を用いて, ネットワークの重みをタスクやドメインにとって適切に初期化を行うことで, 学習の安定化を達成する. 実験結果から, 提案手法がニューラルネットワークにとって望ましい重みの初期値を導出できることを示す.

2. ニューラルネットワークの重み初期化

我々は, 大規模データにおける DNNs の学習の不安定性を, ネットワークの重みの初期値問題の視点から考える. DNNs の学習の安定性および最終的な認識性能はネットワークの重みの初期値に依存する. こうした DNNs の初期化に関する研究はいくつか存在する [3, 4, 5].

2.1 Kaiming Initialization

特に Kaiming He ら [5] が提案した初期化手法は, 現在最も用いられている手法のうちの 1 つである.

ネットワークの順伝播において, 第 i 層の出力は,

$$y^i = W^i x^i + b^i \quad (1)$$

と表現できる. $x_i \in \mathbb{R}^{n_i}$ は i 番目の入力, $W_i \in \mathbb{R}^{d,n}$ は重みフィルタである. 活性化関数 $f(\cdot)$ の出力は,

$$x^{i+1} = f(W^i x^i + b^i) \quad (2)$$

$$= f(y^i) \quad (3)$$

ここで, He らはネットワークの初期値について以下のような仮定を置いている.

連絡先: 佐藤利鷹理, 株式会社 Ridge-i, 東京都千代田区大手町 1-6-1 大手町ビル 442, rsato@ridge-i.com

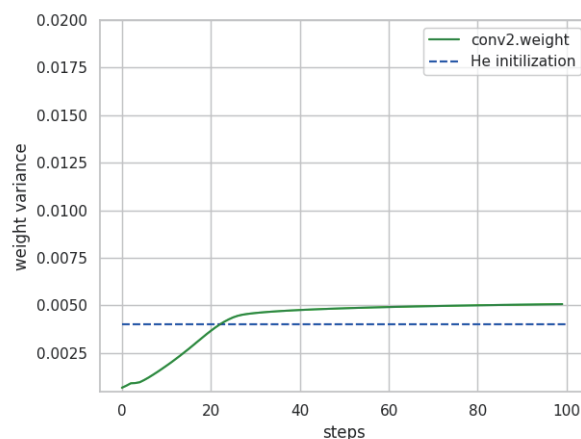


図 1: 部分データ集合を用いたネットワークの重み初期化によって得られる重みの分散. CIFAR10 データセットから 0.1% の部分集合を抽出して, 3 層の畳み込みニューラルネットワークについて初期化を行い, 2 層目の畳み込み層の重みの分散を図示している.

- W_i と x_i の各要素の初期値はそれぞれ独立であり, 同じ分布からサンプリングされる
- W_i と x_i は互いに独立である

これに基づき, He らは第 i レイヤーの重みを 平均 0, 分散 $2/n_i$ の正規分布を用いて初期化している.

$$\text{Var}[W^i] = \frac{2}{n_i} \quad (4)$$

この分布を用いた重みパラメータの初期化は, 従来の初期化手法と比べてより良い実験結果を得ることができている.

2.2 Progressive Data Increasing as the Initializer

与えられた N 件の学習データを $D^{d \times N} = \{(x_i, y_i)\}_{i=1}^N$ とし, x をデータ点, y をラベル, d をデータの次元数とする. 一般的な機械学習のタスクでは, $D^{d \times N}$ を用いて任意の損失関数 L を最小化するようにモデルを学習させる. しかし, デー

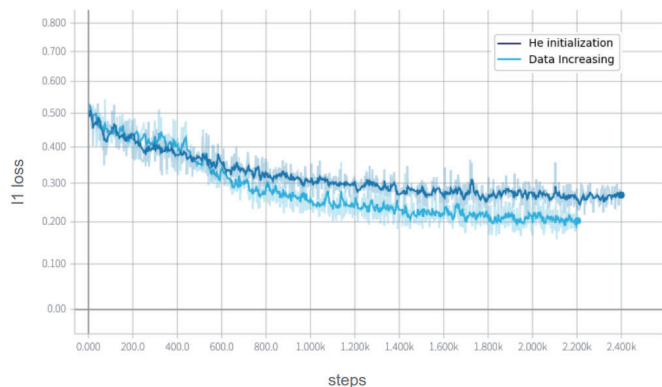


図 2: He Initialization と提案手法の VAE による画像生成の比較実験の結果。提案手法では、CIFAR10 データセットから 0.1% の部分集合を抽出して重みの初期化を行った後、データセット全体を用いて学習させている。

タの次元数 d が大きい場合やネットワーク構造が複雑である場合、学習が不安定になってしまうケースが多く存在する。

我々は、こうした観測がネットワークの重みの初期値が不適切なまま大量のデータで学習を試みていることに起因すると仮定し、学習データの小規模な部分集合を用いて重みを擬似的に初期化する Progressive Data Increasing を提案する。重みの初期化に用いる部分データ集合 S は、

$$S^{d \times M} = \{(\mathbf{x}_i, \mathbf{y}_i) \sim P(\mathbf{x}, \mathbf{y}), \forall (\mathbf{x}_i, \mathbf{y}_i) \in D\}_{i=1}^M \quad (5)$$

と表現できる。ここで M は初期化に用いる部分データ集合のサイズ、 P は任意の分布とする。 $S^{d \times M}$ によって学習されるモデル f は、

$$f_{t^0} = \arg \min_t L(S, f, t) \quad (6)$$

であり、 t^0 は初期化によって得られるモデルのパラメータである。我々はこの t^0 に対するいくつかの分析を行い、 t^0 がニューラルネットワークの重みの初期値として望ましい性質を満たすことを示す。

3. 実験結果

図 2.1 に、提案手法で得られる初期値の分散を図示する。図から、部分データ集合を用いた初期化で得られる重みの分散は、He が導出した、ネットワークにとって望ましい重みの分散とほぼ一致することがわかる。提案手法を用いることで、He らの初期化と同様に、重みの分散は $2/n_i$ の周辺に収束し、学習の安定性を損なわないような初期値を得られている。実験では CIFAR10 データセット *1 を用いている。図では 2 層目の重みをプロットしているが、それ以外の層でも類似した実験結果が得られた。

また、図 2.2 に He らの初期化手法と提案手法を適用した Variational Auto Encoder (VAE) の学習曲線の比較を示す。表 1 は、この実験で得られた最終的なスコアの比較である。実験に用いた VAE のネットワーク構造は 162 層からなる深い Resnet 構造 [6] を用いている。また、生成画像のサイズは 1024×1024 、データセットは CIFAR10 を対象としている。比

表 1: He Initialization と提案手法の VAE による画像生成の比較実験の定量評価。

Initializers	L1 Loss
Kaiming He	0.2418
Progressive Data Increasing	0.1860

較から、複雑なネットワークアーキテクチャによる学習では、提案手法を用いることで、He らの初期化手法と比べてより安定した学習ができることを示している。

4. おわりに

本論文では、学習データセットの部分集合を利用してニューラルネットワークの重みの初期化を擬似的に行い、学習の安定性を向上させる手法を提案した。提案手法はどのようなネットワーク構造、データセットおよびタスクにおいても汎用的に適用可能である。また、実験結果から提案手法が学習の安定化に対して有効であることを示した。

今後の研究課題としては、初期化に用いる部分データ集合の選択を、より効果的に行うことなどが考えられる。現在は与えられた学習データの中から無作為に小規模な部分集合を抽出し、それをネットワークの初期化に用いている。この部分集合をより学習の安定化に寄与するように選択することができれば、さらなる学習の安定化およびネットワークの汎化性能向上につながり得る。

参考文献

- [1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [2] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, Vol. 12, No. Aug, pp. 2493–2537, 2011.
- [3] Dmytro Mishkin and Jiri Matas. All you need is a good init. *arXiv preprint arXiv:1511.06422*, 2015.
- [4] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

*1 <https://www.cs.toronto.edu/~kriz/cifar.html>