Формулировка проблемы

Во время обучения нейронных сетей проблема слишком больших градиентов может проявляться как **взрыв градиентов** (gradient explosion). Это явление приводит к нестабильности обучения, когда обновления весов становятся слишком большими, что может повлиять на:

1. **Плохую сходимость** или её отсутствие.
2. Переполнение чисел при вычислениях (overflow).
3. Потерю эффективности в обновлении весов.

Мы рассмотрим, **как определить слишком большие градиенты** на этапе обучения нейронной сети, а также приведём **эмпирические правила** и способы диагностики.

Анализ задачи

1. Ключевая метрика — норма градиента

При каждой итерации обучения для набора весов $\mathbf{w}$ нейронной сети вычисляется градиент функции потерь L по отношению к $\mathbf{w}$:

$$\mathbf{g} = \nabla_{\mathbf{w}} L.$$

Для оценки величины градиента можно использовать его норму:

- **Евклидова норма (L2-норма):**

$$\|\mathbf{g}\|_2 = \sqrt{\sum_i g_i^2}.$$

- **Максимальная компонента (L∞-норма):**

$$\|\mathbf{g}\|_\infty = \max_i |g_i|.$$

Большие значения нормы указывают на слишком большой градиент.

---

2. Признаки слишком больших градиентов

Критерии, по которым можно диагностировать проблему:

1. **Резкие скачки значения функции потерь.**
   - Если в течение обучения значение функции потерь L резко возрастает (вместо убывания), это может быть признаком взрыва градиентов.
2. **Великая величина нормы градиента.**
   - Значения $\|\mathbf{g}\|_2$ или $\|\mathbf{g}\|_\infty$ становятся на несколько порядков больше обычных значений (например, 10^3, 10^4 или выше).
   - Это можно наблюдать при логировании или визуализации нормы градиента.
3. **Численные ошибки.**
   - Переполнение или генерация NaN (Not a Number) в параметрах весов или функции потерь.
4. **Плохие обновления весов.**
   - Если после применения градиентного шага веса изменяются слишком сильно (например, больше их начальных значений).

---

3. Эмпирические правила для определения слишком больших градиентов

Существует несколько практических подходов для диагностики проблемы:

1. **Мониторинг нормы градиента.**
   - Норма $\|\mathbf{g}\|_2$ должна оставаться в разумных пределах. Типичные значения зависят от модели, но обычно варьируются в диапазоне 10^{-2} до 10^1.
   - Если $\|\mathbf{g}\|_2 > 10^2$ (особенно в ранние итерации), это сигнал к проверке.
2. **Оценка изменения весов.**
   - Если относительное изменение весов на шаге обучения становится слишком большим:

$$\frac{\|\Delta \mathbf{w}\|_2}{\|\mathbf{w}\|_2} > \varepsilon,$$

   где $\varepsilon \approx 10^{-1}$ или 10^{-2}.
3. **Проверка производной функции потерь.**
   - Если производная функции потерь по весам имеет тенденцию резко возрастать на определённых итерациях, это явный признак взрыва градиентов.
4. **Анализ весов.**
   - Если веса становятся слишком большими (например, на порядки выше 1), это может быть результатом слишком больших градиентов.

Способы диагностики и рекомендации

A. Практические способы диагностики:

1. **Логирование нормы градиентов.** На каждой итерации логируйте значения $\|\mathbf{g}\|_2$ или $\|\mathbf{g}\|_\infty$. Визуализация таких значений может помочь выявить тенденции роста.
2. **Трассировка NaN.** Если веса, функция потерь или градиенты становятся NaN, это явный признак численной нестабильности.
3. **Визуализация потерь.** Построение графика значения L от итерации поможет увидеть неожиданные скачки.

B. Эмпирические пределы:

1. **Норма градиента:**

$$10^{-2} \leq \|\mathbf{g}\|_2 \leq 10^1.$$

   Если $\|\mathbf{g}\|_2 > 10^2$, это потенциальная проблема.
2. **Шаг обновления весов:**

$$\frac{\|\Delta \mathbf{w}\|_2}{\|\mathbf{w}\|_2} \leq 0.1.$$

C. Меры для предотвращения взрыва градиентов:

1. **Градиентный клиппинг:** Ограничение нормы градиента:

$$\mathbf{g} \leftarrow \mathbf{g} \cdot \min\left(1, \frac{c}{\|\mathbf{g}\|_2}\right),$$

   где c — предельное значение нормы (например, $c = 1.0$).
2. **Правильная инициализация весов:** Используйте подходящие методы, такие как He или Xavier initialization.
3. **Нормализация данных и входов.** Убедитесь, что входные данные имеют нулевое среднее и единичное стандартное отклонение.
4. **Использование малых шагов обучения.** Слишком большой шаг обучения (η) может усугублять взрыв градиентов.
5. **Использование устойчивых архитектур.** Например, использование LSTM вместо стандартных RNN для предотвращения взрыва градиентов в рекуррентных сетях.

Вывод

Тревогу о взрыве градиентов стоит бить, если:

- $\|\mathbf{g}\|_2$ значительно превышает значения 10^1-10^2.
- Относительное изменение весов ($\|\Delta \mathbf{w}\|_2 / \|\mathbf{w}\|_2$) становится чрезмерным.
- Наблюдаются резкие скачки функции потерь или генерация NaN.

Использование техник, таких как градиентный клиппинг и подходящая инициализация весов, помогает предотвратить эту проблему.