Градиентная маска: как механизм латерального торможения улучшает работу искусственных нейронных сетей

Л. Цзян[1]

В настоящей работе мы предлагаем *градиентную маску*, которая отфильтровывает градиенты шума в процессе обратного распространения. Такое обучение позволяет увеличивать плотность и амплитуду представления сигналов в сети. В работе мы представляем новую меру качества градиента. Мы демонстрируем аналитическими методами, как латеральное торможение в искусственных нейронных сетях улучшает качество распространяемых градиентов. Наконец, проводим несколько различных экспериментов, чтобы изучить, как *градиентная маска* улучшает количественную и качественную производительность сети.

Ключевые слова: латеральное торможение, градиентная маска, свёрточные нейронные сети.

1. Введение

Во время обратного распространения ошибки градиенты генерируются для всех признаков, а затем участвуют в обновлении соответствующих весов. В работе [1] показано, что не все градиенты важны для обучения.

Наша модель придает величину градиентам признаков каждого сверточного слоя с помощью оператора Лапласа Гаусса (LoG), который имеет распределение «Мексиканской шляпы», сходное с модуляцией внимания в биологическом мозге. Во время обратного распространения ошибки LoG подавляет часть менее важных градиентов.

Чтобы обосновать использование латерального торможения мы предлагаем новый критерий для измерения важности каждого признака в зависимости от градиента матрицы признаков.

1.1. Градиентная маска

Для сверточного слоя $l \in \mathbb{R}^{u,v}$, мы равномерно делим матрицу признаков на K непересекающихся матриц — *наборов признаков*. Для каждого

[1] *Цзян Лэй* — аспирант каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: kiwee@outlook.com

Jiang Lei — graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

набора признаков градиенты с одной и той же координатой составляются в вектор, называемый *миниколонкой*, а миниколонку в координатах (i, j) k-го множества мы обозначим как $M_{ij}^l(k)$. Далее вычисляем норму $||M_{ij}^l(k)||_2$, чтобы представить величину градиентов в этой миниколонке. Затем для каждого k мы применяем оператор LoG к матрице, составленной из $||M_{ij}^l(k)||_2$ для всех $0 \leq i \leq u, 0 \leq j \leq v$. Этот процесс выполняется с помощью свертки с ядром:

$$LoG(i,j) = \frac{\partial^2 G_\sigma(i,j)}{\partial i^2} + \frac{\partial^2 G_\sigma(i,j)}{\partial j^2} = -\frac{1}{\pi\sigma^4}[1 - \frac{i^2+j^2}{2\sigma^2}]e^{-\frac{i^2+j^2}{2\sigma^2}} \tag{1}$$

где (i, j) — координаты ядра свертки LoG, а $G_\sigma(i, j)$ — это гауссовская свертка со стандартным отклонением σ. Пусть $\delta_{ij}^l(k)$ обозначает результат свертки LoG на соответствующей части. Установив порог, равный ϵ, мы можем определить *градиентную маску*: $Mask^l(k) = [a_{ij}]_{u\times v}$, где:

$$a_{ij} = \begin{cases} 0 & , |\delta_{ij}^l(k)| < \epsilon \\ 1 & , |\delta_{ij}^l(k)| \geq \epsilon \end{cases} \tag{2}$$

Поскольку маска градиента $Mask^l(k)$ соответствует k-му набору матриц признаков, фильтры, соответствующие этому набору, используют одну и ту же маску $Mask^l(k)$. Во время обратного распространения каждый градиент проходит через градиентную маску. Пусть L обозначает функцию потерь, тогда градиент веса w_{mn}^l на фильтре может быть записан как:

$$\frac{\partial L}{\partial w_{mn}^l} = \underbrace{\sum_{(i,j),a_{ij}=0} \frac{\partial L}{\partial a_{ij}^l}\frac{\partial a_{ij}^l}{\partial w_{mn}^l}}_{=0} + \underbrace{\sum_{(i',j'),a_{i'j'}\neq 0} \frac{\partial L}{\partial a_{i'j'}^l}\frac{\partial a_{i'j'}^l}{\partial w_{mn}^l}}_{\neq 0} \tag{3}$$

Градиенты несущественных признаков приравниваются к 0, что помогает при обучении снизить шум в весах признаков. Как экспериментально показано далее, это улучшает способность модели к обобщению.

1.2. Чувствительность градиентного потока

Будем рассматривать градиентную маску с аналитической точки зрения. Пусть $g_{mn}^l = \frac{\partial L}{\partial w_{mn}^l}$ обозначает градиент веса на l-м сверточном слое, где L — функция потерь, а (x, y) — координаты в матрице признаков этого слоя. В данном разделе мы будем рассматривать только сети, использующие функцию активации ReLU.

Мы определяем *чувствительность градиентного потока* s_{mn}^l как модуль лапласиана градиента в двумерном пространстве:

$$s_{mn}^l := |\Delta g_{mn}^l| = |\frac{\partial^2 g_{mn}^l}{\partial x^2} + \frac{\partial^2 g_{mn}^l}{\partial y^2}| \tag{4}$$

Если мы рассматриваем лапласиан g^l_{mn} как дивергенцию векторного поля, то значение s^l_{mn} в определенной точке матрицы признаков указывает, до какой степени эта точка является «источником» градиента g^l_{mn}. Каждая точка может быть как положительным, так и отрицательным источником. Точки с более высокой чувствительностью к градиентному потоку более важны для градиента g^l_{mn}. При обратном распространении лапласиан градиента можно переписать следующим образом:

$$\Delta g^l_{mn} = \sum_{i,j} \Delta(\frac{\partial L}{\partial a^l_{ij}} \frac{\partial a^l_{ij}}{\partial w^l_{mn}}) = \sum_{i,j} \frac{\partial a^l_{ij}}{\partial w^l_{mn}} \Delta \frac{\partial L}{\partial a^l_{ij}} \tag{5}$$

Здесь $\frac{\partial a^l_{ij}}{\partial w^l_{mn}} \geq 0$ — производная от функции активации, заданная на входе нейрона, которую можно рассматривать как константу. Так как мы рассматриваем сеть с функцией активации ReLU, её производная неотрицательна. Рассматривая заданную точку (i', j') на матрице признаков, мы обозначим чувствительность градиентного потока в этой точке как $s^l_{mn}(i', j')$, а её лапласиан как $\Delta_{(i',j')} g^l_{mn}$, тогда мы имеем

$$\Delta_{(i',j')} g^l_{mn} = \frac{\partial a^l_{i'j'}}{\partial w^l_{mn}} \Delta \frac{\partial L}{\partial a^l_{i'j'}} \qquad s^l_{mn}(i', j') = \frac{\partial a^l_{i'j'}}{\partial w^l_{mn}} |\Delta \frac{\partial L}{\partial a^l_{i'j'}}| \tag{6}$$

Чувствительность градиентного потока в определенной точке на матрице признаков положительно коррелирует с лапласианом градиента признака в этой точке. Так как высокая чувствительность потока указывает на большую значимость для градиента, то и маскирование градиентного потока в точке с более низкими значениями LoG эквивалентно обнулению элементов, которые не важны для обновления веса.

2. Эксперименты

Были проведедены эксперименты по классификации на ImageNet (8 Tesla V100 GPUs) и CIFAR-100 (1 Tesla V100 GPU). Подробная информация о гиперпараметрах обучения: learning rate 0.1, epoch 120, optimizer SGD. Эксперименты показали, что использование градиентной маски улучшает точность классификации. В таблице 1 представлены результаты ResNet на двух наборов данных.

Чтобы выяснить, какие факторы наиболее значимы, были поставлены два эксперимента: в первом вместо латерального торможения используется L_2-норма градиентов внутри миниколонки; во втором латеральное торможение LoG используется для каждой ячейки матрицы без выделения миниколонок. Результаты на CIFAR-100 представлены в таблице 2, где мы видим, что LoG имеет решающее значение для генерации градиентной маски. Использование миниколонок также увеличивает точность.

Модель	CIFAR-100 (ResNet-18)	ImageNet (ResNet-50)
ResNet-18/50	78.2	75.5
ResNet-18/50 с ЛТ	**80.26**	**76.01**

Таблица 1. Точность (%) ResNet-18/50 с/без латеральным торможением (ЛТ) на CIFAR-100 и ImageNet

Модель	ЛТ на миниколонках	Без ЛТ	ЛТ без миниколонок
Accuracy	**80.26**	75.86	78.97

Таблица 2. Точность (%) ResNet-18 с/без латерального торможения (ЛТ) и миниколонок на CIFAR-100

Список литературы

[1] Lan Janice, Liu Rosanne, Zhou Hattie, Yosinski Jason, "Lca: Loss change allocation for neural network training", *Advances in Neural Information Processing Systems*, **32** (2019), 3619–3629.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Gradient Mask: Lateral Inhibition Mechanism Improves Performance in Artificial Neural Networks

Jiang Lei

In this paper we propose **Gradient Mask**, which helps the network to filtering out noisy or unimportant features while training. We propose a new criterion for gradient quality which can be used as a measure during training of various convolutional neural networks (CNNs). We demonstrate analytically how lateral inhibition in artificial neural networks improves the quality of propagated gradients. Finally, we conduct several different experiments to study how **Gradient Mask** improves the performance of the network both quantitatively and qualitatively.

Keywords: lateral inhibition, gradient masking, convolutional neural networks

References

[1] Lan Janice, Liu Rosanne, Zhou Hattie, Yosinski Jason, "Lca: Loss change allocation for neural network training", *Advances in Neural Information Processing Systems*, **32** (2019), 3619–3629.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.