Как я решал и как я не решал конкурс Avito

Евгений Нижибицкий

ВМК МГУ

12 ноября 2014 г.

1. Как я решал Avito
   - Исходные данные
   - Решение
   - Результат

2. Как я не решал Avito
   - «Классическое» компьютерное зрение
   - Сверточные сети
   - Unsupervised feature learning

Как я решал Avito

Исходные данные


ПИЛОМАТЕРИАЛЫ-СПБ
от производителя
тел.: 983-32-43, 942-74-44
ЛУЧШИЕ ЦЕНЫ!
www.pilomaterial-spb.ru
E-MAIL: zakaz@pilomaterial-spb.ru
Бесплатная доставка от 35 куб.м
AVITO.ru


гарантия
100%
результата
АВТО
ЮРИСТ
8(3462) 60-83-75, 72-34-14

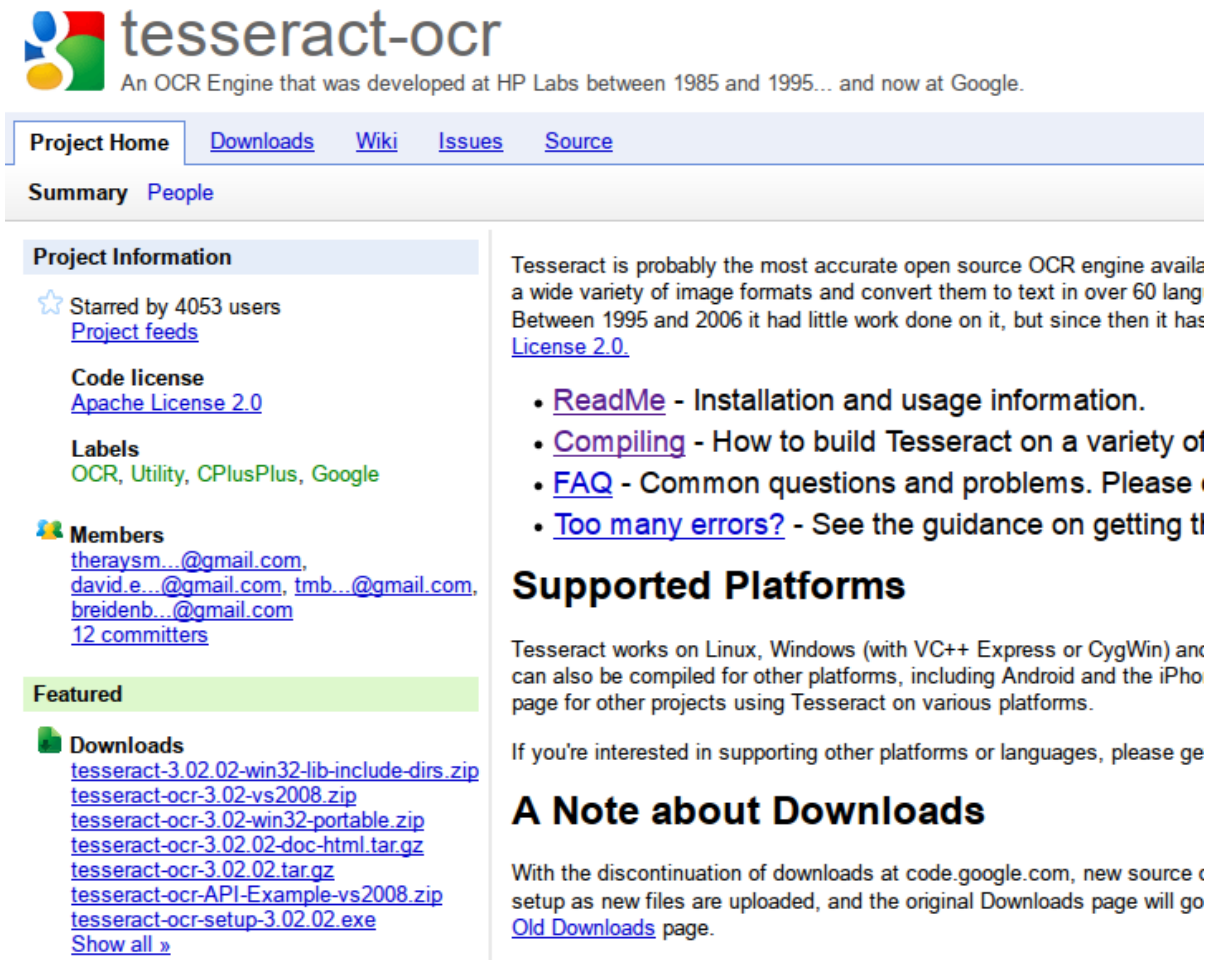
Как я решал Avito

Исходные данные

Как я решал Avito

Исходные данные

Очевидное решение — давайте распознавать текст!



Что плохого, к примеру, в поиске доменов?

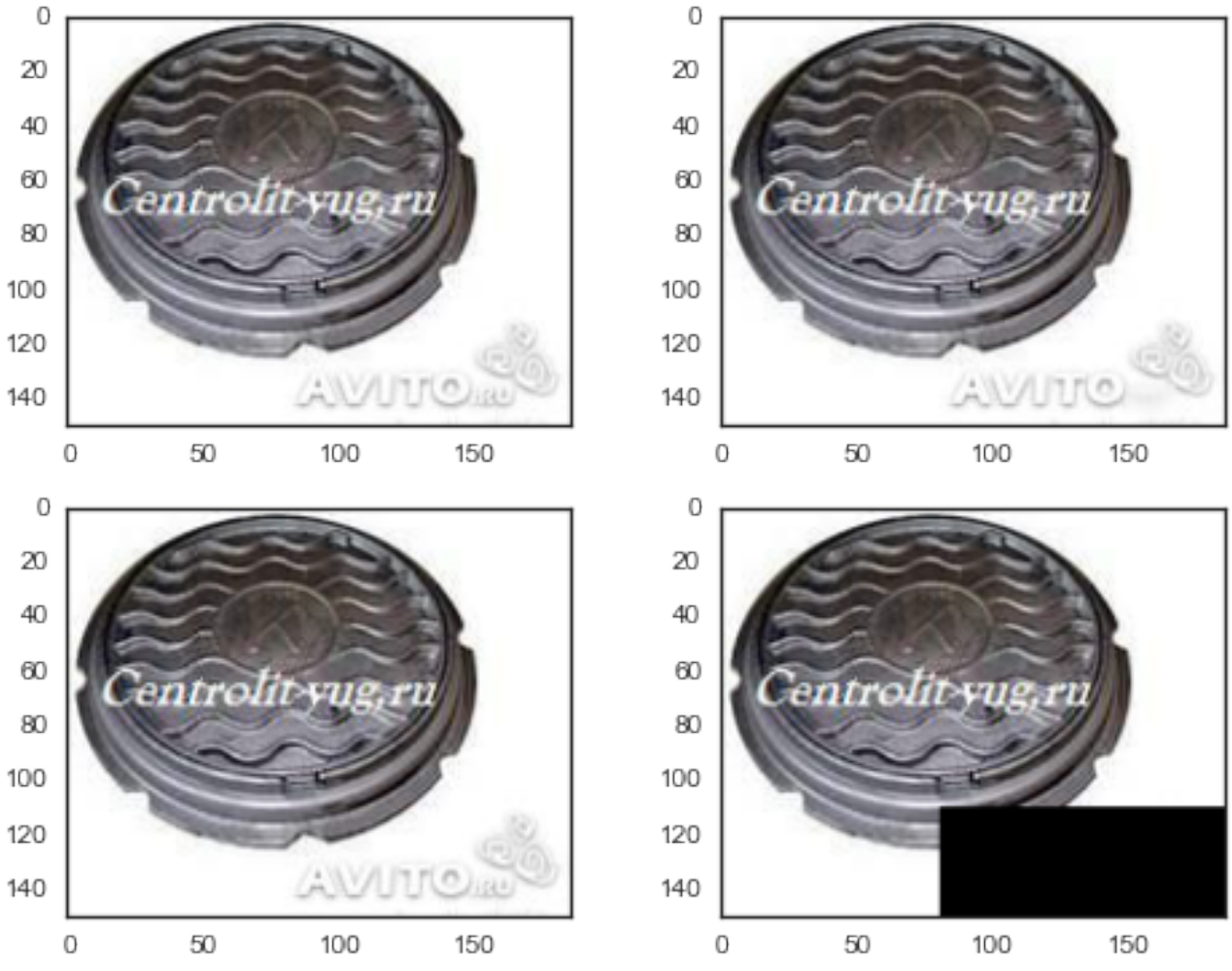
Что плохого, к примеру, в поиске доменов?

Давайте замажем!

```
plt.figure(figsize=(8,6))
im = Image.open('train/129641663_929522737.jpg')
plt.subplot(2,2,1)
plt.imshow(im)
imc = im.crop((im.size[0] - 37, im.size[1] - 15, im.size[0] - 19, im.size[1] - 3))
imc = imc.filter(ImageFilter.MedianFilter(9))
im.paste(imc, (im.size[0] - 37, im.size[1] - 15))
plt.subplot(2,2,2)
plt.imshow(im)

im = Image.open('train/129641663_929522737.jpg')
plt.subplot(2,2,3)
plt.imshow(im)
imc = Image.new('RGB', (105, 40))
im.paste(imc, (im.size[0] - 105, im.size[1] - 40))
plt.subplot(2,2,4)
plt.imshow(im)
plt.show()
```

Давайте замажем!



Результат OCR

In [73]:
```
img = Image.open('blur/130060451_663770998.jpg')

print "tesseract:"
tool = pyocr.get_available_tools()[0]
langs = tool.get_available_languages()
rus = langs[2]
eng = langs[3]
print tool.image_to_string(img, lang=eng, builder=pyocr.builders.TextBuilder())

print "\ncuneiform:"
tool = pyocr.get_available_tools()[1]
langs = tool.get_available_languages()
rus = langs[3]
eng = langs[0]
print tool.image_to_string(img, lang=eng, builder=pyocr.builders.TextBuilder())
```

```
tesseract:
CoaaaeM Kpacmable CeMefiHble
m noKyMeHTanthxe dpMnbeL
Bbl a maan pom.
Wedding wdeo
Farmly vrdeo
vmeo sdmng servmes

www.5fpro. ru

cuneiform:
Coaoaeru ffpacwet ie cebfeaabie
N AourMSHTBJlbubie rpaflbMbr
Bbr e roaeuoa pooa


www sfpro ru
```

Результат OCR

	Id	label	tesseract	txtlen
16668	106555603_200745813	0	АМП TAVE T	10
15992	107731247_203283026	0	| Ад! ‹ шита ЪЁЁI AM! (WWW/g 2’45	36
12124	109780842_207678716	0	Рапазопбс п ь▷ ...	151
238	109875184_207885561	0	у % ПЛИТ ;V e02 “NT ’4	22
7074	110132756_208424448	0	5000руб. 50006 {бкВт „ ЛУИ Т©5000ру...	82

Всего **30%** изображений с нарушениями

Всего **21%** изображений с обнарежнием текста

На примерах без нарушений **15%** обнаружений текста

На примерах с нарушениями **36%** обнаружений текста

Из изображений с обнаружением текста **49%** с нарушениями

Ищем паттерны со 100% Precision:

- Номера: (XX, XX), -XX, XXX + ['8(', '8 (', '(8', '8-8', '+79']
- Части слов из корпусов (без крайних букв)
- Паттерны, придуманные вручную

```
urls = ['.ru', 'vk.', 'http', ':/', '.net', u'.рф', '.ua']
emails = ['@ma', 'e-ma', '@gm']
words = [u'монта', u'тел.', u'звонит', u'кред', u'автом', u'стоим', u'адрес',
         u'конт', u'москва', u'бург', u'красн', u'продае', u'окна', u'окон',
         u'shop', u'под ключ', u'слуг', u'недв']

test = ['abcde']
patterns = test + urls + emails + words # + telephones

scores = []
for pattern in patterns:
    scores.append(find(pattern))
scores = np.array(scores)
for tup in sorted(scores, key=lambda score: float(score[1]), reverse=True):
    print "{}\t{}\t{}".format(tup[0], tup[1], tup[2])
```

```
.ru     262     1.0
тел.    76      1.0
vk.     73      1.0
звонит  65      1.0
:/      54      1.0
монта   46      1.0
http    45      1.0
слуг    43      1.0
недв    38      1.0
@ma     26      1.0
```

Находим около 20% нарушений (100% Precision, 20% Recall)

```
patterns = urls + emails + words + parts + telephones
patterns = list(set(patterns))

ytrue = train['label']
ypred = np.zeros_like(train['label'])
for pattern in patterns:
    ypred[np.array(train['tesseract'].apply(lambda s:
                                            s.lower().encode("utf-8")
                                            .find(pattern.encode("utf-8")) > 0))] += 1
print sum(ypred > 0)
print auc(ytrue, ypred)
print acc(ytrue[ypred > 0], ypred[ypred > 0]*0 + 1)
```

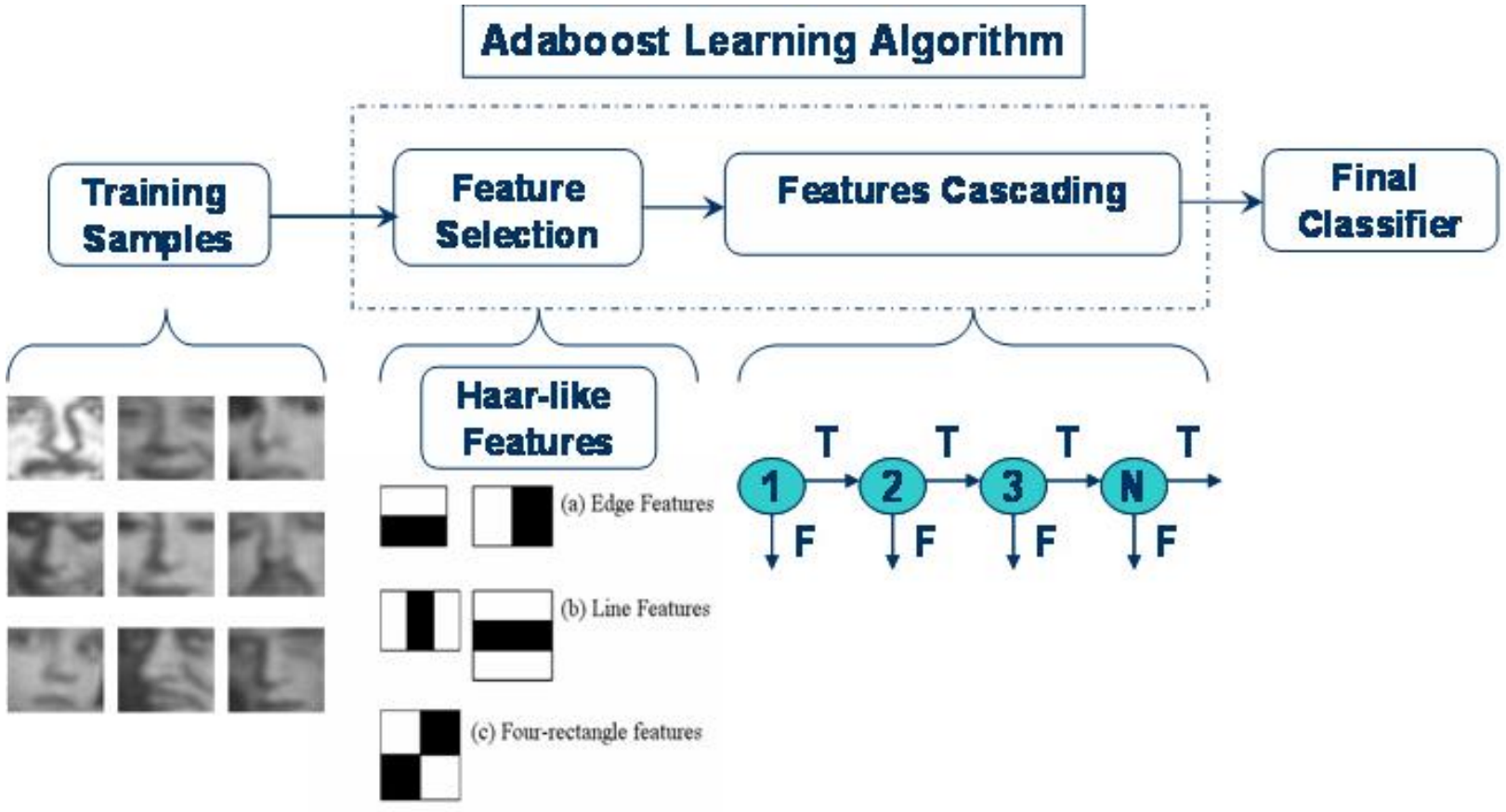
```
2377
0.59878646829
1.0
0.594981166825
```

А сколько (не)реально вообще найти?

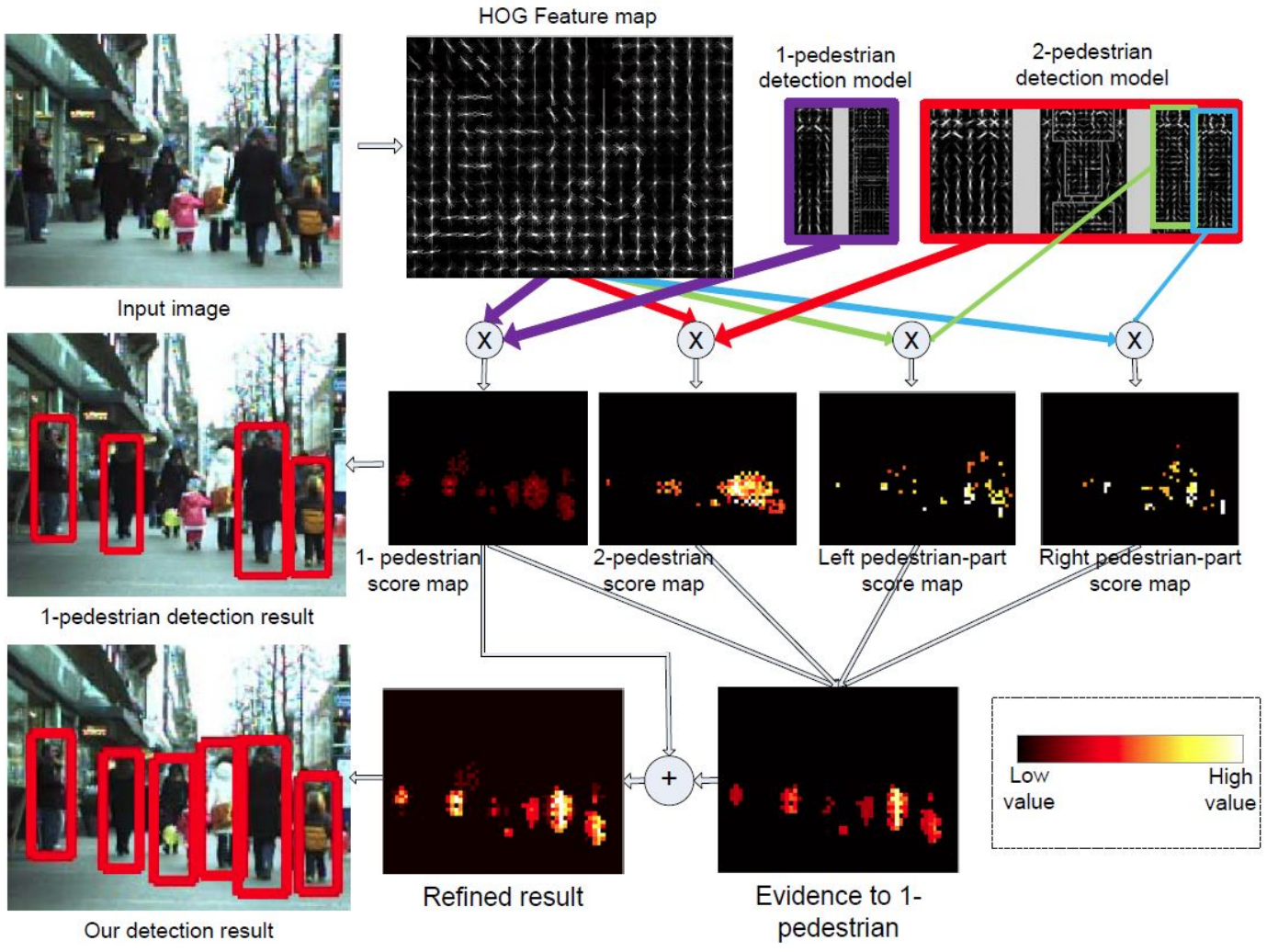
```
ytrue = train['label']
ypred = np.zeros_like(train['label'])
ypred[np.array((train['label'] == 1) & (train['txtlen'] > 0))] = 1
print sum(ypred)
print auc(ytrue, ypred)
print acc(ytrue[ypred > 0], ypred[ypred > 0])
```

```
4408
0.683193417006
1.0
```

Алгоритм Виолы-Джонса



Алгоритм HOG+SVM



AlexNet

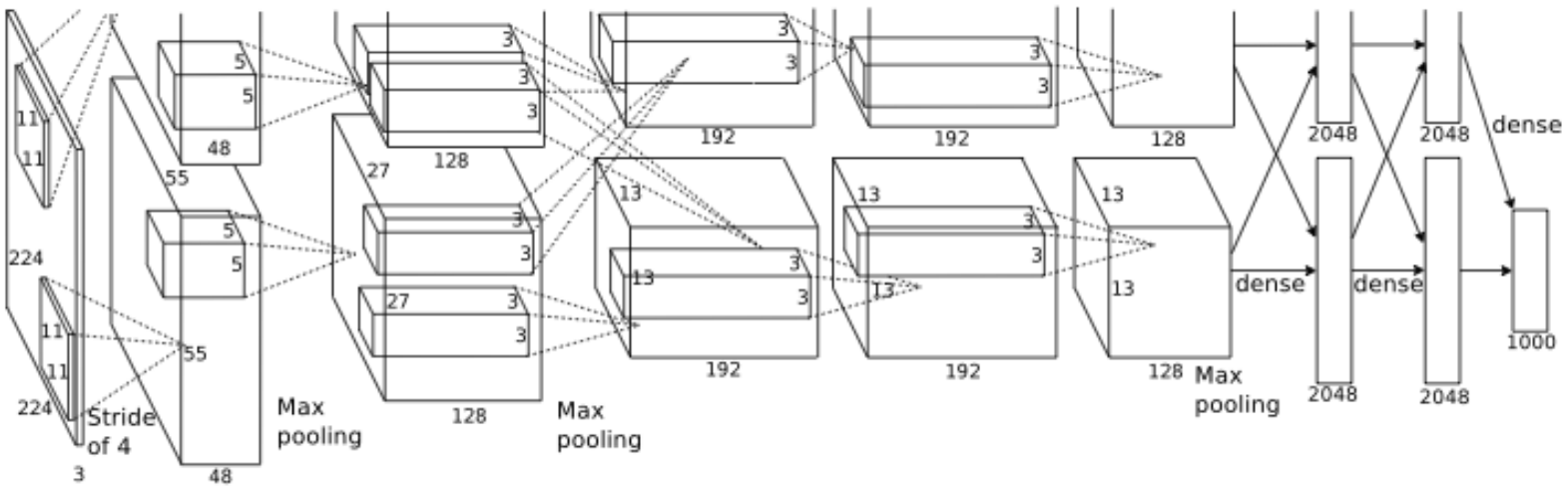
ImageNet Classification with Deep Convolutional Neural Networks [NIPS 2012]

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

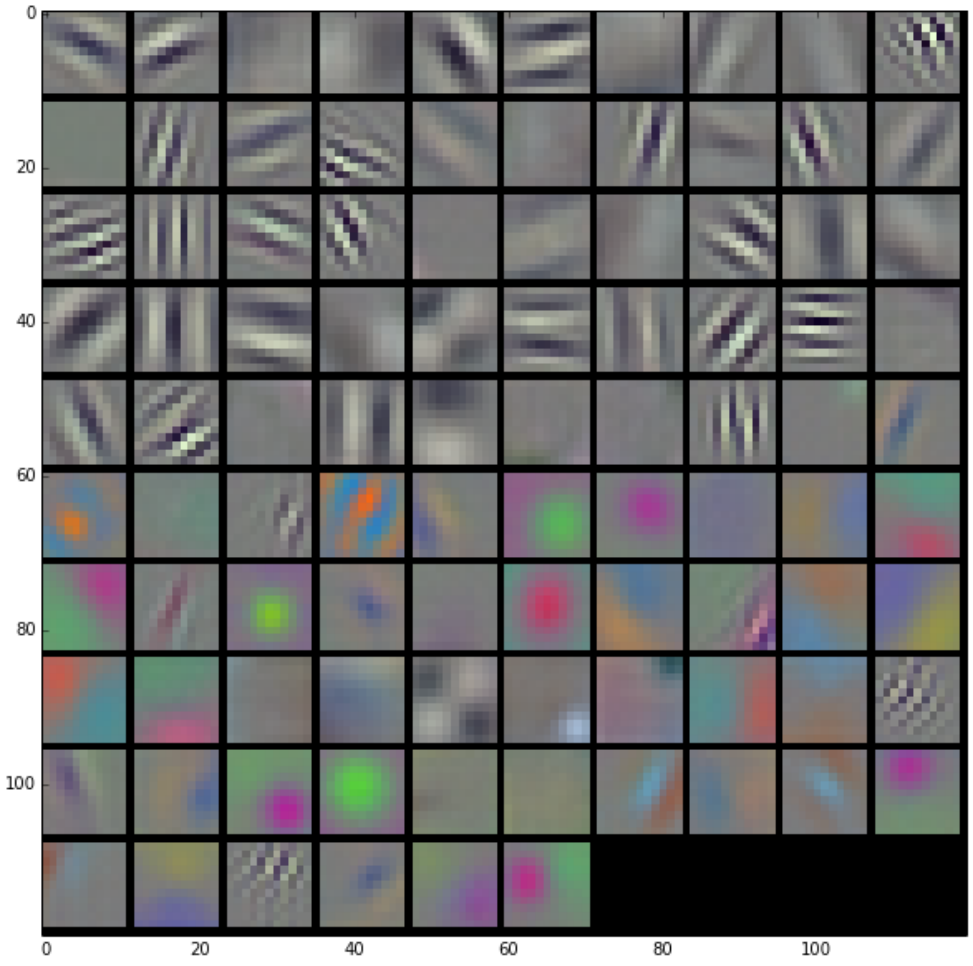
Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

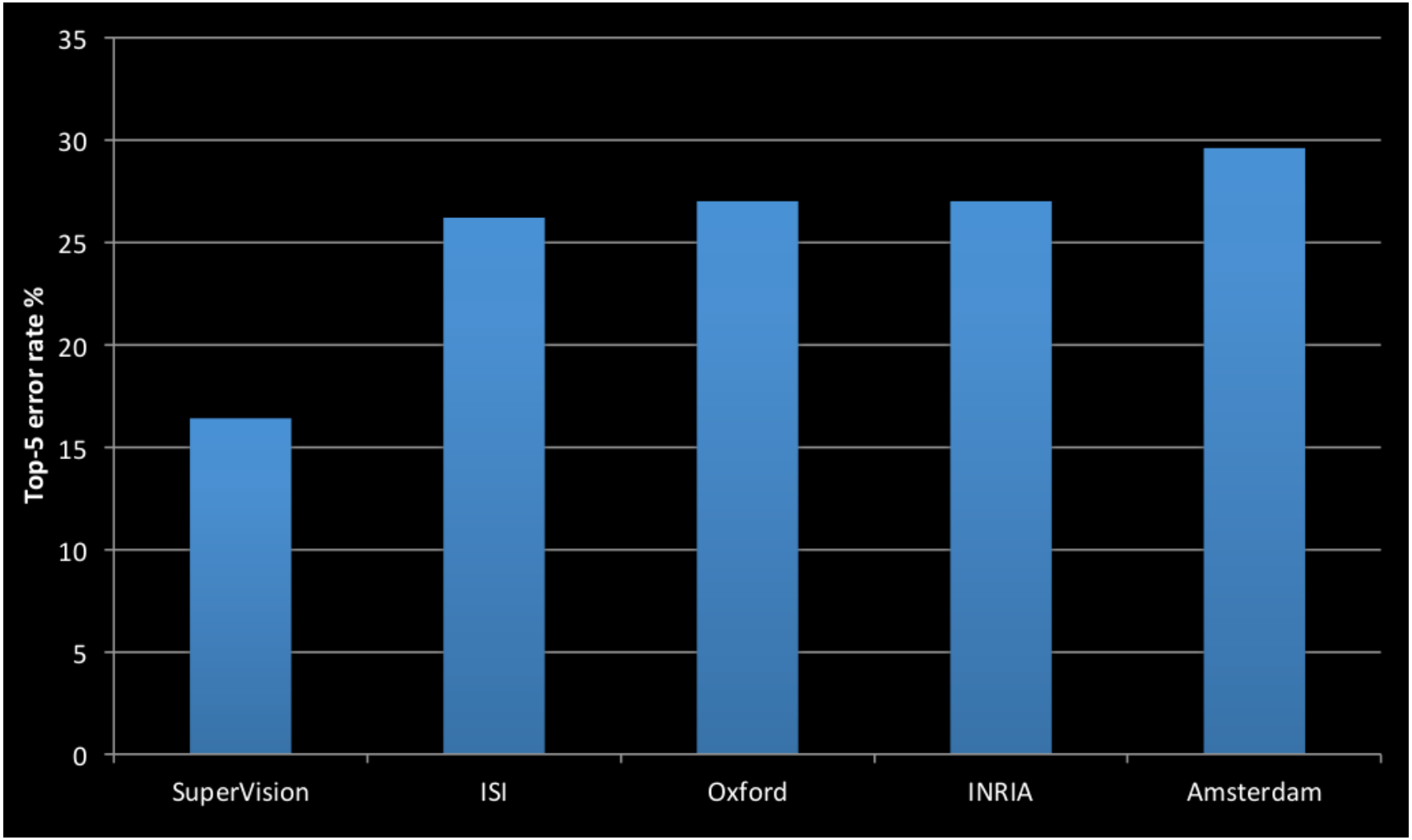
AlexNet


224
224
3
11
11
Stride
of 4
55
48
5
5
55
48
Max
pooling
27
27
128
128
3
3
3
3
3
3
Max
pooling
13
13
13
13
192
192
192
192
128
128
3
3
3
3
3
3
dense
dense
dense
2048
2048
2048
2048
1000
Max
pooling

AlexNet

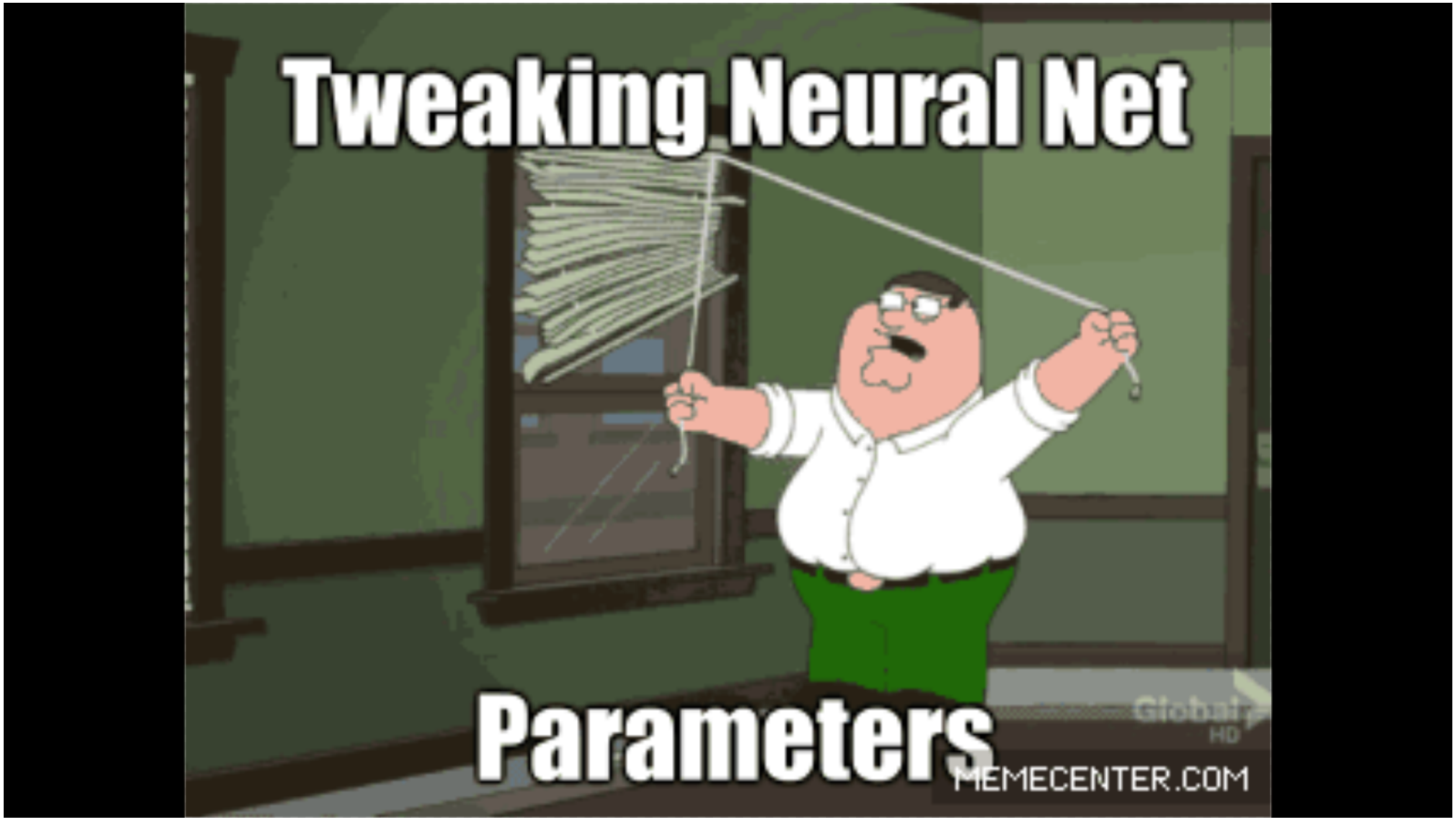


AlexNet


Top-5 error rate %
35
30
25
20
15
10
5
0
SuperVision
ISI
Oxford
INRIA
Amsterdam

Почему может не заработать?

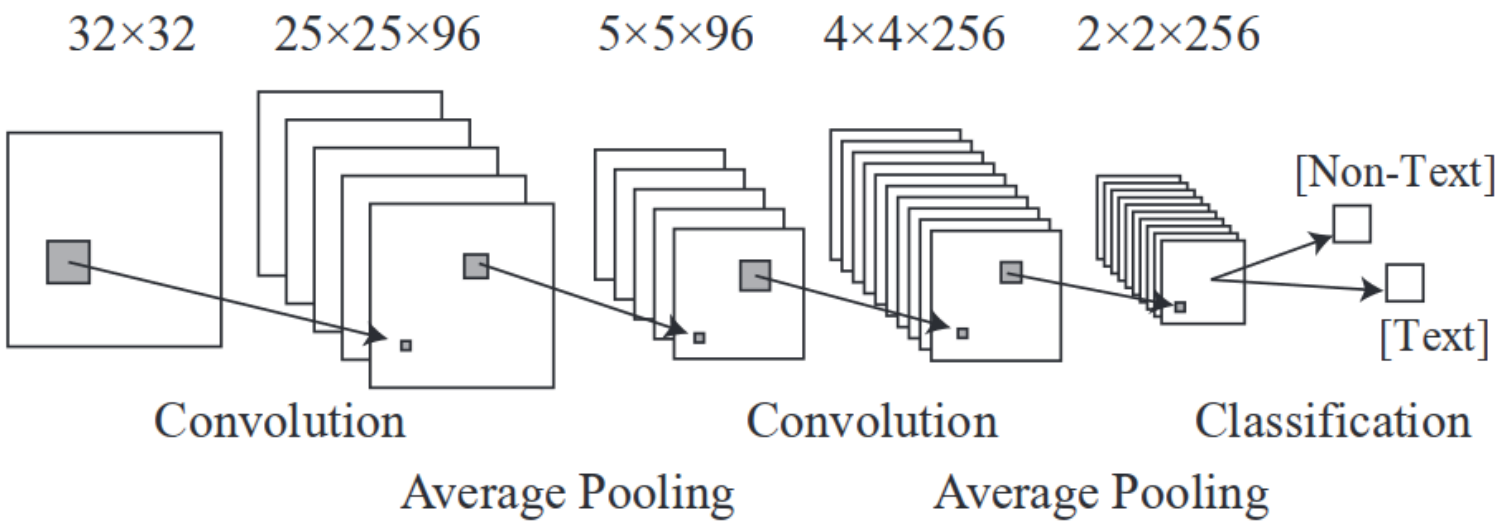
Почему может не заработать?



Обучаем первый уровень сами

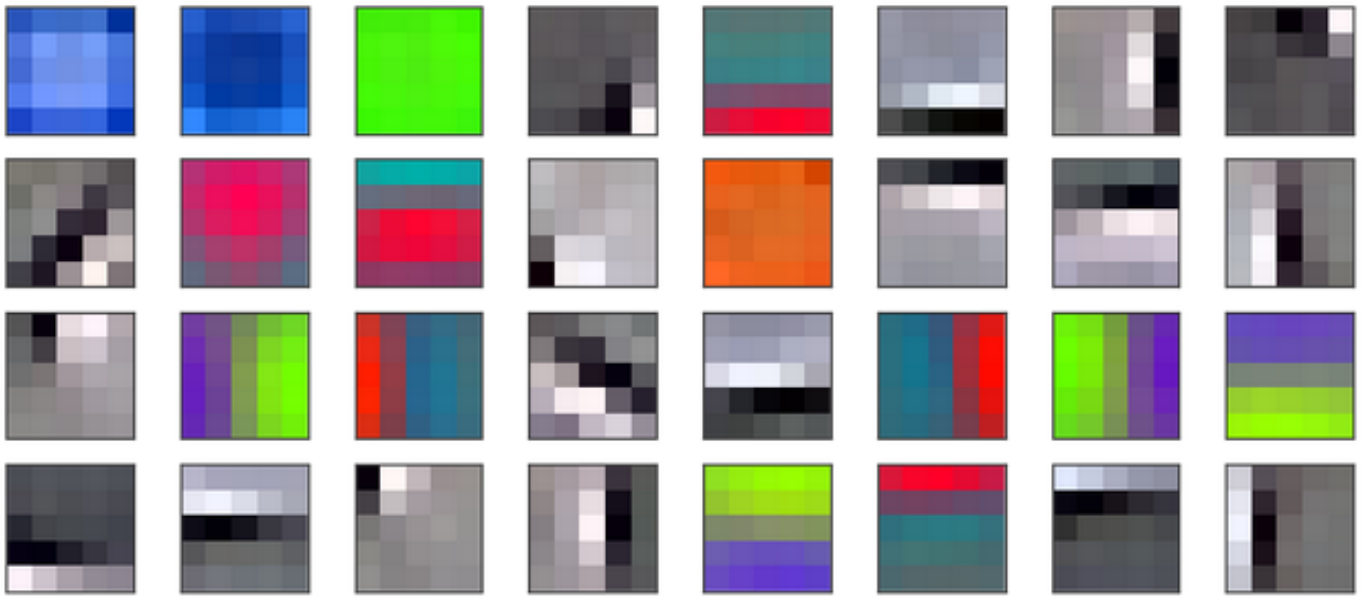
End-to-End Text Recognition with Convolutional Neural Networks

Tao Wang* David J. Wu* Adam Coates Andrew Y. Ng
Stanford University, 353 Serra Mall, Stanford, CA 94305
{*twangcat, dwu4, acoates, ang*}*@cs.stanford.edu*



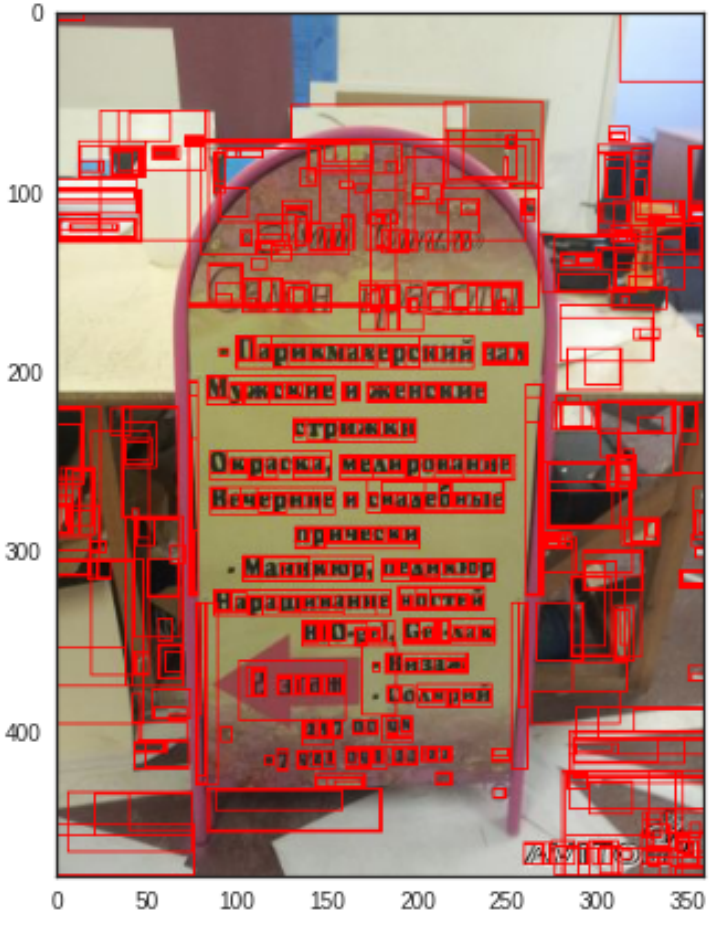
Пример на питоне

```
# display
fig = plt.figure(figsize=(14, 6))
num_col = int(np.ceil(float(NUM_FILTERS)/4))
for i in xrange(NUM_FILTERS):
    ax = fig.add_subplot(4, num_col, i+1)
    filter_ = filters[i,...]
    filter_ -= filter_.min()
    filter_ /= filter_.max()
    ax.imshow(filter_, interpolation='none')
    ax.get_xaxis().set_visible(False)
    ax.get_yaxis().set_visible(False)
plt.show()
```



Регионы для проверки ищем с помощью MSER

```
im = imread("0505.jpg")
mim = mser(im, debug=True)
```



Привет!