Введение в искусственный интеллект. Современное компьютерное зрение

Семинар 3. Несверточные слои

Бабин Д.Н., Иванов И.Е., Петюшко А.А.

кафедра Математической Теории Интеллектуальных Систем

9 марта 2021 г.

План семинара

1. Сведение к свертке

План семинара

1. Сведение к свертке
2. О сигмоиде

Пакетная нормализация при тесте

- Предположим, что мы используем пакет размера $T = 1$ (здесь и далее опустим этот индекс)

Пакетная нормализация при тесте

- Предположим, что мы используем пакет размера $T = 1$ (здесь и далее опустим этот индекс)
- Y_{ij}^{k} — трехмерный тензор значений для некоторого слоя, где

Пакетная нормализация при тесте

- Предположим, что мы используем пакет размера $T = 1$ (здесь и далее опустим этот индекс)
- Y_{ij}^k — трехмерный тензор значений для некоторого слоя, где
  - $1 \leq i \leq H, 1 \leq j \leq W$ — пространственные координаты (ширина и высота),
  - $k = 1 \dots K$ — номер карты признаков.

Пакетная нормализация при тесте

- Предположим, что мы используем пакет размера $T = 1$ (здесь и далее опустим этот индекс)
- Y_{ij}^k — трехмерный тензор значений для некоторого слоя, где
  - $1 \leq i \leq H, 1 \leq j \leq W$ — пространственные координаты (ширина и высота),
  - $k = 1 \dots K$ — номер карты признаков.
- Выход нормализованного слоя: $Z_{ij}^k = \gamma^k \frac{Y_{ij}^k - \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^k$

Пакетная нормализация как линейная операция от входа

- $Z_{ij}^k = \gamma^k \frac{Y_{ij}^k - \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^k$

Пакетная нормализация как линейная операция от входа

- $Z_{ij}^k = \gamma^k \frac{Y_{ij}^k - \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^k$
- Перепишем формулу в другом виде:

Пакетная нормализация как линейная операция от входа

- $Z_{ij}^{k} = \gamma^{k} \frac{Y_{ij}^{k} - \mu_{avg}^{k}}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^{k}$
- Перепишем формулу в другом виде:

$$Z_{ij}^{k} = Y_{ij}^{k} \frac{\gamma^{k}}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} - \frac{\gamma^{k} \mu_{avg}^{k}}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^{k}$$

Пакетная нормализация как линейная операция от входа

- $Z_{ij}^k = \gamma^k \frac{Y_{ij}^k - \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^k$
- Перепишем формулу в другом виде:

$$Z_{ij}^k = Y_{ij}^k \frac{\gamma^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} - \frac{\gamma^k \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^k$$

- Т.о., получаем $Z_{ij}^k = G^k Y_{ij}^k + g^k$, где

Пакетная нормализация как линейная операция от входа

- $Z_{ij}^k = \gamma^k \frac{Y_{ij}^k - \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^k$
- Перепишем формулу в другом виде:

$$Z_{ij}^k = Y_{ij}^k \frac{\gamma^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} - \frac{\gamma^k \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^k$$

- Т.о., получаем $Z_{ij}^k = G^k Y_{ij}^k + g^k$, где
  - Мультипликативный член $G^k = \frac{\gamma^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}}$,

Пакетная нормализация как линейная операция от входа

- $Z_{ij}^k = \gamma^k \frac{Y_{ij}^k - \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^k$
- Перепишем формулу в другом виде:

$$Z_{ij}^k = Y_{ij}^k \frac{\gamma^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} - \frac{\gamma^k \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}} + \beta^k$$

- Т.о., получаем $Z_{ij}^k = G^k Y_{ij}^k + g^k$, где
  - Мультипликативный член $G^k = \frac{\gamma^k}{\sqrt{\sigma_{avg}^{\mathbf{2}k} + \epsilon}}$,
  - Аддитивный член $g^k = \beta^k - \frac{\gamma^k \mu_{avg}^k}{\sqrt{\sigma_{avg}^{\mathbf{2}k} + \epsilon}}$.

Пакетная нормализация как свертка

- $Z_{ij}^k = G^k Y_{ij}^k + g^k$

Пакетная нормализация как свертка

- $Z_{ij}^k = G^k Y_{ij}^k + g^k$
- Значит, пакетная нормализация — это поканальная (depthwise, *см.* предыдущую лекцию) свертка с ядром размера 1×1!

Пакетная нормализация как свертка

- $Z_{ij}^k = G^k Y_{ij}^k + g^k$
- Значит, пакетная нормализация — это поканальная (depthwise, *см.* предыдущую лекцию) свертка с ядром размера 1×1!
- А композиция сверток — тоже свертка (*Указание*: предыдущее ДЗ)

Объединение свертки и пакетной нормализации (1)

- Обычно: сначала свертка, потом пакетная нормализация

Объединение свертки и пакетной нормализации (1)

- Обычно: сначала свертка, потом пакетная нормализация
- По слоям:

Объединение свертки и пакетной нормализации (1)

- Обычно: сначала свертка, потом пакетная нормализация
- По слоям:

$$X_{ij}^{m} \xrightarrow{F_{uv}^{mk}, b^k} Y_{ij}^{k} \xrightarrow{G^k, g^k} Z_{ij}^{k}$$

Объединение свертки и пакетной нормализации (1)

- Обычно: сначала свертка, потом пакетная нормализация
- По слоям:

$$X_{ij}^{m} \xrightarrow{F_{uv}^{mk}, b^k} Y_{ij}^{k} \xrightarrow{G^k, g^k} Z_{ij}^{k}$$

- Выписываем еще раз формулы для свертки:

Объединение свертки и пакетной нормализации (1)

- Обычно: сначала свертка, потом пакетная нормализация
- По слоям:

$$X_{ij}^m \xrightarrow{F_{uv}^{mk},b^k} Y_{ij}^k \xrightarrow{G^k,g^k} Z_{ij}^k$$

- Выписываем еще раз формулы для свертки:

$$Y_{ij}^k = \sum_{m=1}^{M} \sum_{u,v=1}^{p,q} X_{i+u-1,j+v-1}^m \cdot F_{uv}^{mk} + b^k, \quad \forall k = 1 \dots K$$

и для пакетной нормализации:

Объединение свертки и пакетной нормализации (1)

- Обычно: сначала свертка, потом пакетная нормализация
- По слоям:

$$X_{ij}^{m} \xrightarrow{F_{uv}^{mk}, b^k} Y_{ij}^{k} \xrightarrow{G^k, g^k} Z_{ij}^{k}$$

- Выписываем еще раз формулы для свертки:

$$Y_{ij}^{k} = \sum_{m=1}^{M} \sum_{u,v=1}^{p,q} X_{i+u-1,j+v-1}^{m} \cdot F_{uv}^{mk} + b^k, \quad \forall k = 1 \dots K$$

и для пакетной нормализации:

$$Z_{ij}^{k} = G^k Y_{ij}^{k} + g^k$$

.

Объединение свертки и пакетной нормализации (2)

- Объединяя, получим:

Объединение свертки и пакетной нормализации (2)

- Объединяя, получим:

$$Z_{ij}^{k} = \sum_{m=1}^{M} \sum_{u,v=1}^{p,q} X_{i+u-1,j+v-1}^{m} \cdot F_{uv}^{mk} \cdot G^{k} + b^{k} + g^{k}, \quad \forall k = 1 \dots K$$

- Т.о., мы получили свертку с параметрами O_{uv}^{mk}, o^{k}, где (подтягиваем параметры пакетной нормализации):

Объединение свертки и пакетной нормализации (2)

- Объединяя, получим:

$$Z_{ij}^k = \sum_{m=1}^{M} \sum_{u,v=1}^{p,q} X_{i+u-1,j+v-1}^m \cdot F_{uv}^{mk} \cdot G^k + b^k + g^k, \quad \forall k = 1 \dots K$$

- Т.о., мы получили свертку с параметрами O_{uv}^{mk}, o^k, где (подтягиваем параметры пакетной нормализации):
  - Ядро $O_{uv}^{mk} = F_{uv}^{mk} \cdot \frac{\gamma^k}{\sqrt{\sigma_{avg}^{\mathbf{2}k} + \epsilon}}$,

Объединение свертки и пакетной нормализации (2)

- Объединяя, получим:

$$Z_{ij}^k = \sum_{m=1}^{M} \sum_{u,v=1}^{p,q} X_{i+u-1,j+v-1}^m \cdot F_{uv}^{mk} \cdot G^k + b^k + g^k, \quad \forall k = 1 \dots K$$

- Т.о., мы получили свертку с параметрами O_{uv}^{mk}, o^k, где (подтягиваем параметры пакетной нормализации):
  - Ядро $O_{uv}^{mk} = F_{uv}^{mk} \cdot \frac{\gamma^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}}$,
  - Аддитивный член $o^k = b^k + \beta^k - \frac{\gamma^k \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}}$.

Объединение свертки и пакетной нормализации (2)

- Объединяя, получим:

$$Z_{ij}^k = \sum_{m=1}^{M} \sum_{u,v=1}^{p,q} X_{i+u-1,j+v-1}^m \cdot F_{uv}^{mk} \cdot G^k + b^k + g^k, \quad \forall k = 1 \dots K$$

- Т.о., мы получили свертку с параметрами O_{uv}^{mk}, o^k, где (подтягиваем параметры пакетной нормализации):
  - Ядро $O_{uv}^{mk} = F_{uv}^{mk} \cdot \frac{\gamma^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}}$,
  - Аддитивный член $o^k = b^k + \beta^k - \frac{\gamma^k \mu_{avg}^k}{\sqrt{\sigma_{avg}^{2k} + \epsilon}}$.
- Из $X_{ij}^m \xrightarrow{F_{uv}^{mk}, b^k} Y_{ij}^k \xrightarrow{G^k, g^k} Z_{ij}^k$ получили $X_{ij}^m \xrightarrow{O_{uv}^{mk}, o^k} Z_{ij}^k$.

Субдискретизация как свертка

- **Вопрос**: Можно ли maxpooling представить как свертку?

Субдискретизация как свертка

- **Вопрос**: Можно ли maxpooling представить как свертку?
- **Ответ**: Нет, так как операция взятия максимума — нелинейная (в то время как **свертка всегда линейна**)

Субдискретизация как свертка

- **Вопрос**: Можно ли maxpooling представить как свертку?
- **Ответ**: Нет, так как операция взятия максимума — нелинейная (в то время как **свертка всегда линейна**)
- **Вопрос**: Можно ли average pooling представить как свертку?

Субдискретизация как свертка

- **Вопрос**: Можно ли maxpooling представить как свертку?
- **Ответ**: Нет, так как операция взятия максимума — нелинейная (в то время как **свертка всегда линейна**)
- **Вопрос**: Можно ли average pooling представить как свертку?
- **Ответ**: Да, и рассмотрим на примере global average pooling (GAP):

Субдискретизация как свертка

- **Вопрос**: Можно ли maxpooling представить как свертку?
- **Ответ**: Нет, так как операция взятия максимума — нелинейная (в то время как **свертка всегда линейна**)
- **Вопрос**: Можно ли average pooling представить как свертку?
- **Ответ**: Да, и рассмотрим на примере global average pooling (GAP):
  - Пусть двухмерный (не обращаем внимание на карты) вход $X_{ij}, 1 \leq i \leq H, 1 \leq j \leq W$,

Субдискретизация как свертка

- **Вопрос**: Можно ли maxpooling представить как свертку?
- **Ответ**: Нет, так как операция взятия максимума — нелинейная (в то время как **свертка всегда линейна**)
- **Вопрос**: Можно ли average pooling представить как свертку?
- **Ответ**: Да, и рассмотрим на примере global average pooling (GAP):
  - Пусть двухмерный (не обращаем внимание на карты) вход $X_{ij}, 1 \leq i \leq H, 1 \leq j \leq W$,
  - $GAP2D(X) = \frac{1}{HW} \sum_{i,j=1}^{H,W} X_{ij}$,

Субдискретизация как свертка

- **Вопрос**: Можно ли maxpooling представить как свертку?
- **Ответ**: Нет, так как операция взятия максимума — нелинейная (в то время как **свертка всегда линейна**)
- **Вопрос**: Можно ли average pooling представить как свертку?
- **Ответ**: Да, и рассмотрим на примере global average pooling (GAP):
  - Пусть двухмерный (не обращаем внимание на карты) вход $X_{ij}, 1 \leq i \leq H, 1 \leq j \leq W$,
  - $GAP2D(X) = \frac{1}{HW} \sum_{i,j=1}^{H,W} X_{ij}$,
  - Тогда свертка, соответствующая $GAP2D(X)$ — это свертка с ядром $F_{GAP} = \frac{1}{HW} \mathbb{1}_{i,j=1}^{H,W}$ без аддитивного члена, с размером, как у входа $H \times W$, применяемая без добивки (паддинга) и в режиме "VALID"

О сигмоиде

- Вспомним три основных вида активации:

О сигмоиде

- Вспомним три основных вида активации:
  1. Сигмоида $\sigma(x) = \frac{1}{1+\exp(-x)}$,

О сигмоиде

- Вспомним три основных вида активации:
  1. Сигмоида $\sigma(x) = \frac{1}{1+\exp(-x)}$,
  2. Гиперболический тангенс $\tanh(x) = 2\sigma(2x) - 1$,

О сигмоиде

- Вспомним три основных вида активации:
  1. Сигмоида $\sigma(x) = \frac{1}{1+\exp(-x)}$,
  2. Гиперболический тангенс $\tanh(x) = 2\sigma(2x) - 1$,
  3. Rectified Linear Unit $ReLU(x) = \max(0, x)$.

О сигмоиде

- Вспомним три основных вида активации:
  1. Сигмоида $\sigma(x) = \frac{1}{1+\exp(-x)}$,
  2. Гиперболический тангенс $\tanh(x) = 2\sigma(2x) - 1$,
  3. Rectified Linear Unit $ReLU(x) = \max(0, x)$.
- Изначально все использовали $\sigma(x)$. Тем не менее, сейчас он почти не встречается. Почему?

О сигмоиде

- Вспомним три основных вида активации:
  1. Сигмоида $\sigma(x) = \frac{1}{1+\exp(-x)}$,
  2. Гиперболический тангенс $\tanh(x) = 2\sigma(2x) - 1$,
  3. Rectified Linear Unit $ReLU(x) = \max(0, x)$.
- Изначально все использовали $\sigma(x)$. Тем не менее, сейчас он почти не встречается. Почему?
- **Проблема**: выход $\sigma(x)$ — не центрирован в нуле.

О сигмоиде

- Вспомним три основных вида активации:
  1. Сигмоида $\sigma(x) = \frac{1}{1+\exp(-x)}$,
  2. Гиперболический тангенс $\tanh(x) = 2\sigma(2x) - 1$,
  3. Rectified Linear Unit $ReLU(x) = \max(0, x)$.
- Изначально все использовали $\sigma(x)$. Тем не менее, сейчас он почти не встречается. Почему?
- **Проблема**: выход $\sigma(x)$ — не центрирован в нуле.
- **Решение**: использовать $\tanh(x)$.

О сигмоиде

- Вспомним три основных вида активации:
  1. Сигмоида $\sigma(x) = \frac{1}{1+\exp(-x)}$,
  2. Гиперболический тангенс $\tanh(x) = 2\sigma(2x) - 1$,
  3. Rectified Linear Unit $ReLU(x) = \max(0, x)$.
- Изначально все использовали $\sigma(x)$. Тем не менее, сейчас он почти не встречается. Почему?
- **Проблема**: выход $\sigma(x)$ — не центрирован в нуле.
- **Решение**: использовать $\tanh(x)$.
- Однако это не избавляет от главной проблемы — **исчезающих градиентов**:
  1. Производная $\sigma'(x) = \sigma(x)(1 - \sigma(x))$,

О сигмоиде

- Вспомним три основных вида активации:
  1. Сигмоида $\sigma(x) = \frac{1}{1+\exp(-x)}$,
  2. Гиперболический тангенс $\tanh(x) = 2\sigma(2x) - 1$,
  3. Rectified Linear Unit $ReLU(x) = \max(0, x)$.
- Изначально все использовали $\sigma(x)$. Тем не менее, сейчас он почти не встречается. Почему?
- **Проблема**: выход $\sigma(x)$ — не центрирован в нуле.
- **Решение**: использовать $\tanh(x)$.
- Однако это не избавляет от главной проблемы — **исчезающих градиентов**:
  1. Производная $\sigma'(x) = \sigma(x)(1 - \sigma(x))$,
  2. Для любых больших по модулю x $\sigma(x)$ стремится к 1 или 0, и соответственно его производная — всегда к нулю.

- Вспомним три основных вида активации:
  1. Сигмоида $\sigma(x) = \frac{1}{1+\exp(-x)}$,
  2. Гиперболический тангенс $\tanh(x) = 2\sigma(2x) - 1$,
  3. Rectified Linear Unit $ReLU(x) = \max(0, x)$.
- Изначально все использовали $\sigma(x)$. Тем не менее, сейчас он почти не встречается. Почему?
- **Проблема**: выход $\sigma(x)$ — не центрирован в нуле.
- **Решение**: использовать $\tanh(x)$.
- Однако это не избавляет от главной проблемы — **исчезающих градиентов**:
  1. Производная $\sigma'(x) = \sigma(x)(1 - \sigma(x))$,
  2. Для любых больших по модулю x $\sigma(x)$ стремится к 1 или 0, и соответственно его производная — всегда к нулю.

О ReLU[1]

- $ReLU(x) = \max(0, x)$ дает нулевую производную только при отрицательных x,

[1] `https://stats.stackexchange.com/a/422579`

О ReLU[1]

- $ReLU(x) = \max(0, x)$ дает нулевую производную только при отрицательных x,
- $ReLU(x)$ при $x > 0$ дает константную производную (равную 1),

[1] https://stats.stackexchange.com/a/422579

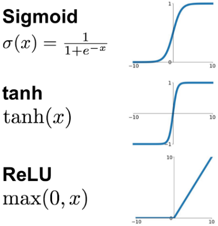
О ReLU[1]

- $ReLU(x) = \max(0, x)$ дает нулевую производную только при отрицательных x,
- $ReLU(x)$ при $x > 0$ дает константную производную (равную 1),
- $ReLU(x)$ потрясающе эффективен в реализации на конечном устройстве.

[1] https://stats.stackexchange.com/a/422579

О ReLU[1]

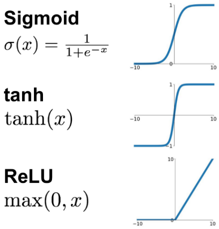
- $ReLU(x) = \max(0, x)$ дает нулевую производную только при отрицательных x,
- $ReLU(x)$ при $x > 0$ дает константную производную (равную 1),
- $ReLU(x)$ потрясающе эффективен в реализации на конечном устройстве.
- Иллюстрация:

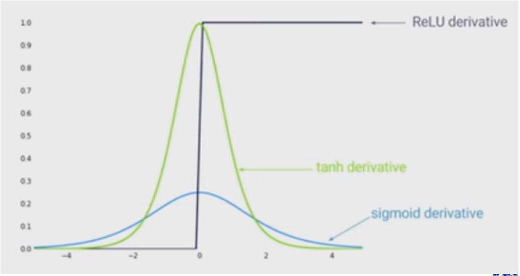


[1] https://stats.stackexchange.com/a/422579

О ReLU[1]

- $ReLU(x) = \max(0, x)$ дает нулевую производную только при отрицательных x,
- $ReLU(x)$ при $x > 0$ дает константную производную (равную 1),
- $ReLU(x)$ потрясающе эффективен в реализации на конечном устройстве.
- Иллюстрация:





[1] https://stats.stackexchange.com/a/422579

Спасибо за внимание!