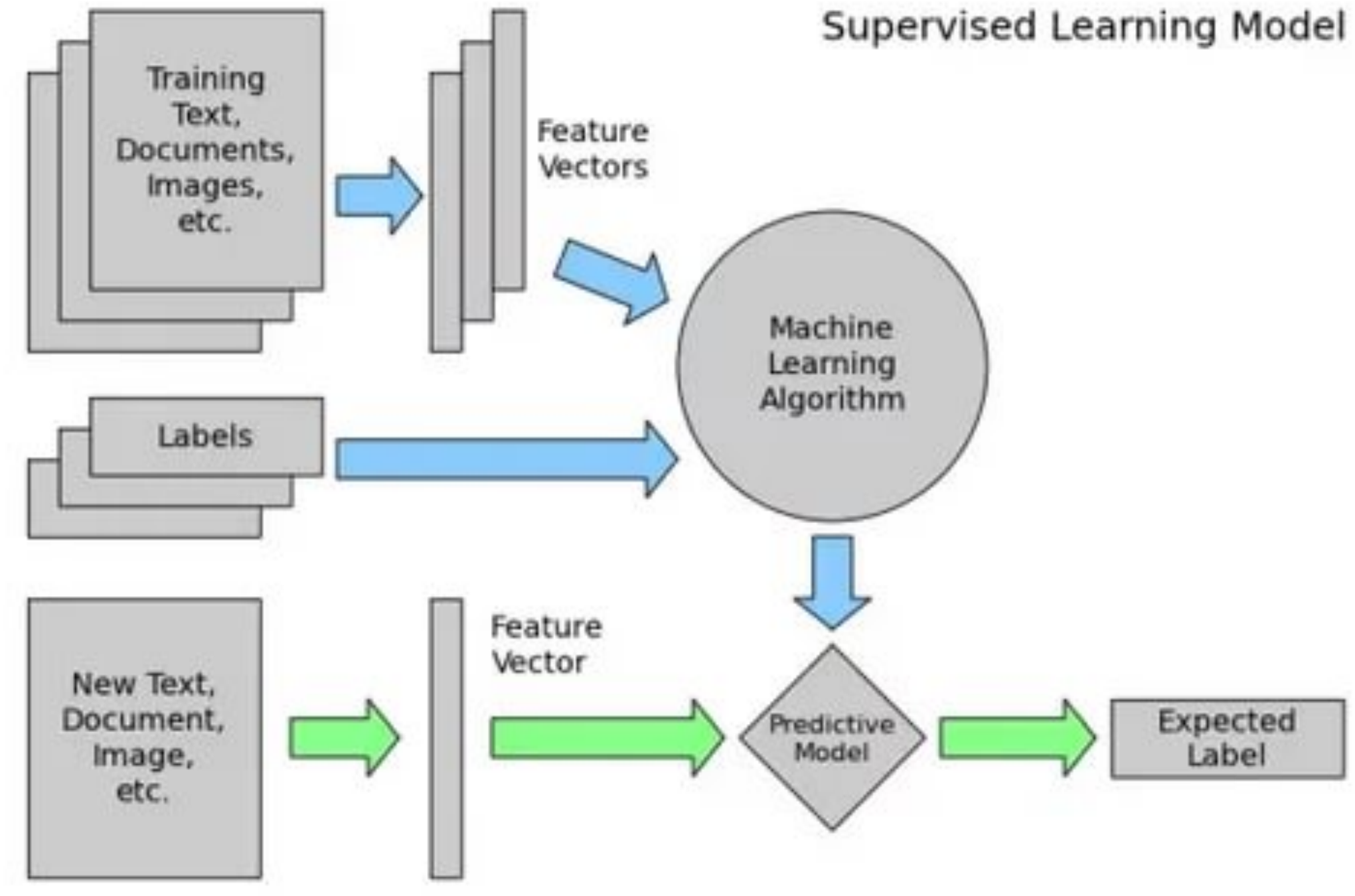
Конволюционные сети для последовательностей

Смирнов Павел БПМИ172


Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

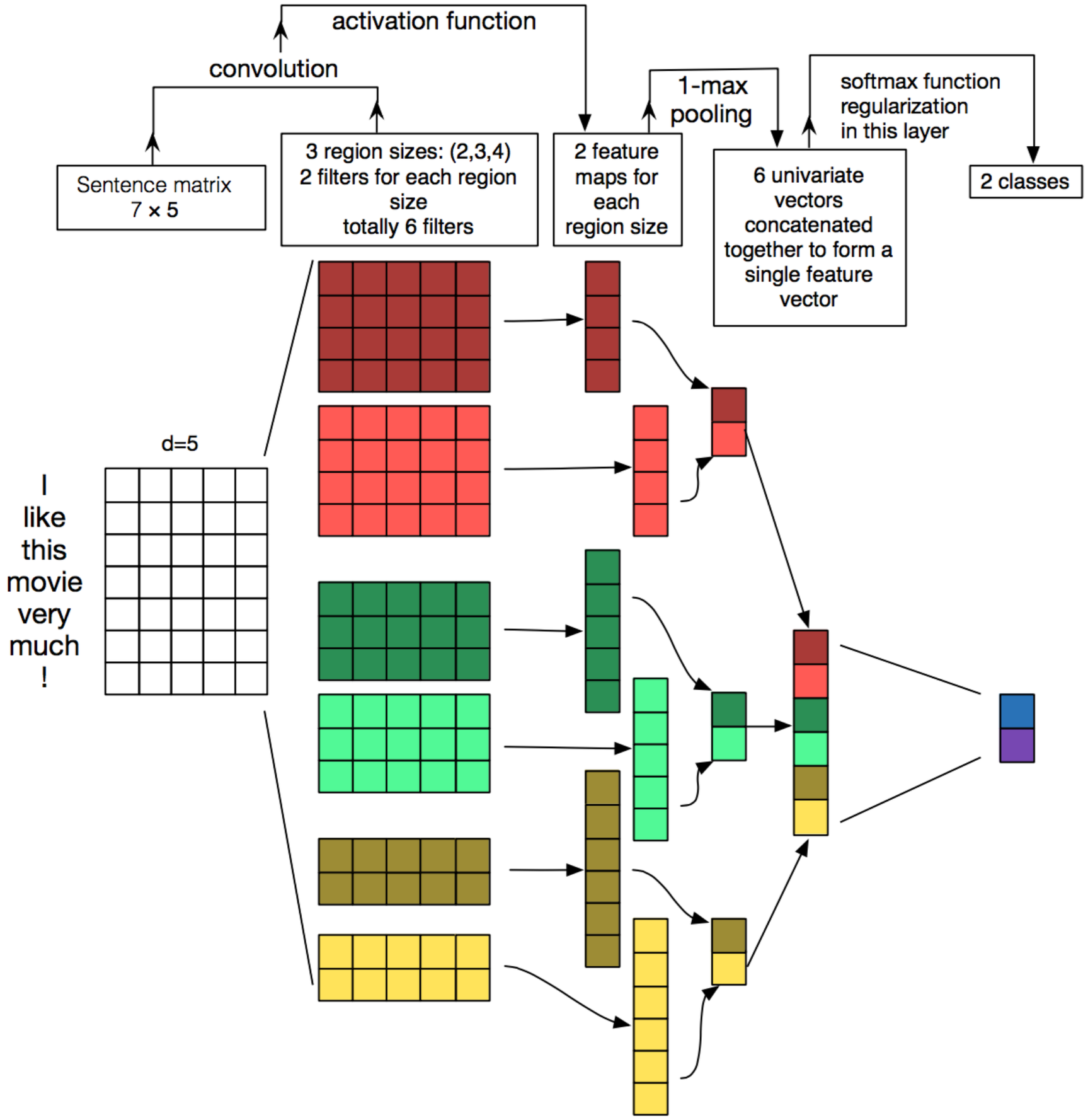
Softmax

- $\sigma(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$-Функция преобразует вектор z размерности K в вектор σ той же размерности, где каждая координата σ_i полученного вектора представлена вещественным числом в интервале [0,1] и сумма координат равна 1.

- $z = w^T x - \theta$ где x — вектор-столбец признаков объекта размерностиM×1;wT — транспонированная матрица весовых коэффициентов признаков, имеющая размерность K×M;
θ — вектор-столбец с пороговыми значениями размерности K×1, где K— количество классов объектов, а M — количество признаков объектов.

CNN в текстах

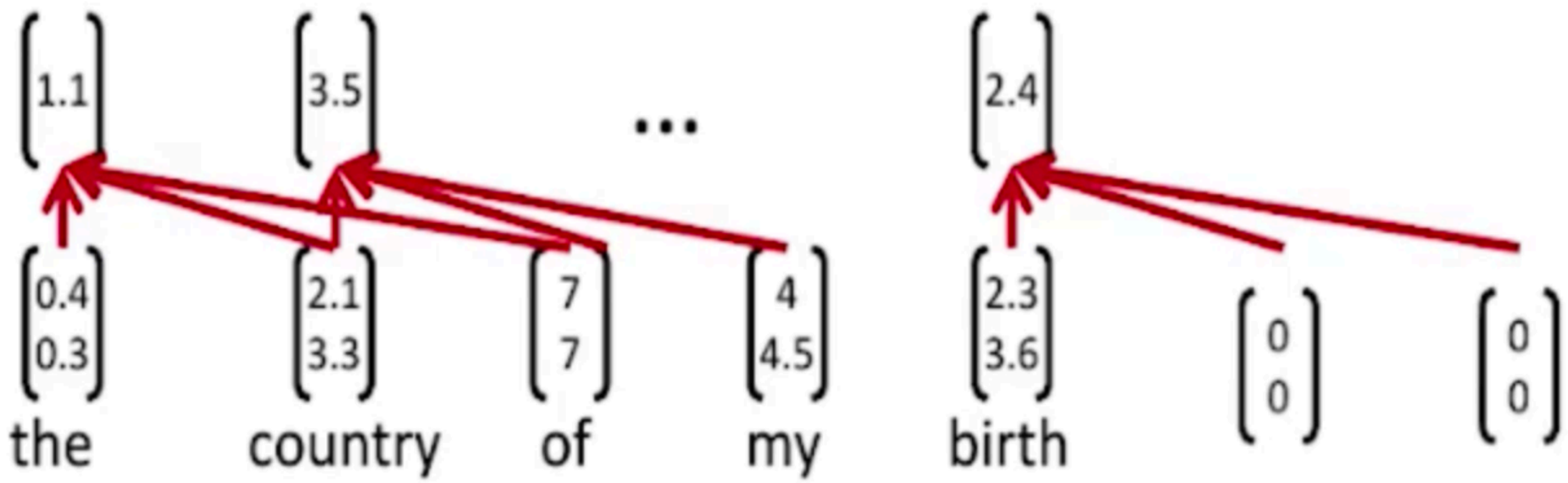
- Съешь ещё этих мягких французских булок:
«Съешь ещё», «ещё этих», … , «Съешь ещё этих», «ещё этих мягких», …, «Съешь ещё этих мягких», «ещё этих мягких французских», … , «Съешь ещё этих мягких французских булок»


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
!

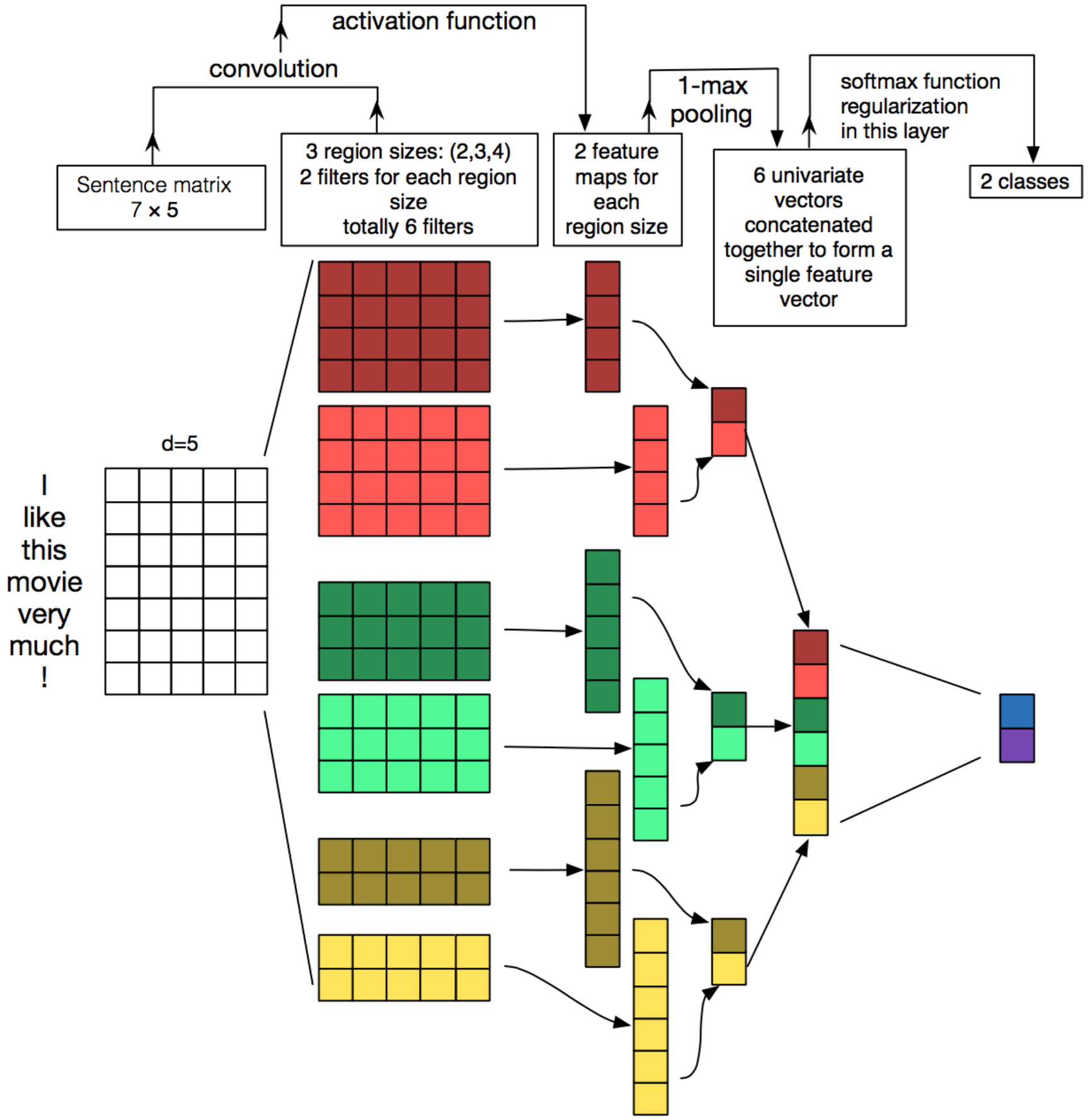
Проблема с фиксированным входом

$$c_i = f\left(\mathbf{w}^T \mathbf{x}_{i:i+h-1} + b\right)$$

Result is a feature map: $\mathbf{c} = [c_1, c_2, \ldots, c_{n-h+1}] \in \mathbb{R}^{n-h+1}$

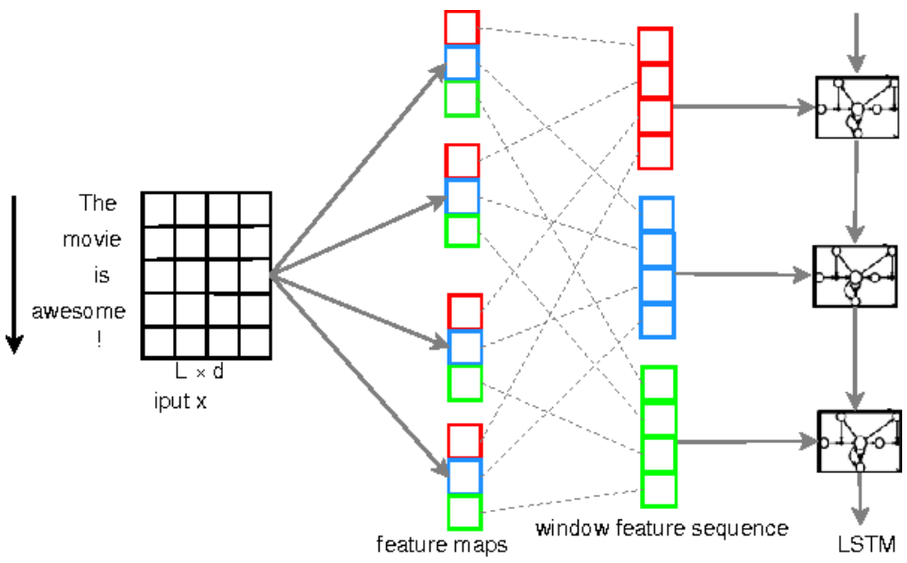


$$\hat{c} = max_i(c_i)$$


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
!

Современные CNN для NLP

- CLSTM



- VDCNN аналог ResNet для текстов (2016) нужны большие input

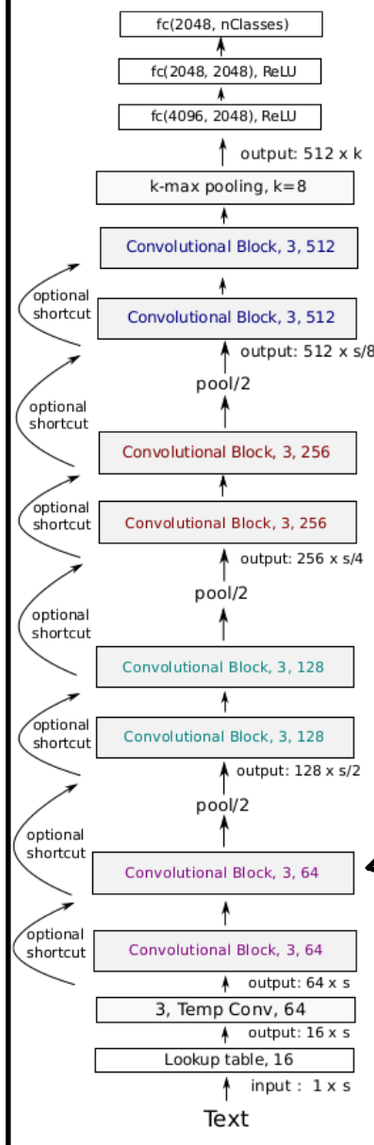
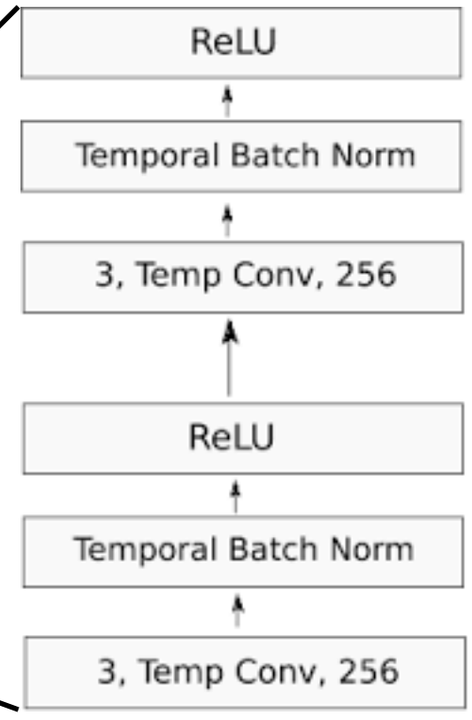
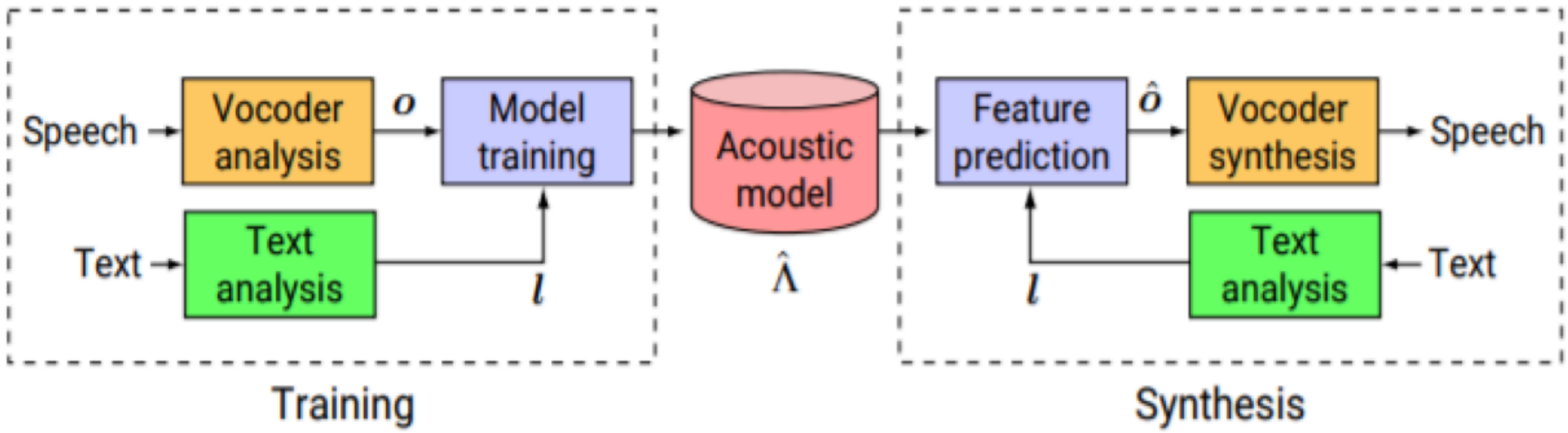


Figure 1: VDCNN architecture.

Каждый convolutional block это:



WaveNet



- Совместная вероятность волны $x = (x_1, \ldots, x_T)$ описывается как произведение условных вероятностей уравнением:

$$p(x) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1})$$

Свертка

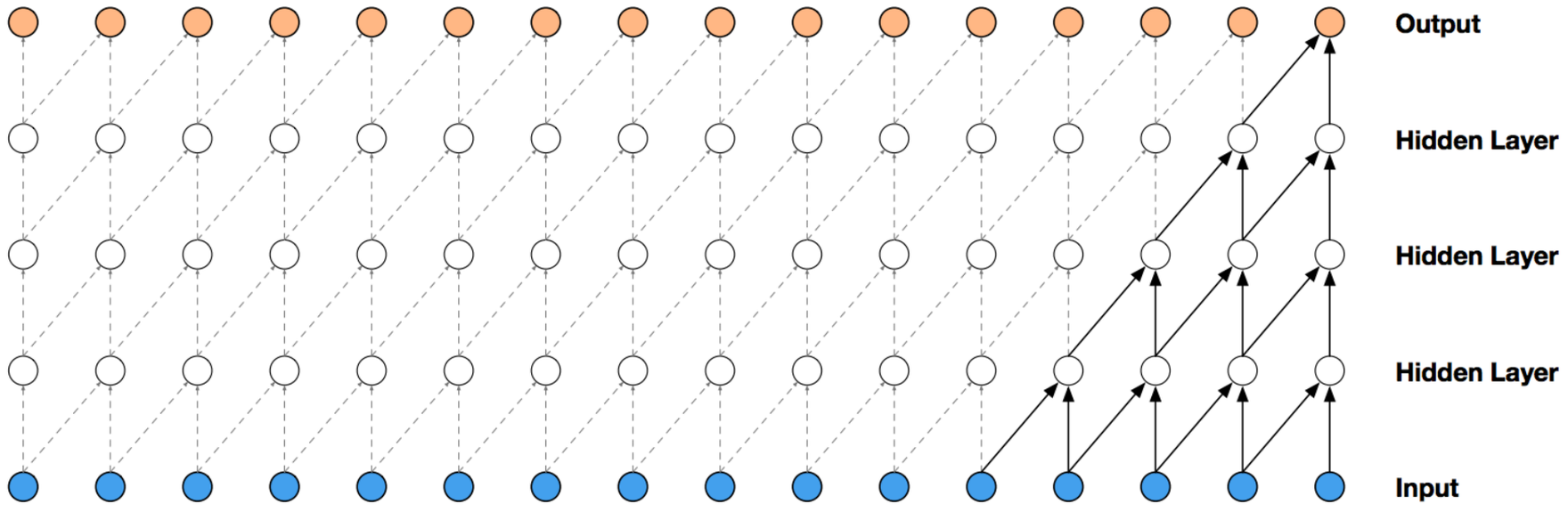
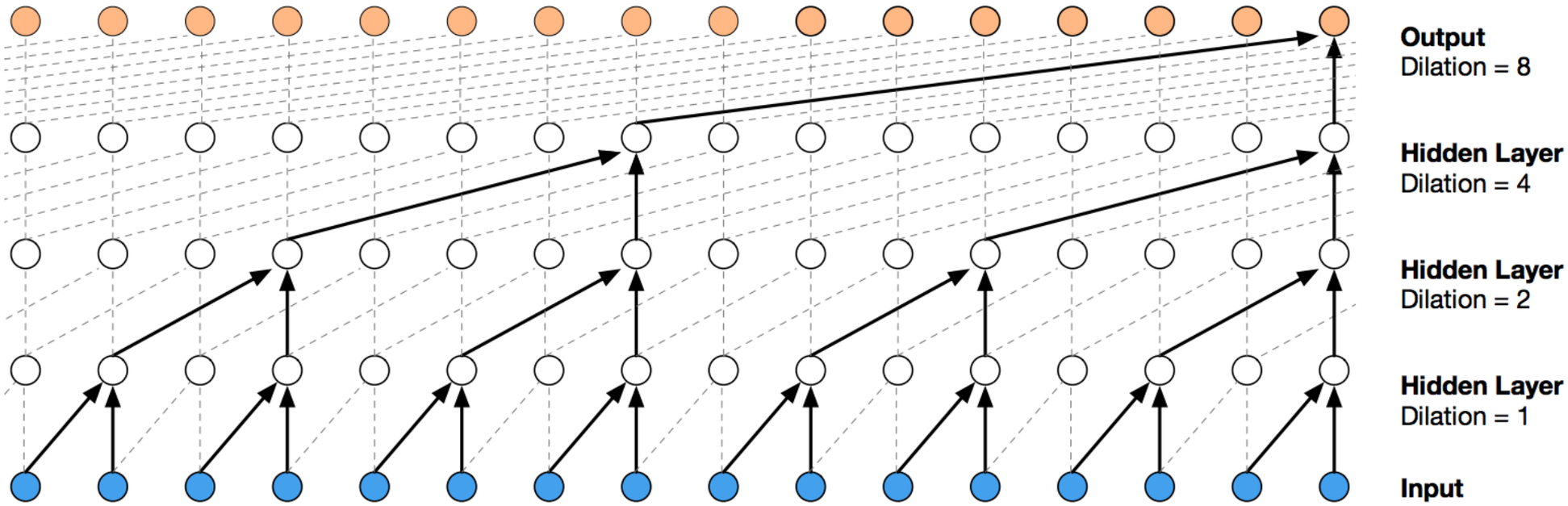


Figure 2: Visualization of a stack of causal convolutional layers.

-

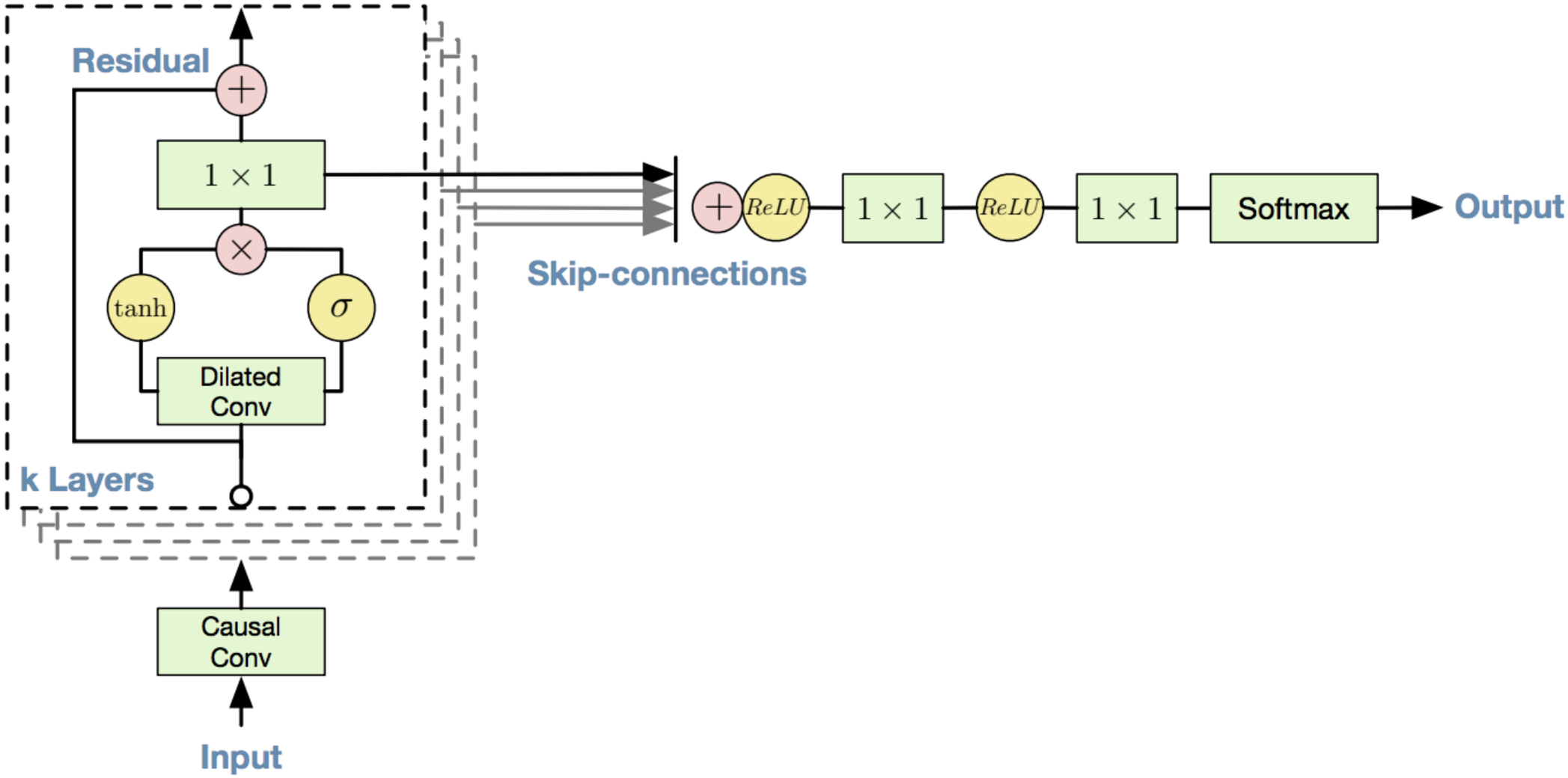
Дырявые свертки



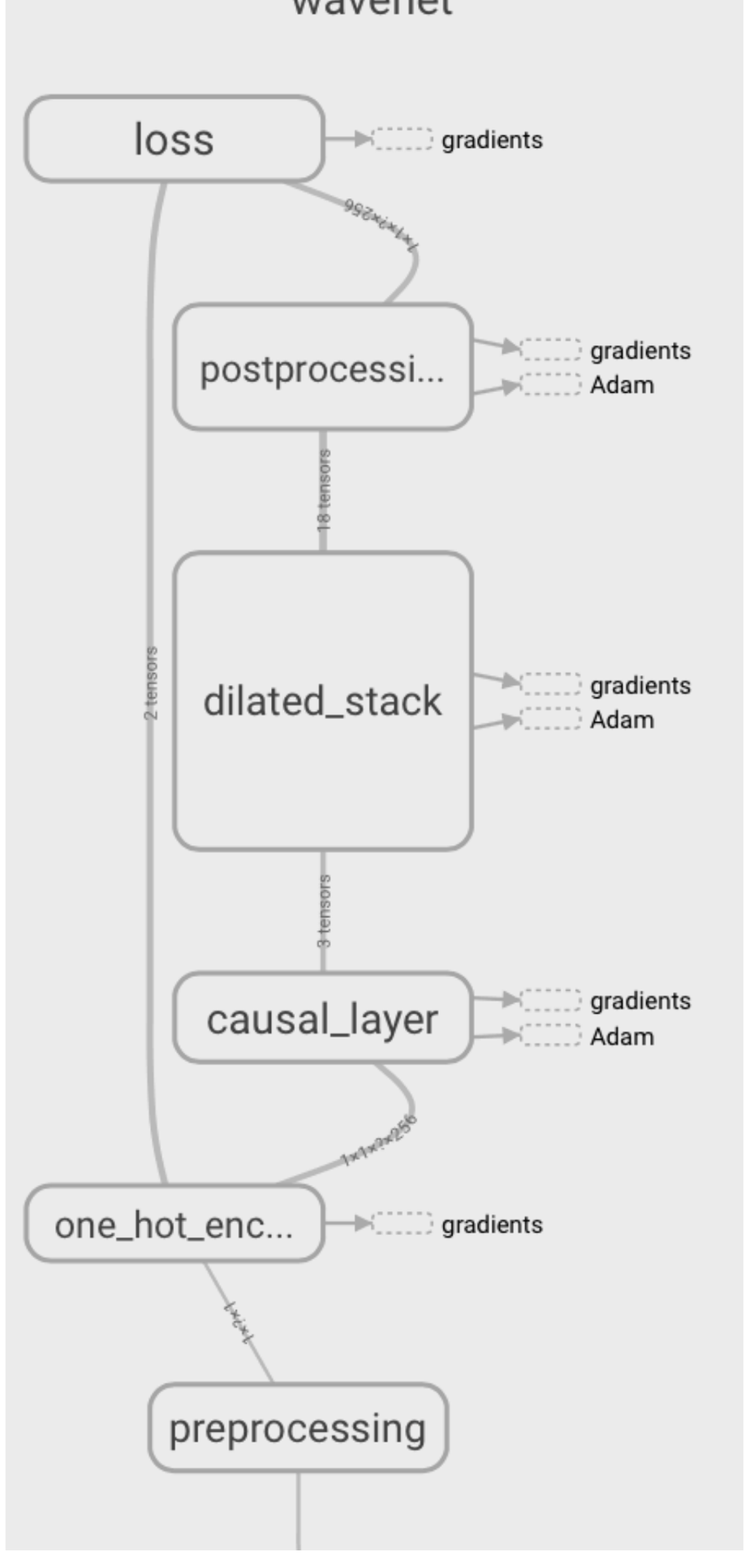
Функция активации

$$z = tanh(W_{f,k} * x) \odot \sigma(W_{g,k} * x)$$

где $*$ обозначает операцию свёртки, $\odot$ обозначает поэлементное умножение, σ(·) сигмоида, k номер слоя, f и g обозначают фильтр и gate соответственно, и W обучаемый свёрточный фильтр.



$$z = tanh(W_{f,k} * x) \odot \sigma(W_{g,k} * x)$$


wavenet
loss
gradients
1×1×256
postprocessi...
gradients
Adam
18 tensors
2 tensors
dilated_stack
gradients
Adam
3 tensors
causal_layer
gradients
Adam
1×1×256
one_hot_enc...
gradients
1×256
preprocessing

WaveNet с условием

- Если дан дополнительный вход h как условие, WaveNet способен моделировать условное распределение $p(x|h) = \prod_{t=1}^{T} p(x_t|x_1, \ldots, x_{t-1}, \mathbf{h}))$ аудио по этому входу.

- Функция активации для глобального условия:
  $z = tanh(W_{f,k} * x + V^T h) \odot \sigma(W_{g,k} * x + V^T h))$
  Где V является обучаемой линейной моделью

- Функция активации для локального условия:
  $z = tanh(W_{f,k} * x + Vh) \odot \sigma(W_{g,k} * x + Vh))$, где $V_{*,k}$ это свёртка 1×1

• Локальное условие .

– Временной ряд, той же длины что и данные. Качество, изменяющееся по времени.

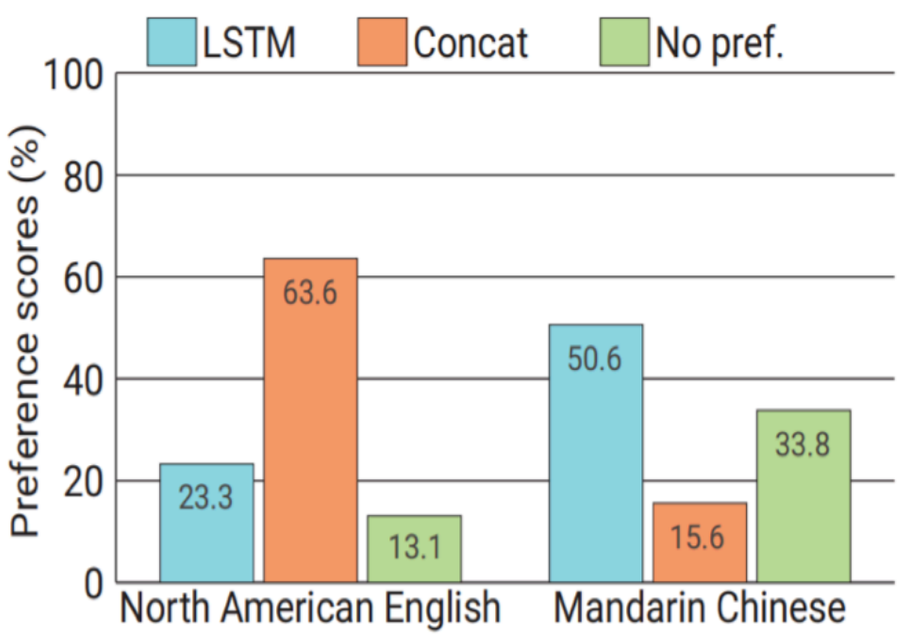
• Глобальное условие.

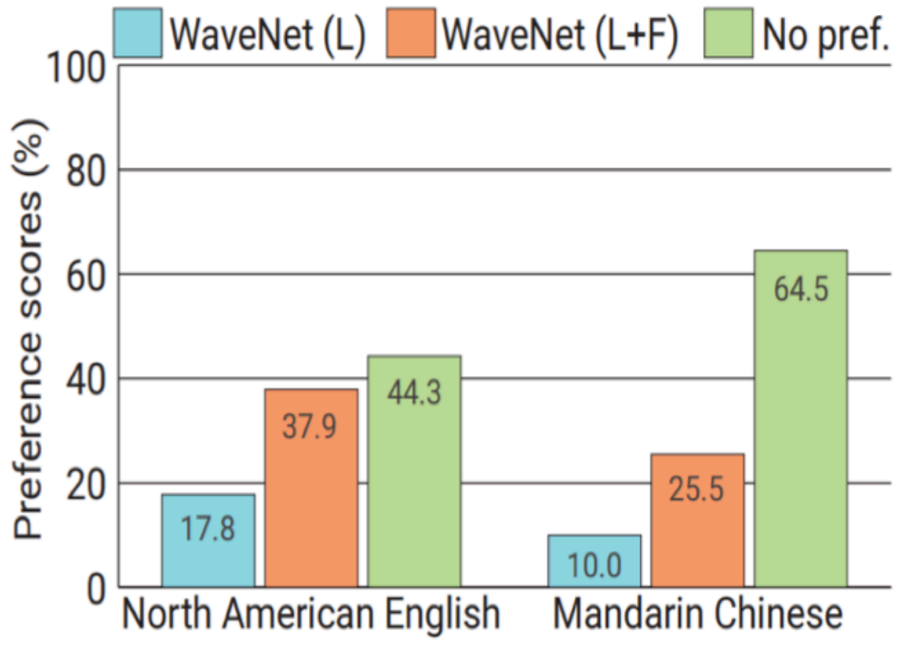
– Качество говорящего, не зависящее от времени. Не меняет своего значения в процессе обучения/генерации.

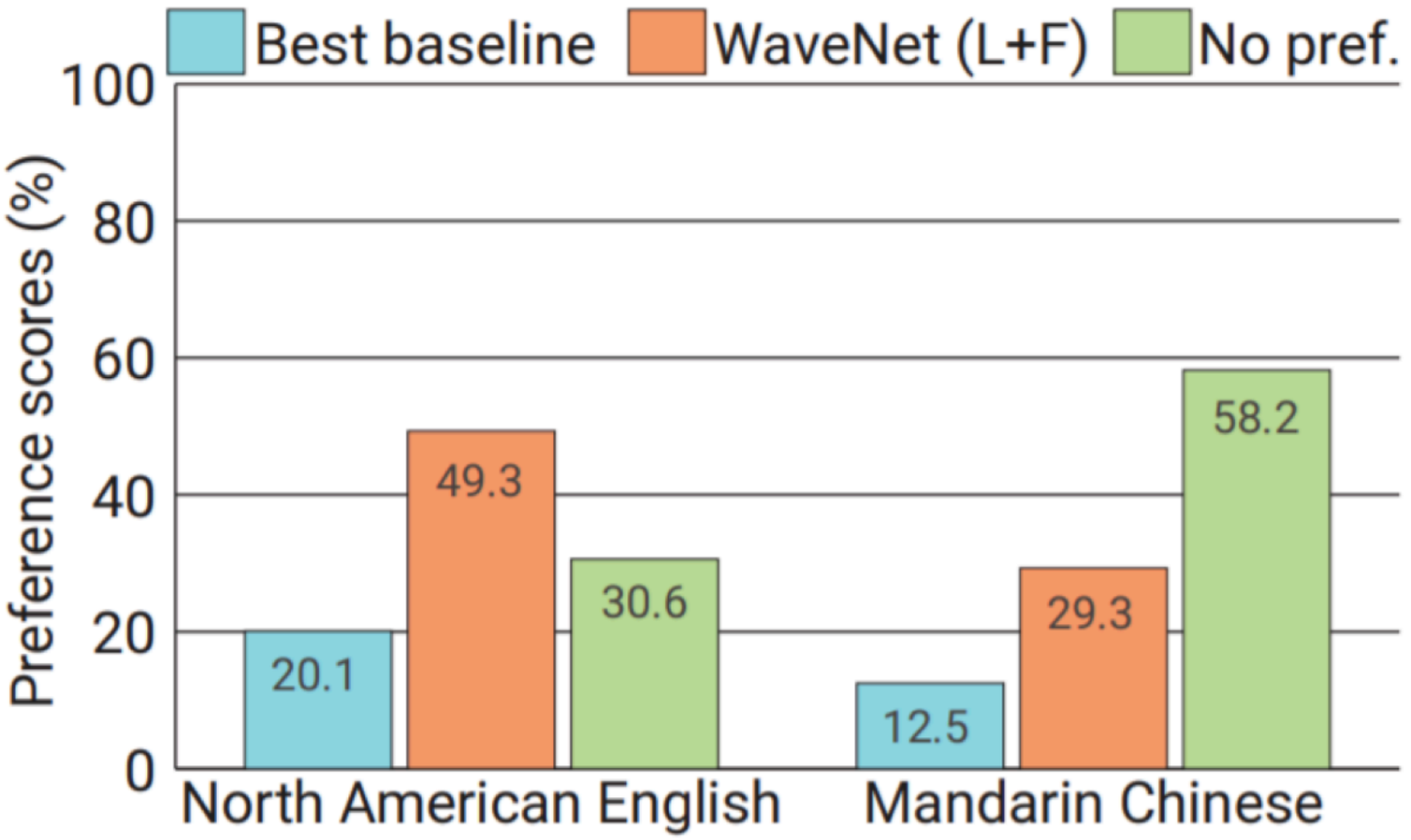
Сравнение качества результатов

Конфигурация WaveNet	**Данные**	**Результат**
Без модификаций	весть корпус	речеподобный звук
Без модификаций	одна фраза, много голосов	высокий гул
Без модификаций	один голос, много фраз	шум
Глобальное условие: ID говорящего	весь корпус	речеподобный звук
Глобальное условие: ID говорящего	одна фраза	шум
Глобальное условие: пол говорящего	два голоса, много фраз	шум
Глобальное условие: пол говорящего	два голоса, одна фраза	шум
Локальное условие: текст	весь корпус	шум
Локальное условие:текст	одна фраза	шум
Локальное условие: yandex-speech	весь корпус	речеподобный звук лучшего качества
Локальное условие: yandex-speech	одна фраза	высокий гул
Глобальное условие ID говорящего + локальное условие yandex-speech	весь корпус	шум
Глобальное условие ID говорящего + локальное условие yandex-speech	одна фраза	шум

	Subjective 5-scale MOS in naturalness	
Speech samples	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071


LSTM
Concat
No pref.
Preference scores (%)
100
80
60
40
20
0
23.3
63.6
13.1
50.6
15.6
33.8
North American English
Mandarin Chinese


WaveNet (L)
WaveNet (L+F)
No pref.
Preference scores (%)
100
80
60
40
20
0
17.8
37.9
44.3
10.0
25.5
64.5
North American English
Mandarin Chinese


Best baseline
WaveNet (L+F)
No pref.
Preference scores (%)
100
80
60
40
20
0
20.1
49.3
30.6
12.5
29.3
58.2
North American English
Mandarin Chinese

Производительность

На GPU Tesla K80 с 11 гигабайтами видеопамяти с поправками на виртуализацию.

Чтобы добиться значений функции потерь хотя бы как на рисунке требуется около 4 суток. Генерация пяти секунд аудио занимает около часа.

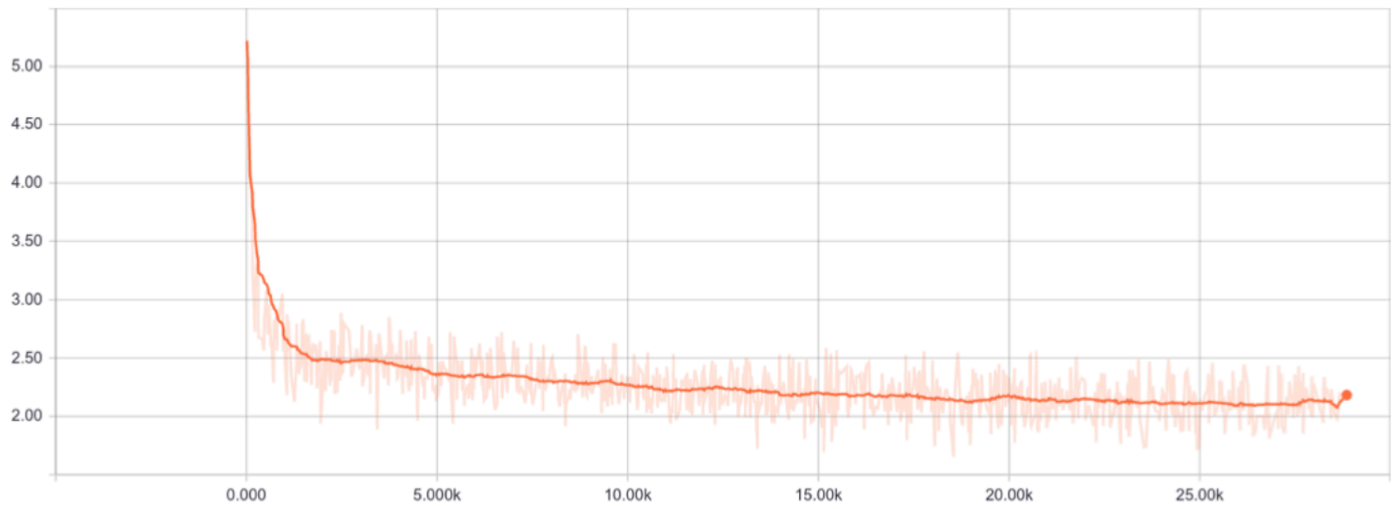
- 

Рис. 11: Изменение функции потерь в процессе обучения

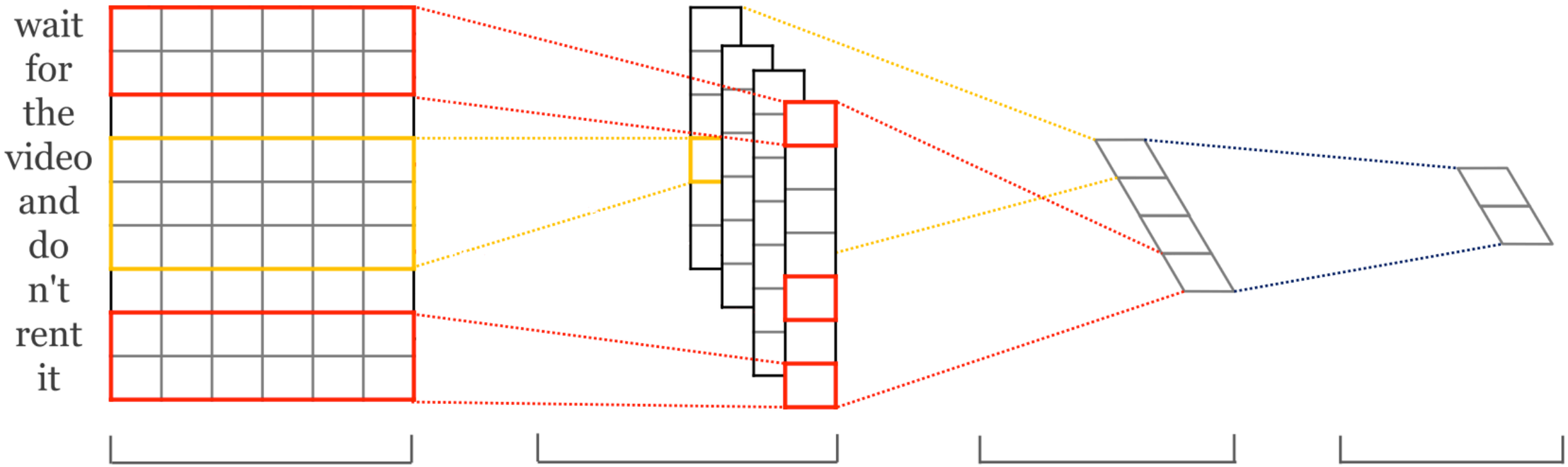
Вопросы

- Напишите функцию активации с условием h?
- В чем преимущество дырявых сверток(delated casual convolutions) над casual convolutions?
- Зачем нужны условия в WaveNet?

Ссылки

- http://web.eng.tau.ac.il/deep_learn/wp-content/uploads/2018/01/WaveNet.pdf
- http://mit.spbau.ru/sewiki/images/e/ed/Kurbanov_diploma_master2017.pdf
- https://habr.com/ru/company/ods/blog/353060/
- https://www.youtube.com/watch?v=GYMfcIqMlOU
- https://www.youtube.com/watch?v=v-7zVVQqCAs
- https://ru.wikipedia.org/wiki/Softmax

Применение сверток к текстам



- первый шаг - это этап предварительной обработки звука, после квантования входной формы волны до фиксированного целочисленного диапазона. Целочисленные амплитуды затем кодируются для получения тензора формы (num_samples, num_channels).

- Свернутый слой, который получает доступ только к текущим и предыдущим входам, затем уменьшает размер канала.

- Ядро сети построено в виде стека причинно-следственных расширенных слоев, каждый из которых представляет собой расширенную свертку (свертку с дырами), которая обращается только к текущим и прошлым аудиоданным.

- Выходы всех уровней объединяются и расширяются до исходного числа каналов с помощью серии плотных слоев постобработки, за которыми следует функция softmax для преобразования выходов в категориальное распределение.

- Функция потерь представляет собой перекрестную энтропию между выходом для каждого временного интервала и входом на следующем временном интервале.