Final prediction
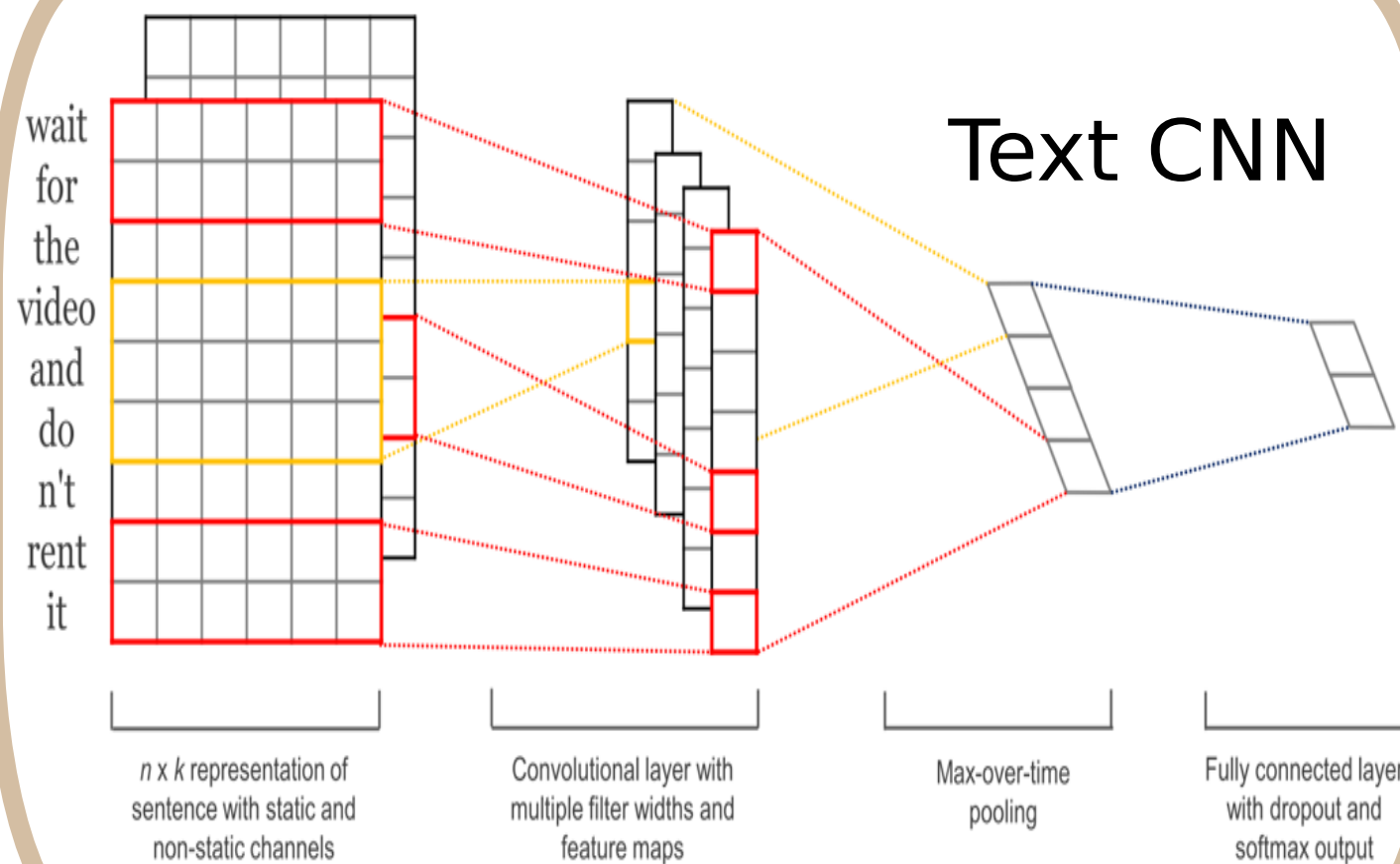
Policy

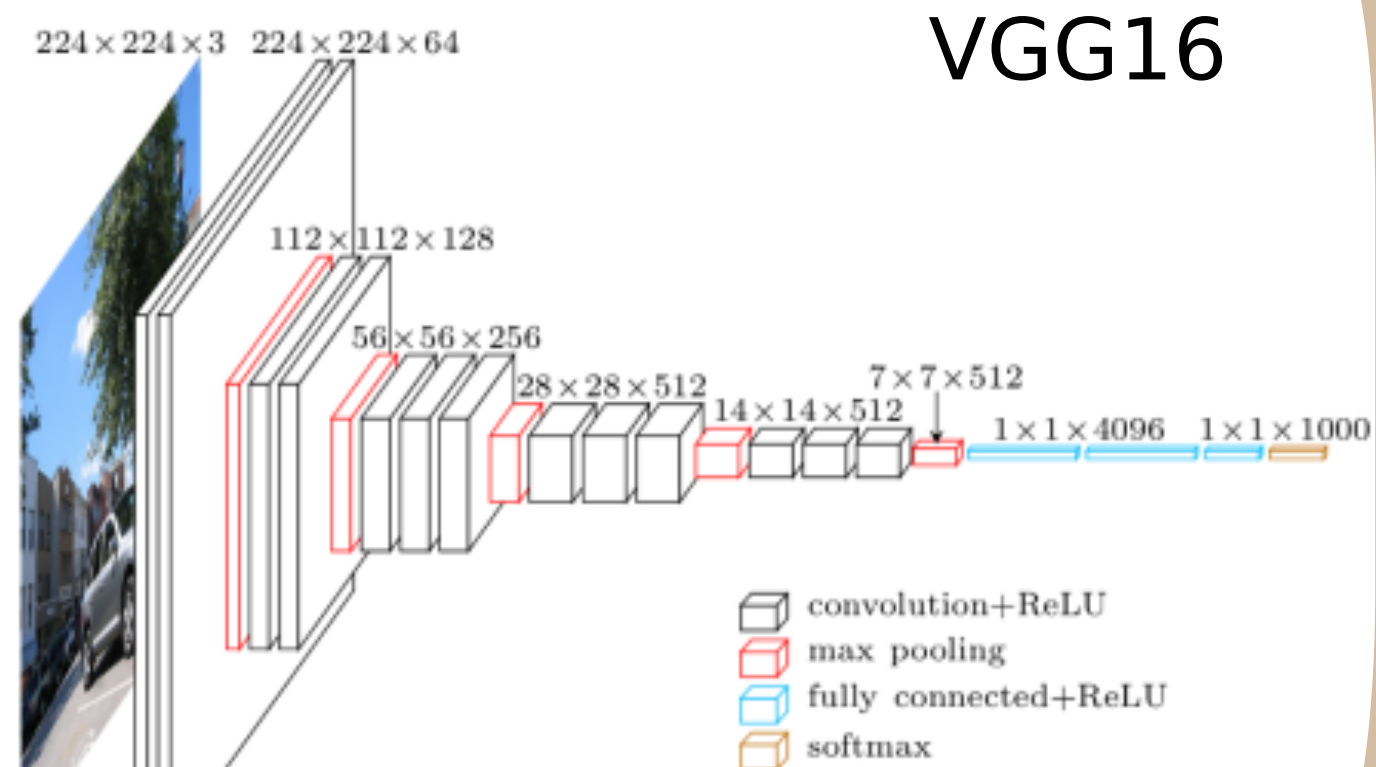Prediction ↑ | ↑ Class probabilities | Class probabilities ↑ | Prediction ↑

Text CNN

wait
for
the
video
and
do
n't
rent
it

VGG16

$224 \times 224 \times 3$ $224 \times 224 \times 64$

$112 \times 112 \times 128$

$56 \times 56 \times 256$

$28 \times 28 \times 512$

$14 \times 14 \times 512$

$7 \times 7 \times 512$

$1 \times 1 \times 4096$ $1 \times 1 \times 1000$

convolution+ReLU
max pooling
fully connected+ReLU
softmax

n x k representation of
sentence with static and
non-static channels

Convolutional layer with
multiple filter widths and
feature maps

Max-over-time
pooling

Fully connected layer
with dropout and
softmax output

Text

Image