Классификация и CNN

Маша Шеянова, masha.shejanova@gmail.com

Задачи NLP в МО

Как можно?

- на уровне текстов
- на уровне предложений
- на уровне слов
- speech, OCR, image captioning

На уровне текстов

Классификация текста:

- spam detection
- жанры
- sentiment analysis (тональность)
- предсказание темы

Кластеризация текстов:

- новости
- topic modelling

Предложения

- Paraphrase
- Textual entailment
- QA systems
- machine translation

Слова

- POS-tagging
- named entity recognition

(и то и другое решается как задача sequence tagging)

Dropout

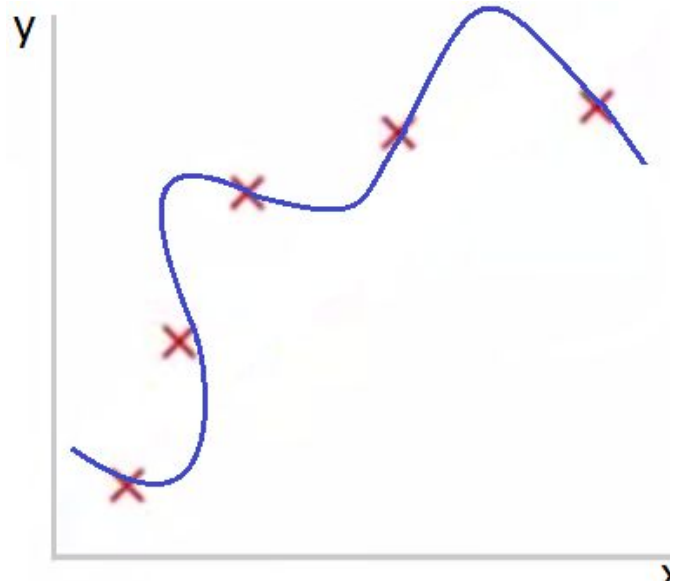
Что такое переобучение?

Переобучение означает, что модель обращает слишком много внимания на незначительные признаки, чтобы “слишком хорошо” подстроиться под данные на обучающей выборке.

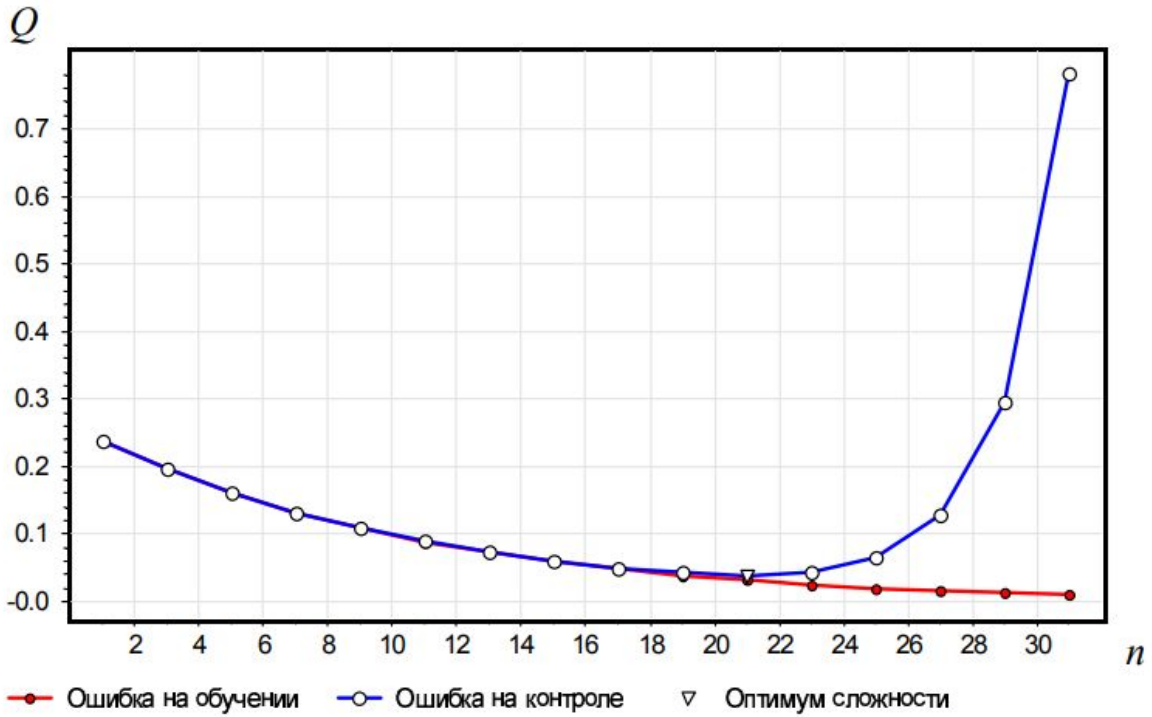
При этом модель становится излишне сложной (посмотрите на эти ненужные загогулины).

Скорее всего, она покажет плохой результат на тестовой выборке.

Как понять, что модель переобучилась?



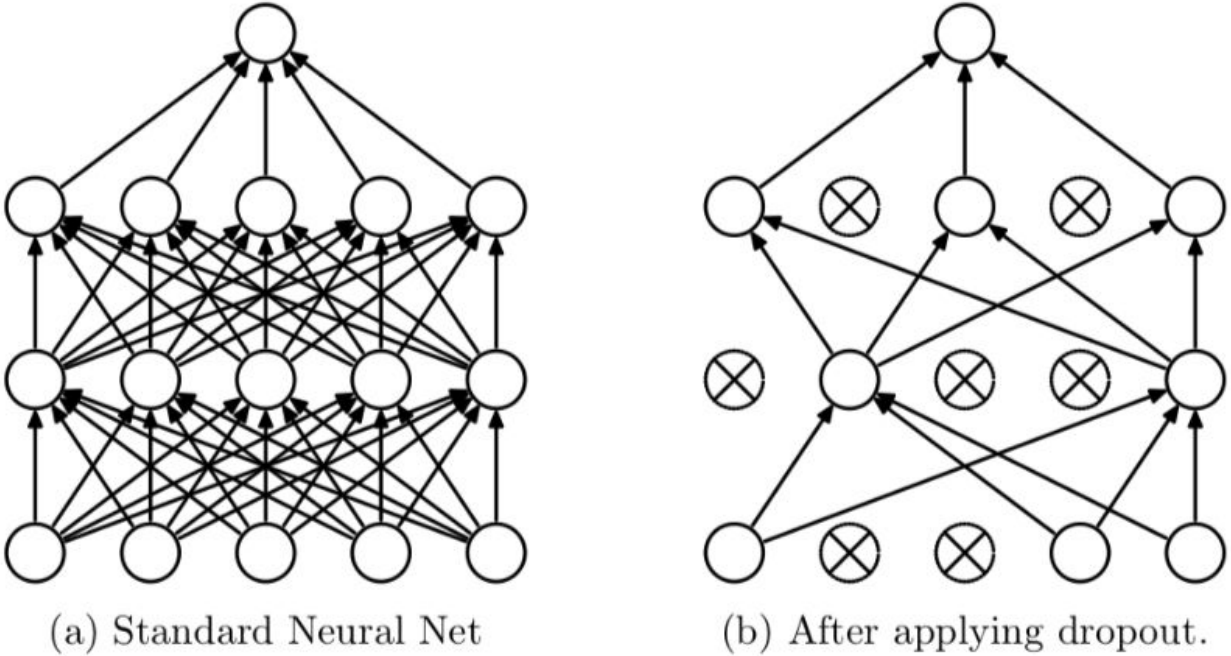
Переобучение — это когда



Источник картинки.

Иными словами, если качество модели на обучении продолжает расти, а на тесте — падает, модель переобучилась.

Dropout

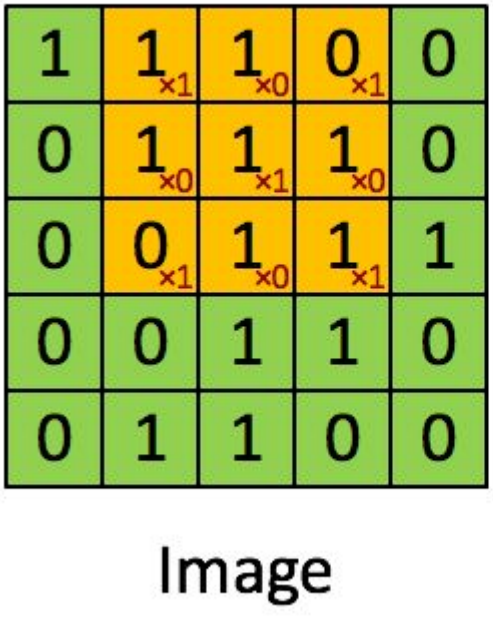


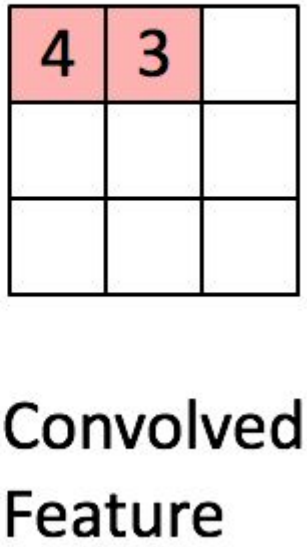
(a) Standard Neural Net

(b) After applying dropout.

CNN

Свёртка (convolution)





- проходимся по матрице скользящим окном (его называют ядро, kernel)
- перемножаем каждое попавшее в окно число матрицы на числа ядра
- складываем результат и записываем в ячейку новой матрицы

анимация

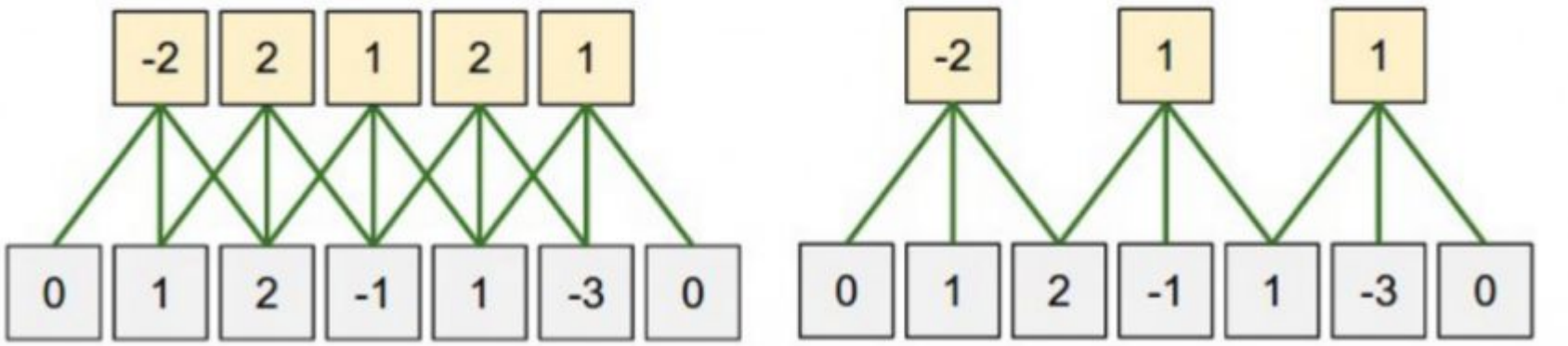
CNN

- идея за свёрткой — из сырых фичей извлечь полезные признаки
- вообще говоря, числа в ядре могут быть фиксированными
  - например, все единички
- однако, тогда придётся заранее их выбирать
- универсальный оптимальный вариант выбрать нельзя
- но можно обучить нейросеть подбирать цифры в ядре самостоятельно!

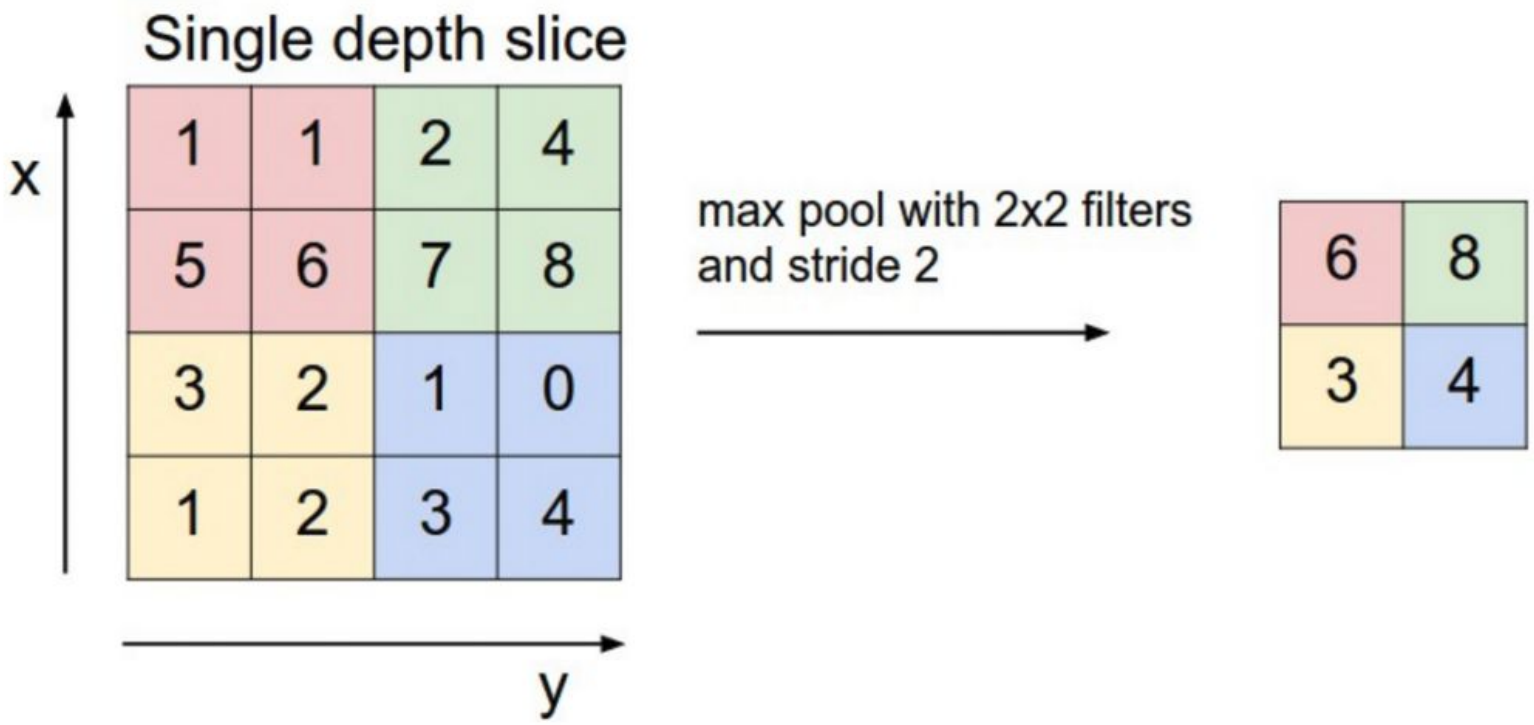
Stride

Можно проходиться, как на картинке выше, каждый раз сдвигаясь на 1.

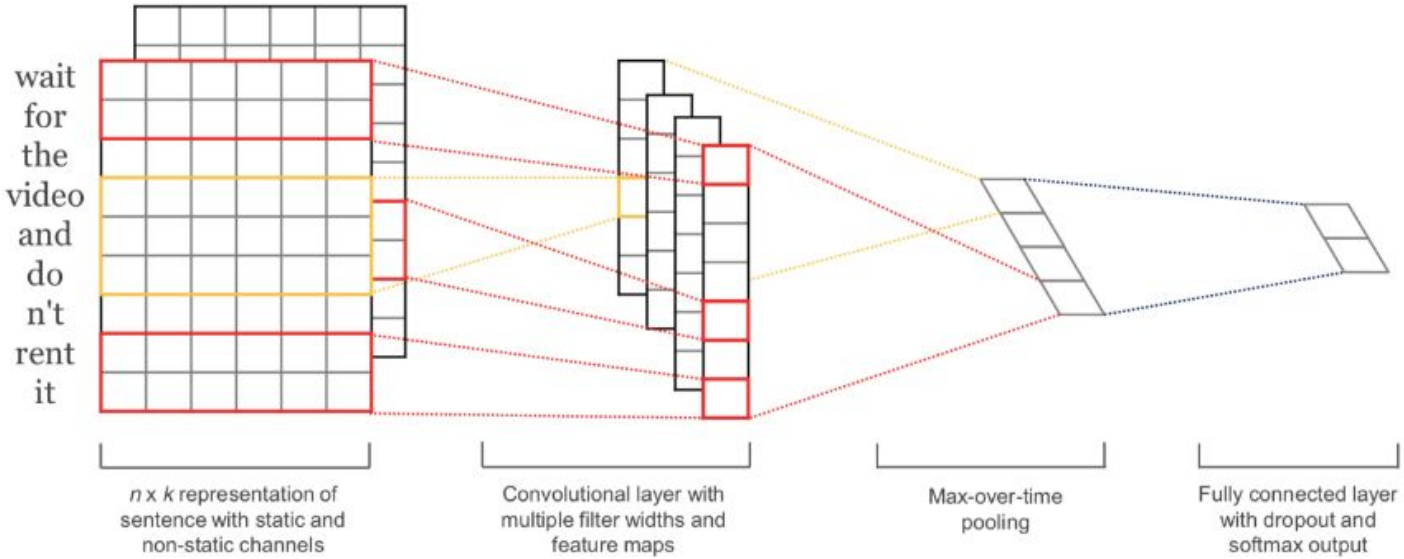
Но можно сдвигаться сильнее и делать меньшее перекрытие.

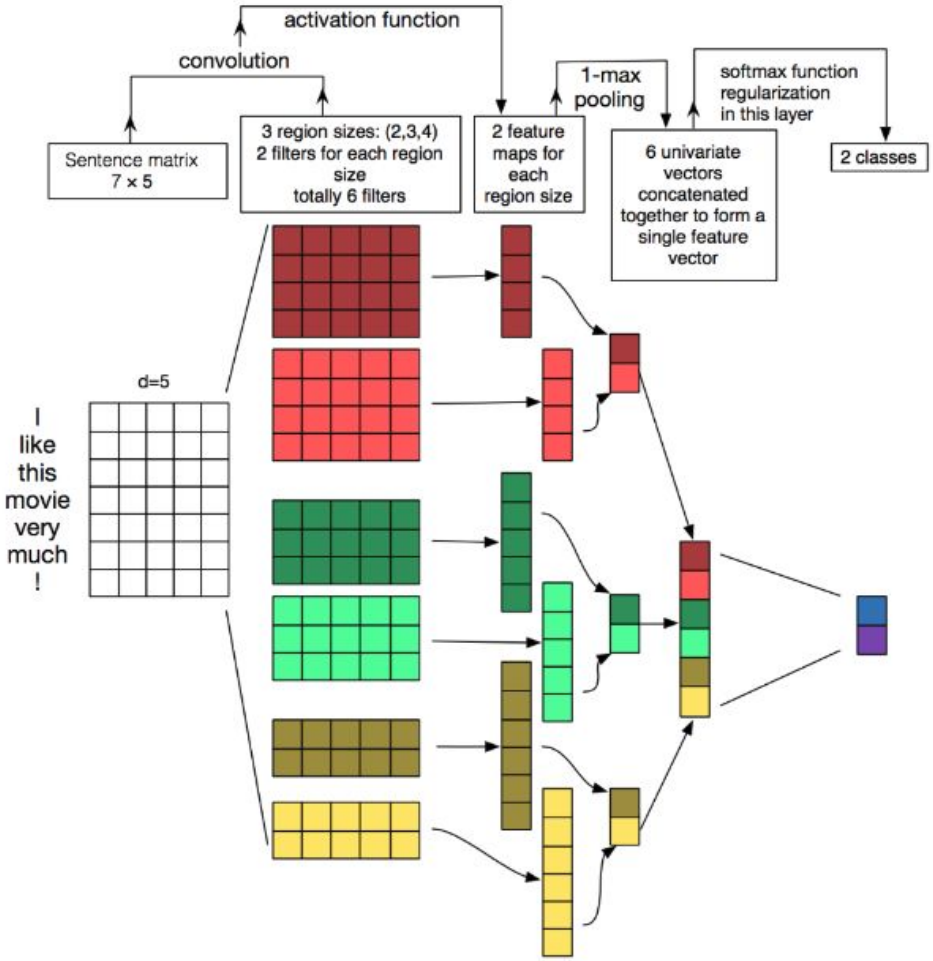


Max pooling


Single depth slice
x
y
1 1 2 4
5 6 7 8
3 2 1 0
1 2 3 4
max pool with 2x2 filters
and stride 2
6 8
3 4

CNN для классификации предложений




activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!