Анализ чувствительности сверточных нейронных сетей

Рысьмятова Анастасия

ВМК МГУ 417 группа

11.11.2015

О чем пойдет речь

1. Yoon Kim.
   2014.
   Convolutional neural networks for sentence classification.
   В данной статье описывается как с помощью однослойной сверточной нейронной решить задачу классификации предложений.

2. Ye Zhang, Byron C. Wallace.
   2015.
   A Sensitivity Analysis of Convolutional Neural Networks for Sentence Classification
   В данной статье исследуется зависимость результата классификации предложений от различных параметров однослойной сверточной нейронной сети.

Данные

1. **MR**
   Отзывы о фильме в одном предложении. Классификация на 2 класса: положительные / отрицальные.
2. **SST-1**
   Расширение MR. Классификация на 5 классов.
3. **SST-2**
   SST-1 с удаленными нейтральными отзывами.
4. **Subj**
   Классифицировать предложения на 2 класса.
5. **TREC**
   Классифицировать вопросы на 6 типов.
6. **CR**
   Отзывы покупателей о продукте. Классифицировать на 2 класса.
7. **MPQA**
   Положительные/отрицательные отзывы.

Данные

Dataset	Average length	Maximum length
MR	20	56
SST-1	18	53
SST-2	19	53
Subj	23	120
TREC	10	37
CR	19	105
MPQA	3	36

Table 1: Average length and maximum length of the 7 datasets

Data	c	l	N	$\|V\|$	$\|V_{pre}\|$	*Test*
MR	2	20	10662	18765	16448	CV
SST-1	5	18	11855	17836	16262	2210
SST-2	2	19	9613	16185	14838	1821
Subj	2	23	10000	21323	17913	CV
TREC	6	10	5952	9592	9125	500
CR	2	19	3775	5340	5046	CV
MPQA	2	3	10606	6246	6083	CV

Table 1: Summary statistics for the datasets after tokenization. c: Number of target classes. l: Average sentence length. N: Dataset size. $|V|$: Vocabulary size. $|V_{pre}|$: Number of words present in the set of pre-trained word vectors. *Test*: Test set size (CV means there was no standard train/test split and thus 10-fold CV was used).

Архитектура сети

Пусть x_i - вектор $i-$ого слова в предложении.

$$x_{1:n} = x_1 \oplus x_2 \oplus ... \oplus x_n$$

Здесь $\oplus$ операция объединения

Сверточный слой:

$$c_i = f(w \cdot x_{i:i+h-1} + b)$$
$$c = [c_1, c_2, ..., c_{n-h+1}]$$

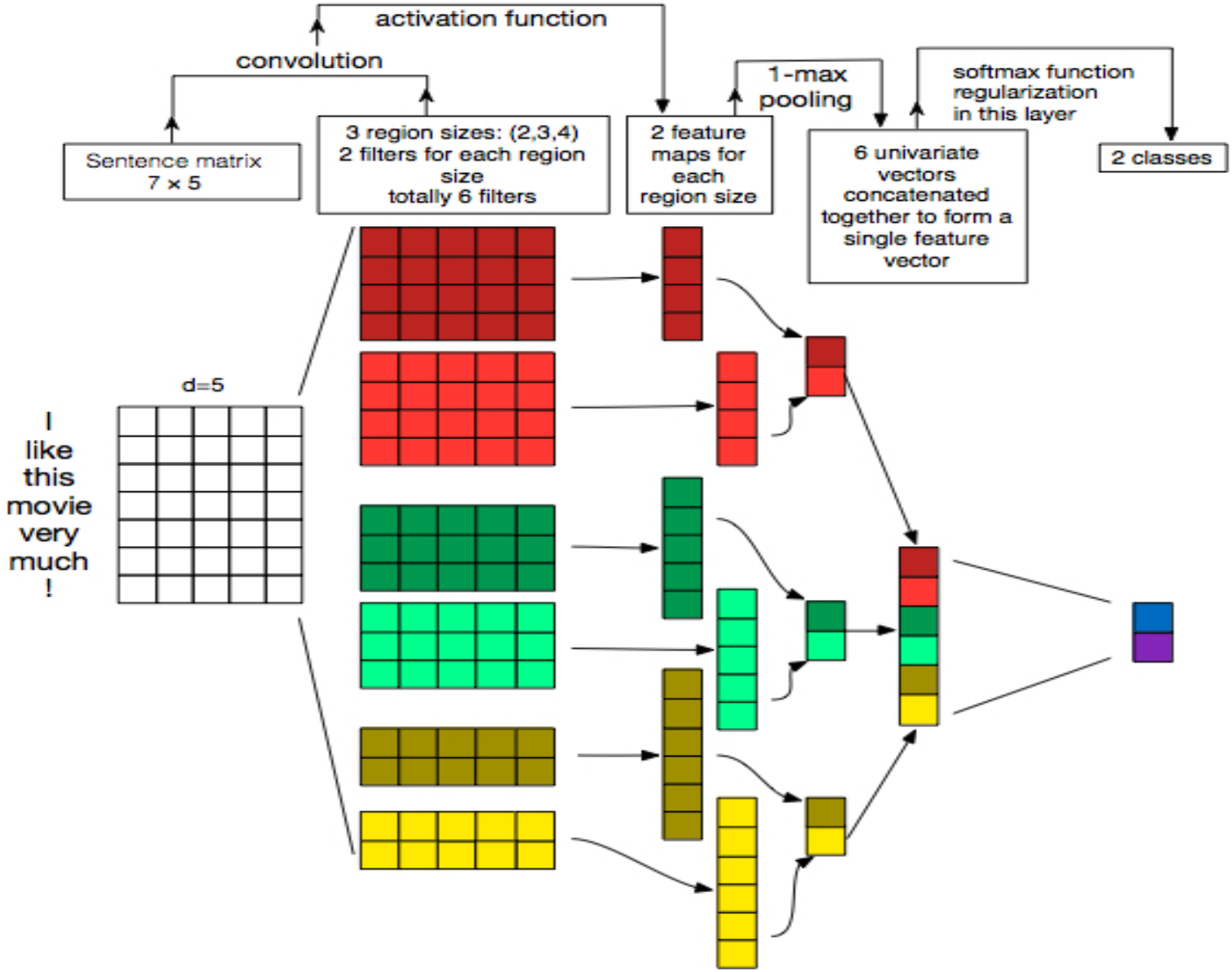
MAX-pooling слой:

$$\widehat{c} = max\{c\}$$

Dropout слой:

$$y = w(z \cdot r) + b$$

Архитектура сети



Word2Vec

Word2Vec — это технология от гугл, которая заточена на статистическую обработку больших массивов текстовой информации.

W2V собирает статистику по совместному появлению слов в фразах, после чего методами нейронных сетей решает задачу снижения размерности и выдает на выходе компактные векторные представления слов, в максимальной степени отражающие отношения этих слов в обрабатываемых текстах.

W2V обучен на более 100 миллиардов слов.

Модели сети

1. CNN-rand: Слова инициализируются случайным образом.
2. CNN-static: Слова переводятся в вектор с помощью Word2Vec.
3. CNN-non-static: Слова переводятся в вектор с помощью Word2Vec, для каждой задачи отдельно.
4. CNN-multichannel: Для каждого слоя существует 2 вектора.

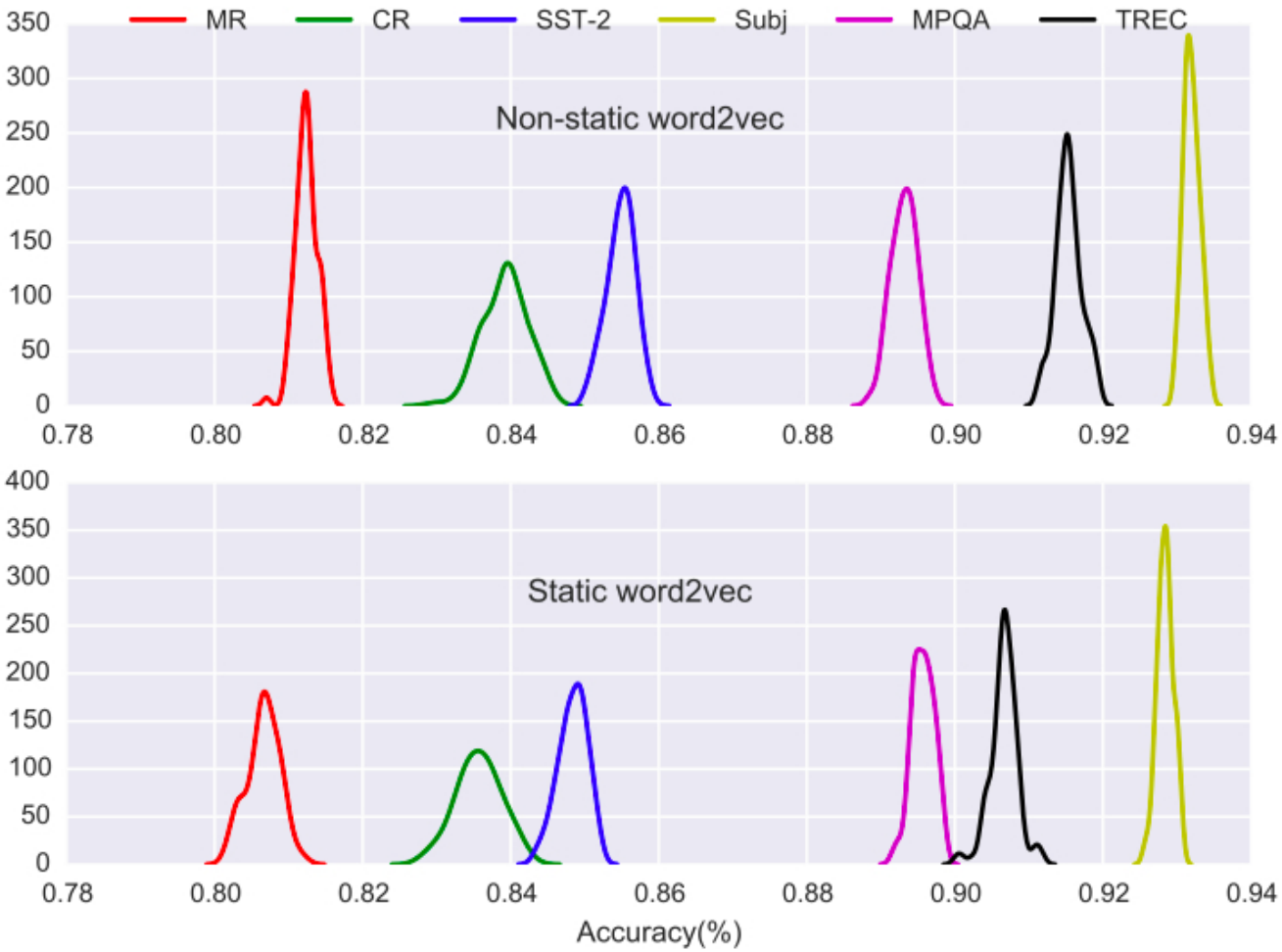
Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	76.1	45.0	82.7	89.6	91.2	79.8	83.4
CNN-static	81.0	45.5	86.8	93.0	92.8	84.7	**89.6**
CNN-non-static	**81.5**	48.0	87.2	93.4	93.6	84.3	89.5
CNN-multichannel	81.1	47.4	**88.1**	93.2	92.2	**85.0**	89.4
RAE (Socher et al., 2011)	77.7	43.2	82.4	–	–	–	86.4
MV-RNN (Socher et al., 2012)	79.0	44.4	82.9	–	–	–	–
RNTN (Socher et al., 2013)	–	45.7	85.4	–	–	–	–
DCNN (Kalchbrenner et al., 2014)	–	48.5	86.8	–	93.0	–	–
Paragraph-Vec (Le and Mikolov, 2014)	–	**48.7**	87.8	–	–	–	–
CCAE (Hermann and Blunsom, 2013)	77.8	–	–	–	–	–	87.2
Sent-Parser (Dong et al., 2014)	79.5	–	–	–	–	–	86.3
NBSVM (Wang and Manning, 2012)	79.4	–	–	93.2	–	81.8	86.3
MNB (Wang and Manning, 2012)	79.0	–	–	**93.6**	–	80.0	86.3
G-Dropout (Wang and Manning, 2013)	79.0	–	–	93.4	–	82.1	86.1
F-Dropout (Wang and Manning, 2013)	79.1	–	–	**93.6**	–	81.9	86.3
Tree-CRF (Nakagawa et al., 2010)	77.3	–	–	–	–	81.4	86.1
CRF-PR (Yang and Cardie, 2014)	–	–	–	–	–	82.7	–
SVM_S (Silva et al., 2011)	–	–	–	–	**95.0**	–	–

Базовая конфигурация нейронной сети

Description	Values
input word vectors	Google word2vec
filter region size	(3,4,5)
feature maps for each region size	100
activation function	ReLU
pooling	1-max pooling
dropout rate	0.5
$l2$ norm constraint on weight vector	3

Table 3: Baseline configuration.

Базовая конфигурация нейронной сети



Влияние входного вектора слова

Использование GloVe

Dataset	Non-static Word2vec-CNN	Static Word2vec-CNN
MR	81.24 (80.69, 81.56)	80.66 (80.16, 81.22)
SST-1	47.08 (46.42,48.01)	45.54 (45.03,46.27)
SST-2	85.49 (85.03, 85.90)	84.84 (84.34,85.20)
Subj	93.20 (92.97, 93.45)	92.84 (92.56,93.06)
TREC	91.54 (91.15, 91.92)	90.66 (90.02, 91.18)
CR	83.92 (82.95, 84.56)	83.57 (82.78, 84.28)
MPQA	89.32 (88.84, 89.73)	89.57 (89.18, 89.85)

Dataset	Non-static GloVe-CNN	Static GloVe-CNN
MR	81.03 (80.68,81.48)	80.10 (79.55,80.51)
SST-1	45.65 (45.09,45.94)	44.76 (44.09,45.09)
SST-2	85.22 (85.04,85.48)	84.15 (83.94,84.33)
Subj	93.64 (93.51,93.77)	93.44 (93.28,93.60)
TREC	90.38 (90.19,90.59)	89.68 (89.26,90.05)
CR	84.33 (84.00,84.67)	83.50 (82.84,83.98)
MPQA	89.57 (89.31,89.78)	89.17 (88.98,89.46)

Table 5: Performance of GloVe-CNN

Влияние входного вектора слова

Использование GloVe

Dataset	Non-static Word2vec-CNN	Static Word2vec-CNN
MR	81.24 (80.69, 81.56)	80.66 (80.16, 81.22)
SST-1	47.08 (46.42,48.01)	45.54 (45.03,46.27)
SST-2	85.49 (85.03, 85.90)	84.84 (84.34,85.20)
Subj	93.20 (92.97, 93.45)	92.84 (92.56,93.06)
TREC	91.54 (91.15, 91.92)	90.66 (90.02, 91.18)
CR	83.92 (82.95, 84.56)	83.57 (82.78, 84.28)
MPQA	89.32 (88.84, 89.73)	89.57 (89.18, 89.85)

Dataset	Non-static GloVe+Word2vec CNN
MR	81.02 (80.75,81.32)
SST-1	45.98 (45.49,46.65)
SST-2	85.45 (85.03,85.82)
Subj	93.66 (93.39,93.87)
TREC	91.37 (91.13,91.62)
CR	84.65 (84.21,84.96)
MPQA	89.55 (89.22,89.88)

Table 6: Performance of non-static GloVe + word2vec CNN

Влияние входного вектора слова

Использование One-hote кодировки

Dataset	Non-static Word2vec-CNN	Static Word2vec-CNN
MR	81.24 (80.69, 81.56)	80.66 (80.16, 81.22)
SST-1	47.08 (46.42,48.01)	45.54 (45.03,46.27)
SST-2	85.49 (85.03, 85.90)	84.84 (84.34,85.20)
Subj	93.20 (92.97, 93.45)	92.84 (92.56,93.06)
TREC	91.54 (91.15, 91.92)	90.66 (90.02, 91.18)
CR	83.92 (82.95, 84.56)	83.57 (82.78, 84.28)
MPQA	89.32 (88.84, 89.73)	89.57 (89.18, 89.85)

Dataset	One-hot vector CNN
MR	77.83 (76.56,78.45)
SST-1	41.96 (40.29,43.46)
SST-2	79.80 (78.53,80.52)
Subj	91.14 (90.38,91.53)
TREC	88.28 (87.34,89.30)
CR	78.22 (76.67,80.00)
MPQA	83.94 (82.94,84.31)

Table 7: Performance of one-hot vector CNN

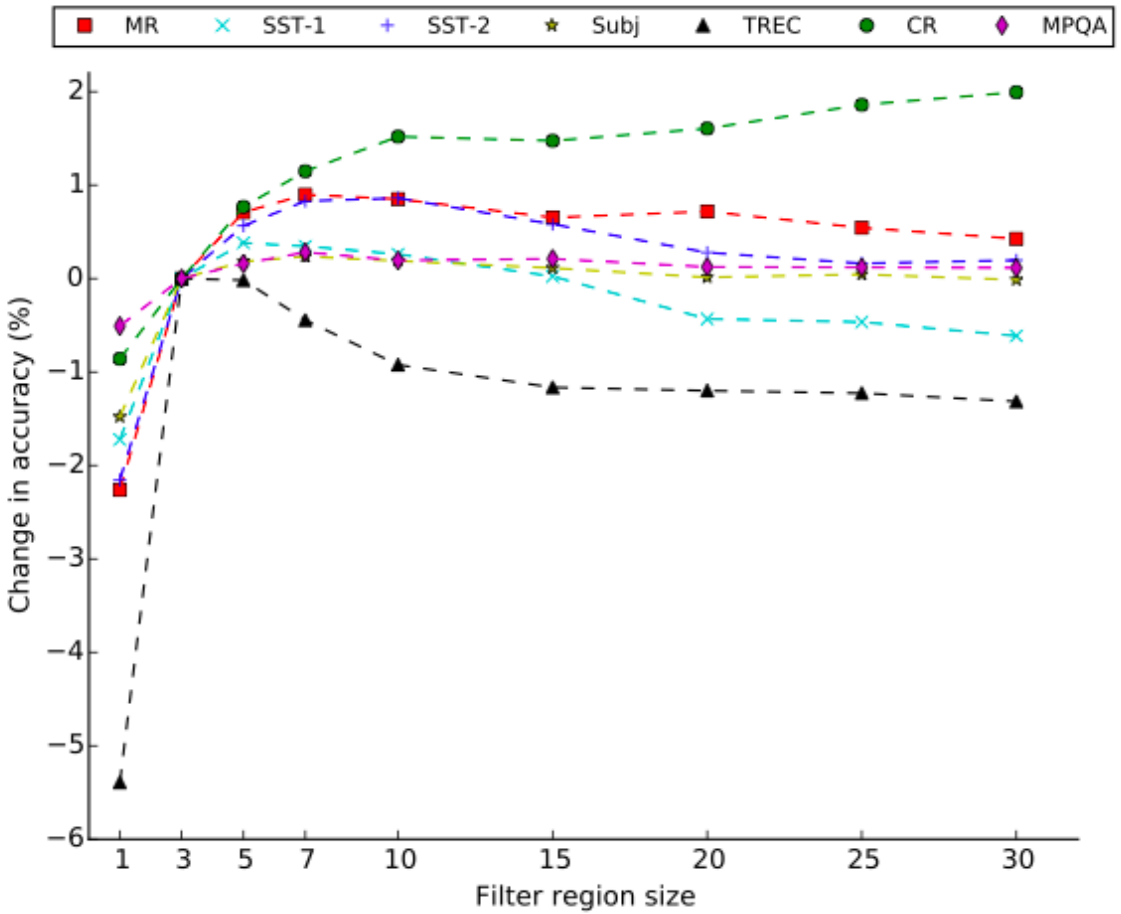
Влияние входного вектора слова

Выводы:

Результаты классификации про использовании *GloVe* и *Word2Vec* зависят от выборки. Нельзя одназначно определить лучший метод.

One-hot кодировка работает заметно хуже чем *GloVe* и *Word2Vec*.

Влияние размера ядер свертки



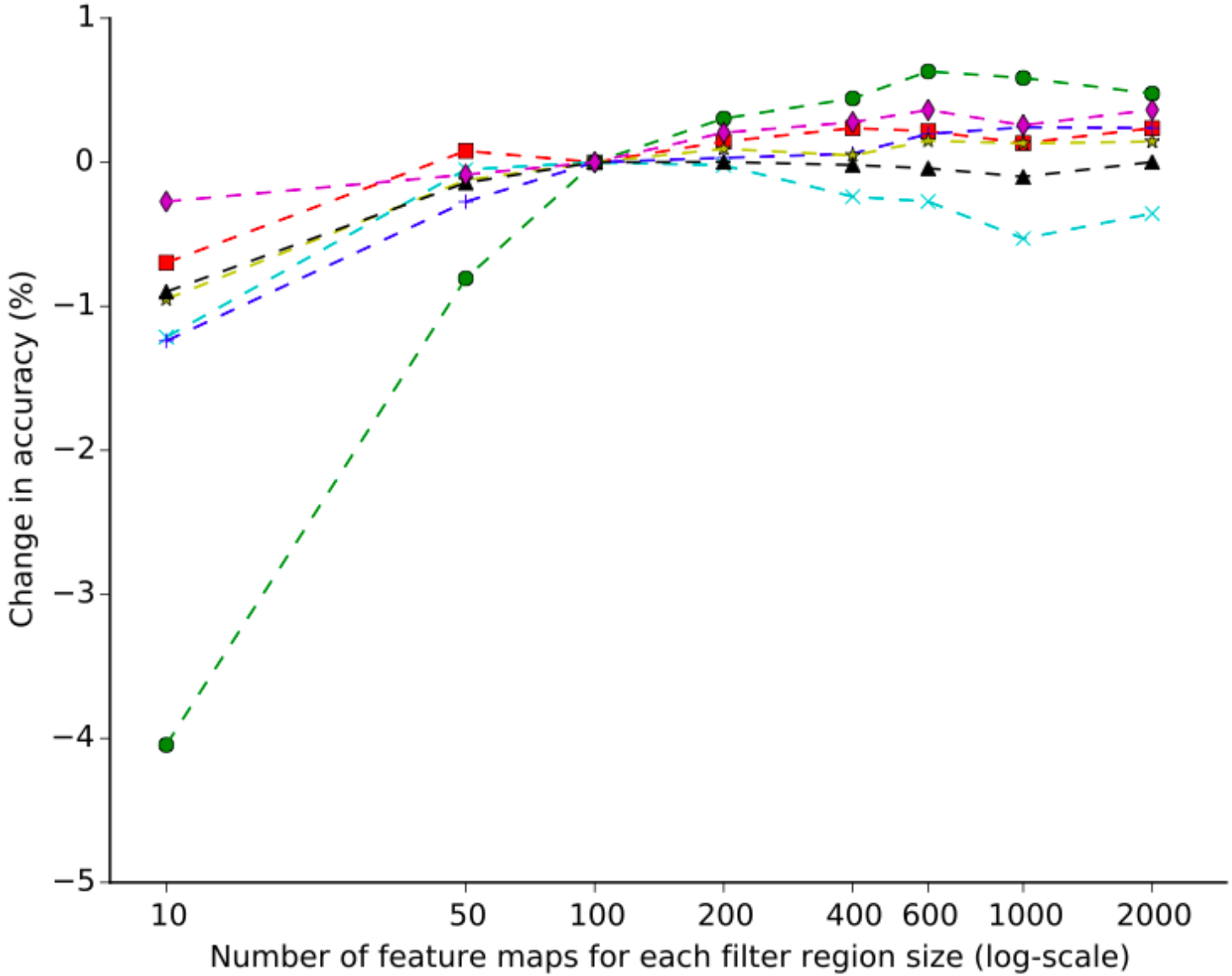
Влияние размера ядер свертки

Выводы:

Оптимальный размер ядра свертки для каждой выборки свой.

У всех выборок кроме CR оптимальный размер ядра свертки от 5 до 25. Это можно объяснить тем, что в CR предложения длинне, чем в остальных выборках.

Влияние количества ядер свертки для каждого размера



Влияние количества ядер свертки для каждого размера

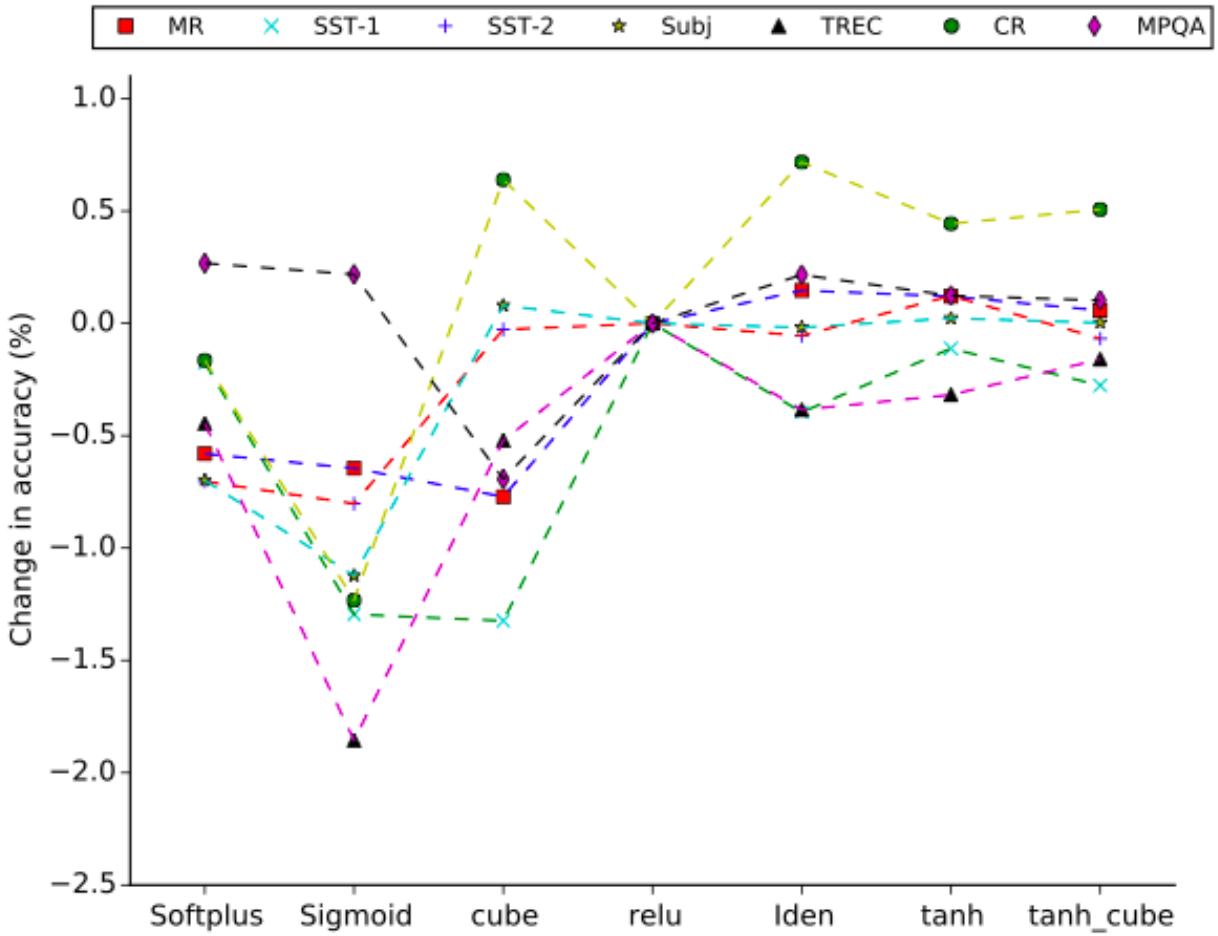
Выводы:

Оптимальное количество ядер свертки для каждой выборки разное.

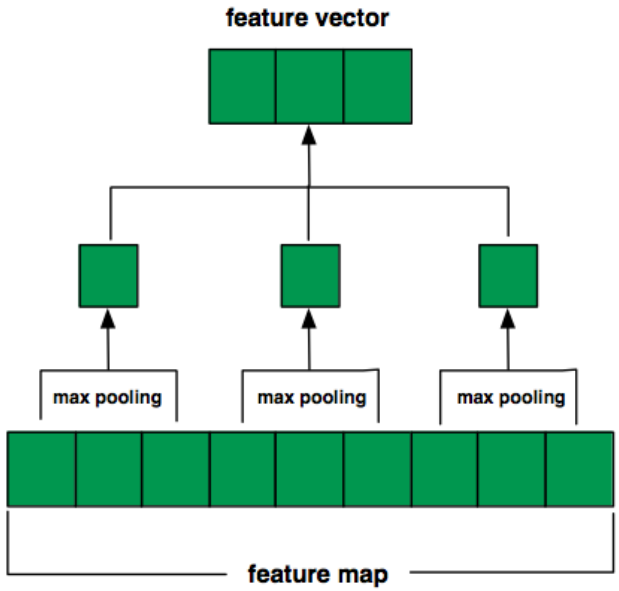
Оптимальное количество ядер свертки лежало в диапазоне от 50 до 600.

Выбирая значение больше 600 происходило переобучение.

Влияние функции активации

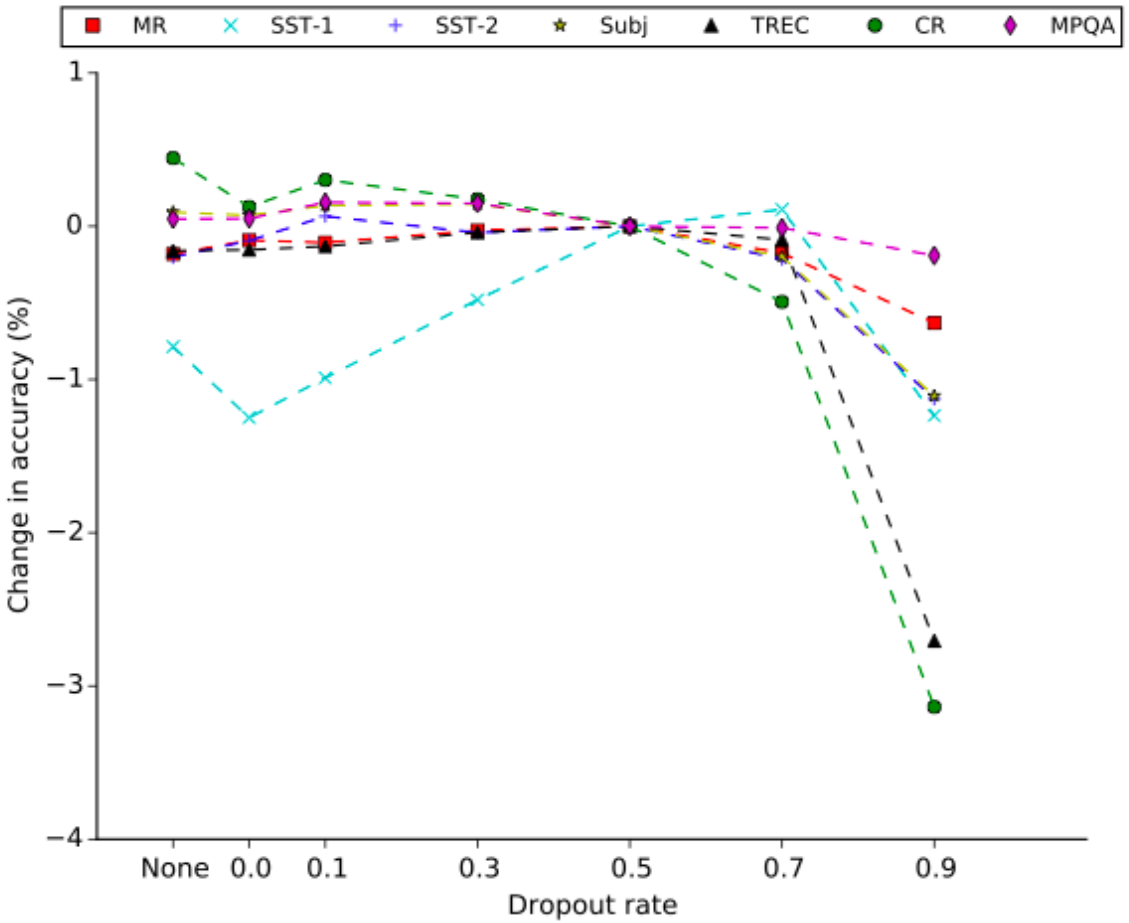


Влияние выбора стратегии объединения



	max,3	max,10	max,20	max,30	max,all (1-max)
MR	79.75 (79.47,80.03)	80.20 (80.02,80.35)	80.68 (80.14,81.21)	80.99 (80.65,81.30)	81.28 (81.16,81.54)
SST-1	44.98 (44.06,45.68)	46.10(45.37,46.84)	46.75 (46.35,47.36)	47.02 (46.59,47.59)	47.00 (46.54,47.26)
SST-2	83.69(83.46,84.07)	84.63 (84.44,84.88)	85.18 (84.64,85.59)	85.38 (85.31,85.49)	85.50 (85.31,85.83)
Subj	92.60 (92.28,92.76)	92.87 (92.69,93.17)	93.06 (92.81,93.19)	93.13 (92.79,93.32)	93.20 (93.00,93.36)
TREC	90.29 (89.93,90.61)	91.42 (91.16,91.71)	91.52 (91.23,91.72)	91.47 (91.15,91.64)	91.56 (91.67,91.88)
CR	81.72 (81.21,82.20)	82.71 (82.06,83.30)	83.44(83.06,83.90)	83.70 (83.31,84.25)	83.93 (83.48,84.39)
MPQA	89.15 (88.83,89.47)	89.39 (89.14,89.56)	89.30 (89.16,89.60)	89.37 (88.99,89.61)	89.39 (89.04,89.73)

Влияние регуляризации



Рекомендации по выбору параметров для данной задачи

1. Использовать Word2Vec или GloVe.
2. Оптимальный размер ядра свертки выбирать в диапазоне 2-10. (Но для классификации более объемных текстов возможно надо рассматривать другой диапазон)
3. Количество ядер свертки каждого размера выбирать в диапазоне от 50 до 600.
4. Выбирать функцию активации ReLu или tanh.
5. Использовать общее объединение по масимуму. (1 Max-poolin)
6. Dropout выбирать в диапазоне от 0.0 до 0.5.