Cześć!

NLP

natural language processing

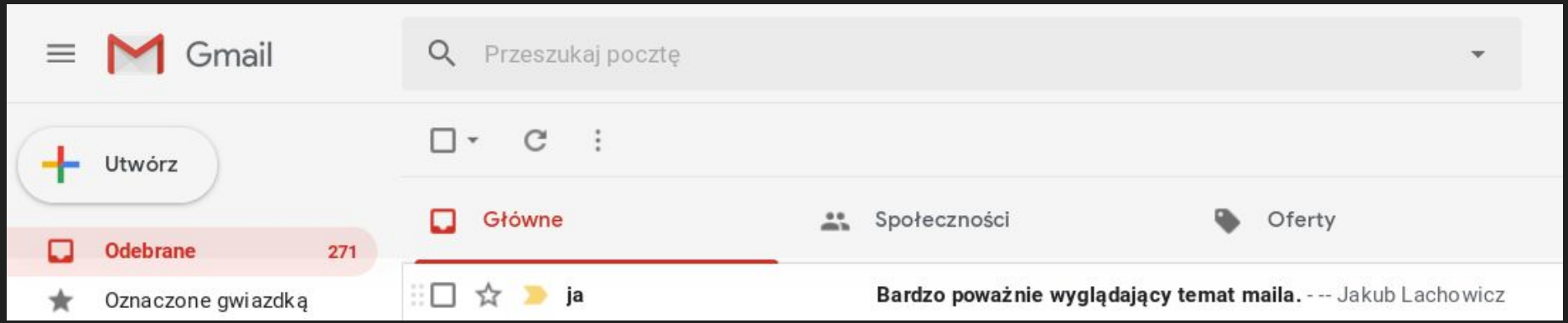
Sentiment Analysis

“Czas realizacji zamówienia zbyt długi. Niedbale spakowana przesyłka.”

“Czas realizacji zamówienia zbyt długi. Niedbale spakowana przesyłka.”

“Profesjonalna obsługa. Towar dotarł szybko i zgodnie z zamówieniem.”

Text Classification


Gmail
Przeszukaj pocztę
Utwórz
Odebrane 271
Oznaczone gwiazdką
Główne
Społeczności
Oferty
ja
Bardzo poważnie wyglądający temat maila. - -- Jakub Lachowicz

Kenneth Herron (Reporter)

Description • 13 years ago

This was found through a coverity scan of the firefox source code. Please refer
to the sample URL.

At lines 323-325, |pt_PostNotifyToCvar| checks |notified->link| for NULL, calls |PR_NEWZAP| to allocate a _PT_Notified structure, and assigns the resulting pointer to |notified|. |PR_NEWZAP| wraps |PR_Calloc|, which may call |calloc|, which may return NULL. This allocation isn't checked so |notified| may contain NULL after line 325.

Jak komputer “rozumie” tekst?

Bag Of Words

Czyli metody zliczające

```
1       from sklearn.feature extraction.text import CountVectorizer
```

```
1       from sklearn.feature_extraction.text import CountVectorizer
2       corpus = [
3            'This is the first document.',
4            'This document is the second document.',
5            'And this is the third one.',
6            'Is this the first document?',
7       ]
```

```
1       from sklearn.feature_extraction.text import CountVectorizer
2       corpus = [
3            'This is the first document.',
4            'This document is the second document.',
5            'And this is the third one.',
6            'Is this the first document?',
7       ]
8       vectorizer = CountVectorizer()
9       X = vectorizer.fit transform(corpus)
```

```
1       from sklearn.feature_extraction.text import CountVectorizer
2       corpus = [
3            'This is the first document.',
4            'This document is the second document.',
5            'And this is the third one.',
6            'Is this the first document?',
7       ]
8       vectorizer = CountVectorizer()
9       X = vectorizer.fit_transform(corpus)
10      print(vectorizer.get_feature_names())
11      ['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third', 'this']
```

```
from sklearn.feature_extraction.text import CountVectorizer
corpus = [
     'This is the first document.',
     'This document is the second document.',
     'And this is the third one.',
     'Is this the first document?',
]
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(corpus)
print(vectorizer.get_feature_names())
['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third', 'this']
print(X.toarray())
[[0, 1, 1, 1, 0, 0, 1, 0, 1]
 [0, 2, 0, 1, 0, 1, 1, 0, 1]
 [1, 0, 0, 1, 1, 0, 1, 1, 1]
 [0, 1, 1, 1, 0, 0, 1, 0, 1]]
```

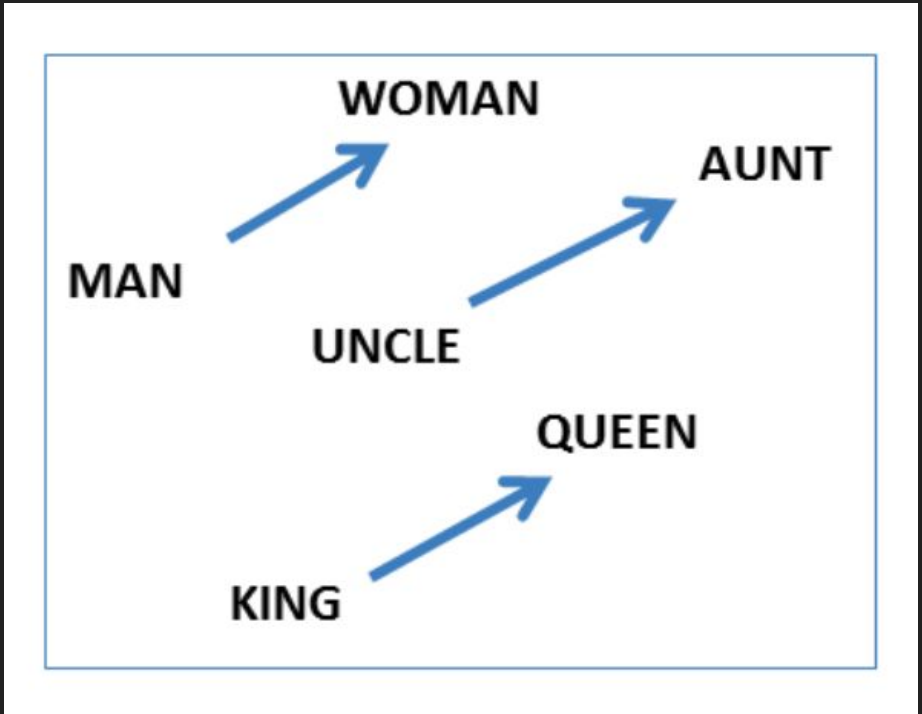
“Produkt mnie nie rozczarował i cieszę się z zakupu.”

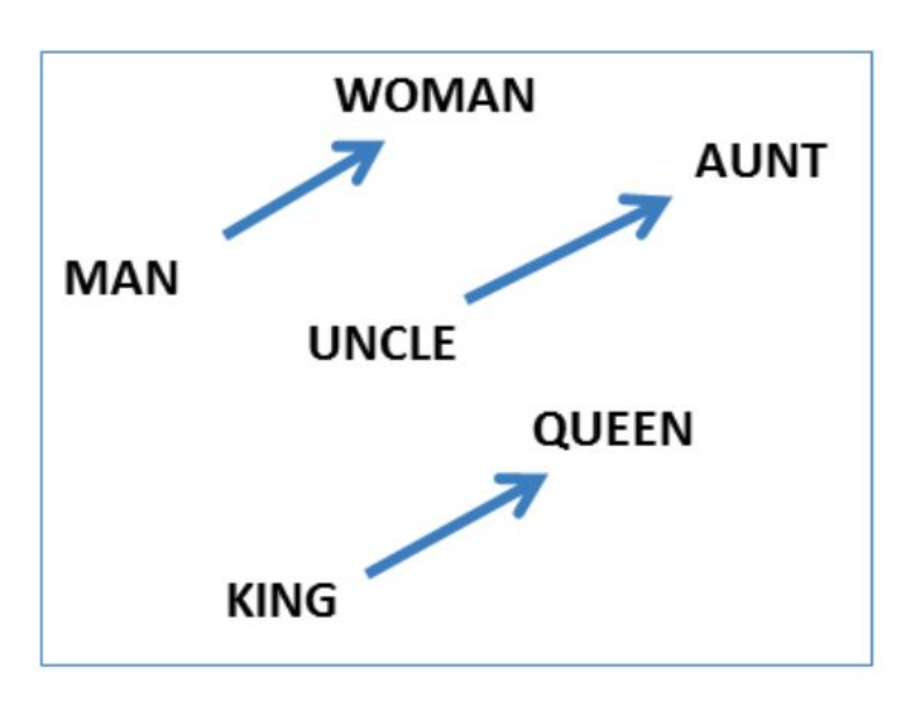
“Produkt mnie nie rozczarował i cieszę się z zakupu.”

“Produkt mnie rozczarował i nie cieszę się z zakupu.”

Word embeddings

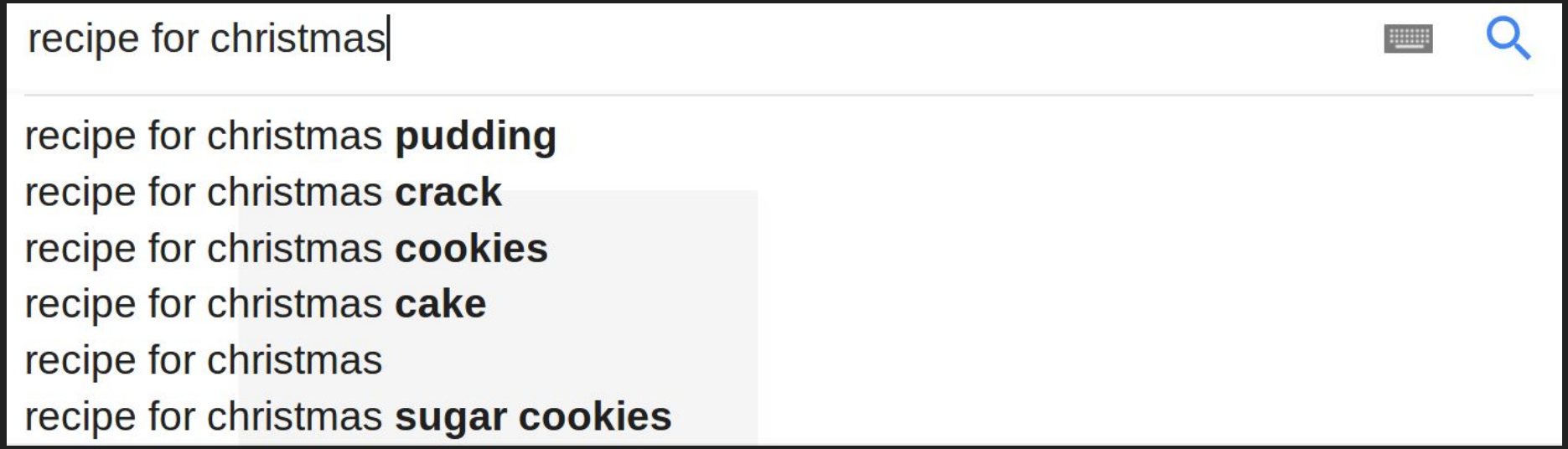
```
w -0.623933 2.184556 -2.251491 -1.645580 -0.536714
i -0.927808 0.920390 -1.175119 -1.256960 -1.419652
na 0.149684 3.621207 -3.882869 -2.175517 -0.109260
z 0.727299 2.601258 -3.080432 -3.438949 0.080732
się -0.777044 6.364845 0.160861 1.954626 1.527357
nie -1.945414 5.001998 0.304228 1.160726 1.649076
do -0.868247 2.539189 -3.158081 -0.869770 -0.214892
to 0.721761 1.916624 0.537992 0.418699 0.408665
że -4.128553 5.105642 -1.619554 -0.684059 0.044076
```


WOMAN
AUNT
MAN
UNCLE
QUEEN
KING



King – Man + Woman = ?

Language Model


recipe for christmas
recipe for christmas pudding
recipe for christmas crack
recipe for christmas cookies
recipe for christmas cake
recipe for christmas
recipe for christmas sugar cookies

Zdanie: Wczoraj byłem w kinie i obejrzałem film.

1. P(Wczoraj|<bos>)
2. P(Byłem|<bos> Wczoraj)
3. P(w|<bos> Wczoraj byłem)
4. P(kinie|<bos> Wczoraj byłem w)

…

7. P(film|<bos> Wczoraj byłem w kinie i obejrzałem)

<bos> - “begin of the sentence”

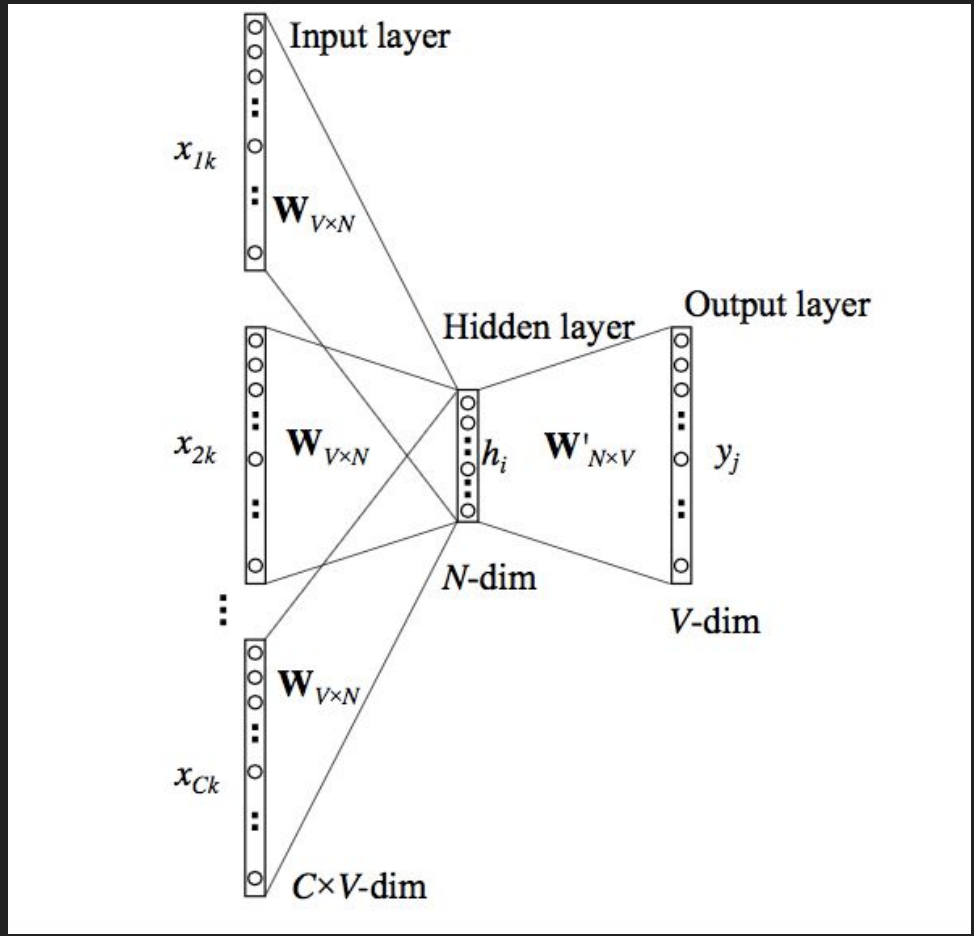
Zdanie: Wczoraj byłem w kinie i obejrzałem film.

1. P(Wczoraj|<bos>)
2. P(Byłem|<bos> Wczoraj)
3. P(w|<bos> Wczoraj byłem)
4. P(kinie|<bos> Wczoraj byłem w)

…

7. P(film|<bos> Wczoraj byłem w kinie i obejrzałem)

<bos> - “begin of the sentence”



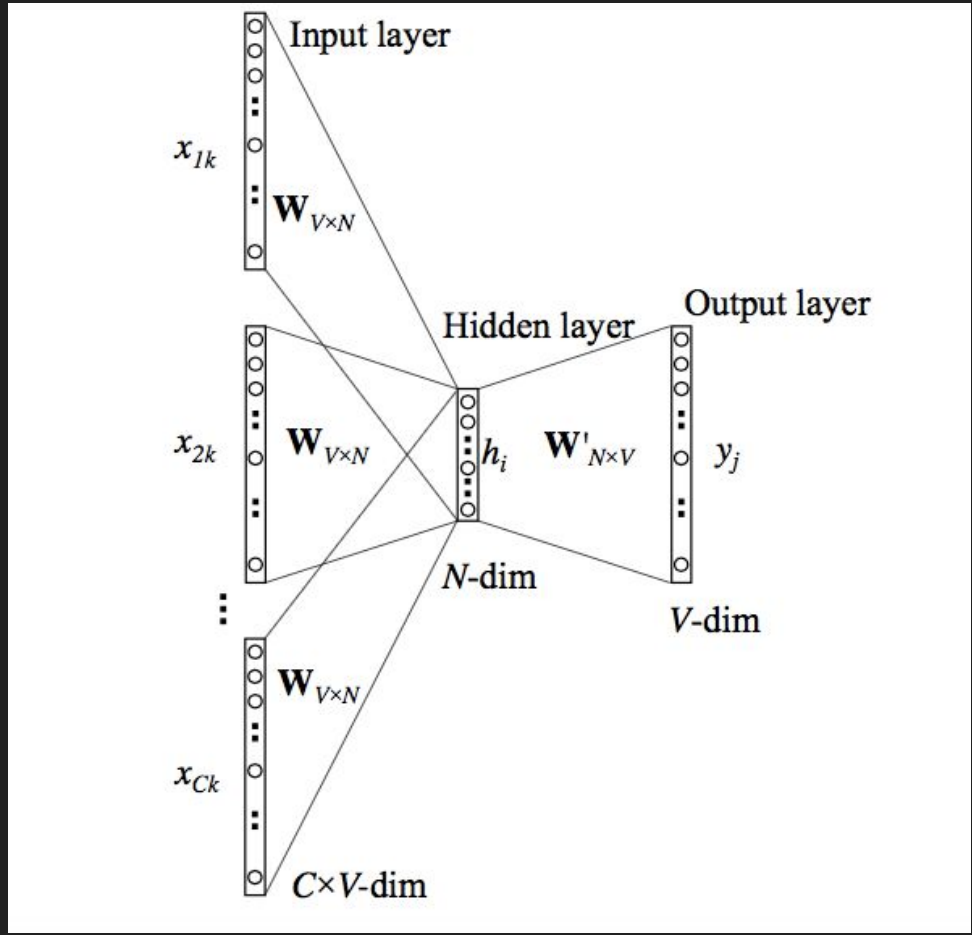
Zdanie: Wczoraj byłem w kinie i obejrzałem film.

1. P(Wczoraj|<bos>)
2. P(Byłem|<bos> Wczoraj)
3. P(w|<bos> Wczoraj byłem)
4. P(kinie|<bos> Wczoraj byłem w)

…

<bos> - “begin of the sentence”

W - macierz zawierająca word embeddings. Wymiar VxN
V - to liczba wszystkich słów których używamy
N - to wymiar wektora dla 1 słowa.



dsmodels.nlp.ipipan.waw.pl

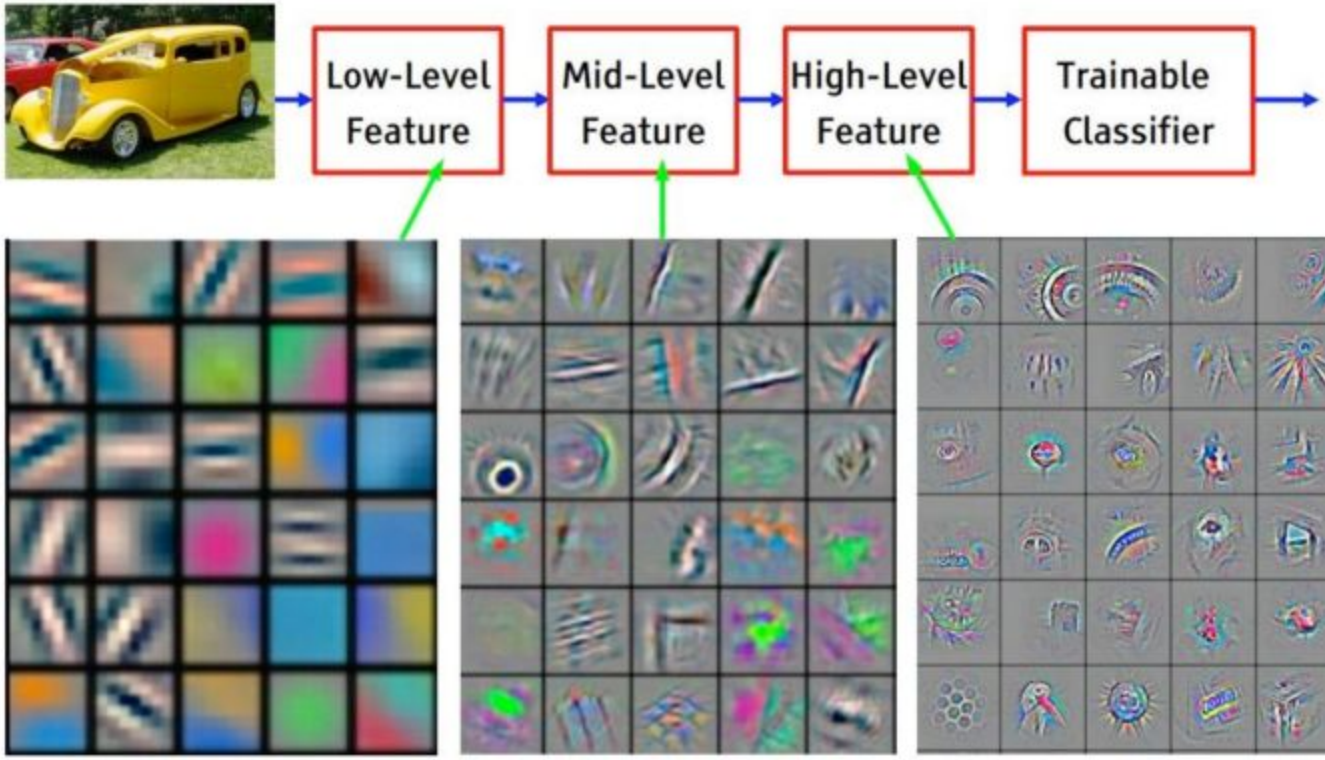
Word Embeddings + LSTM:

github.com/Ermlab/pl-sentiment-analysis

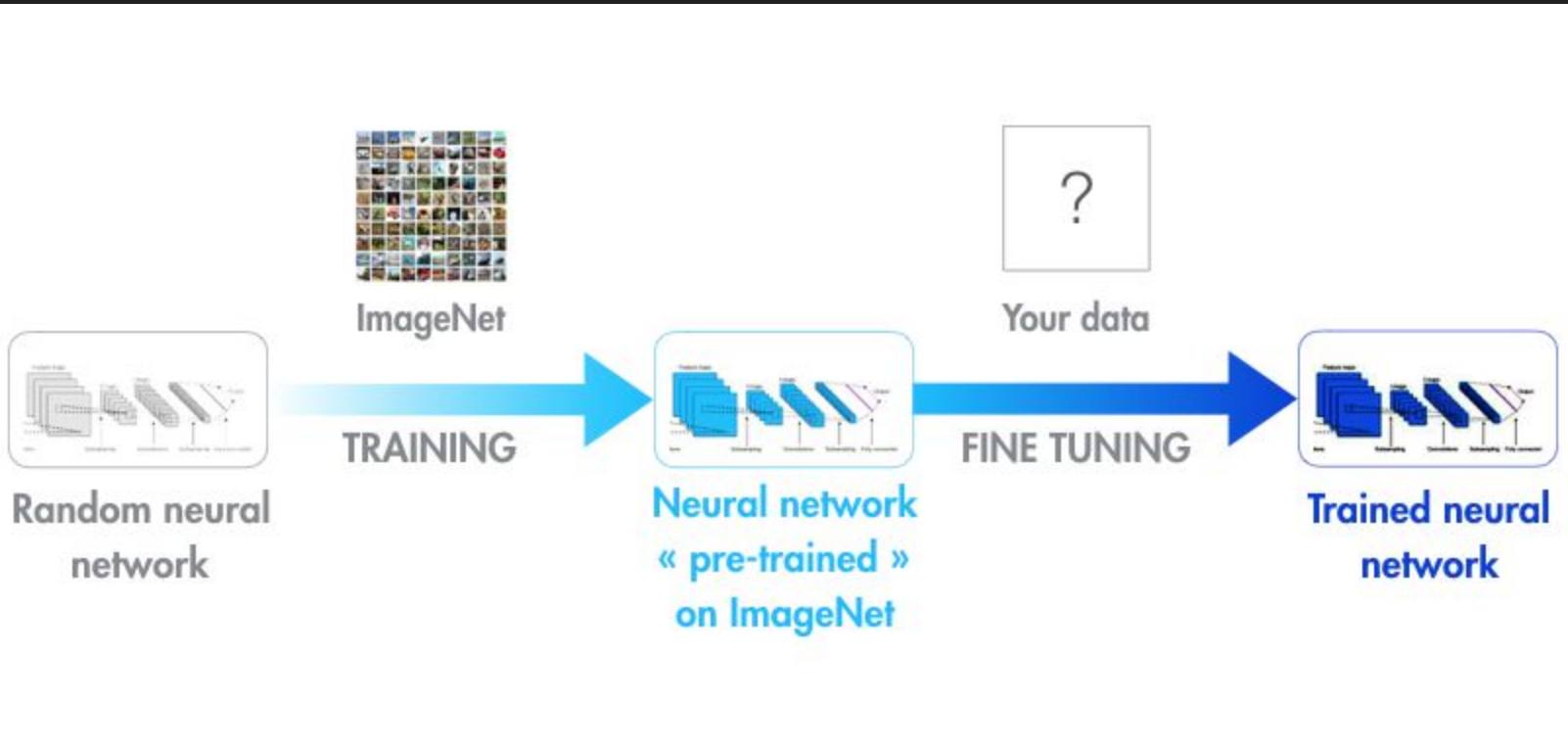
3-5 dni na CPU

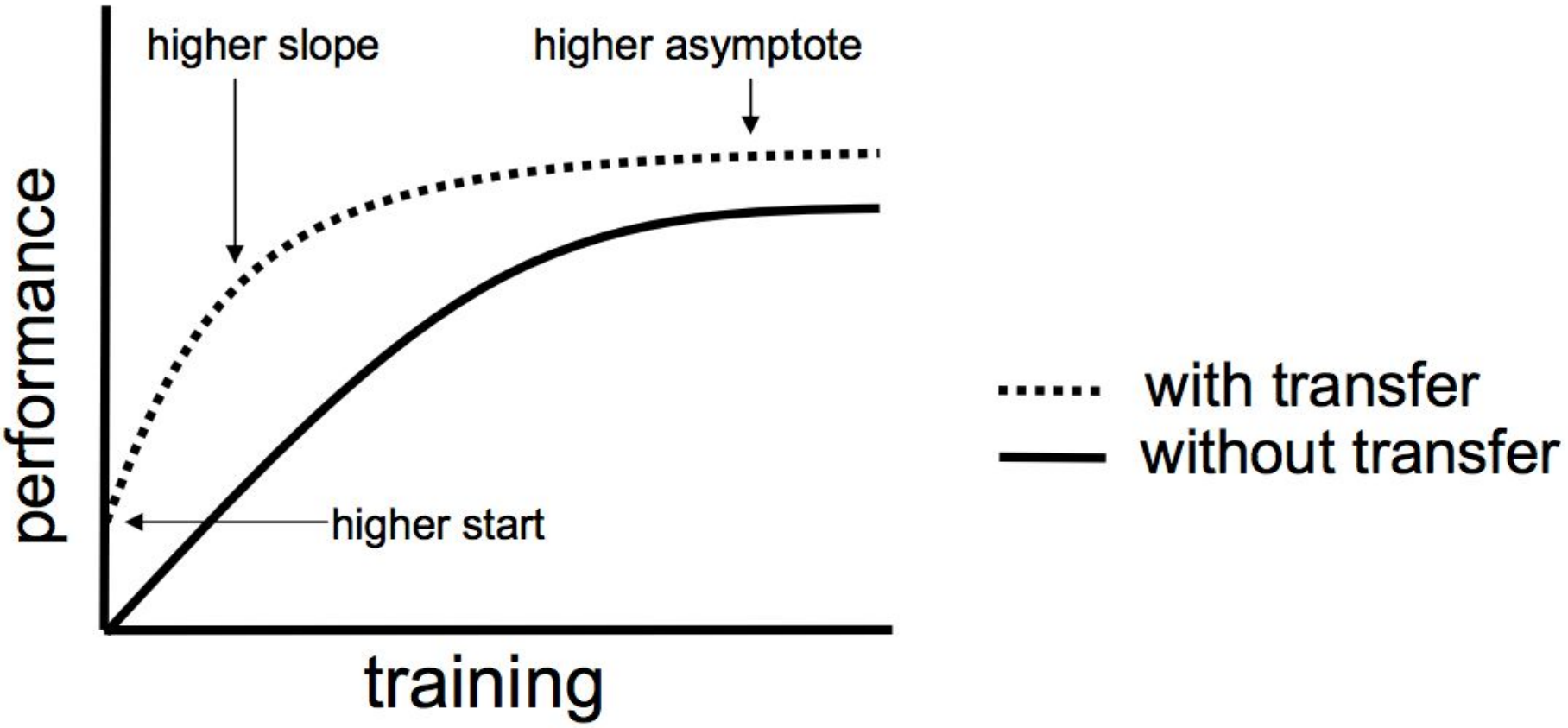
½ dnia na GPU

Transfer Learning



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]


ImageNet
?
Your data
TRAINING
FINE TUNING
Random neural network
Neural network « pre-trained » on ImageNet
Trained neural network


higher slope
higher asymptote
performance
higher start
training
with transfer
without transfer

Word embeddings

to też transfer learning

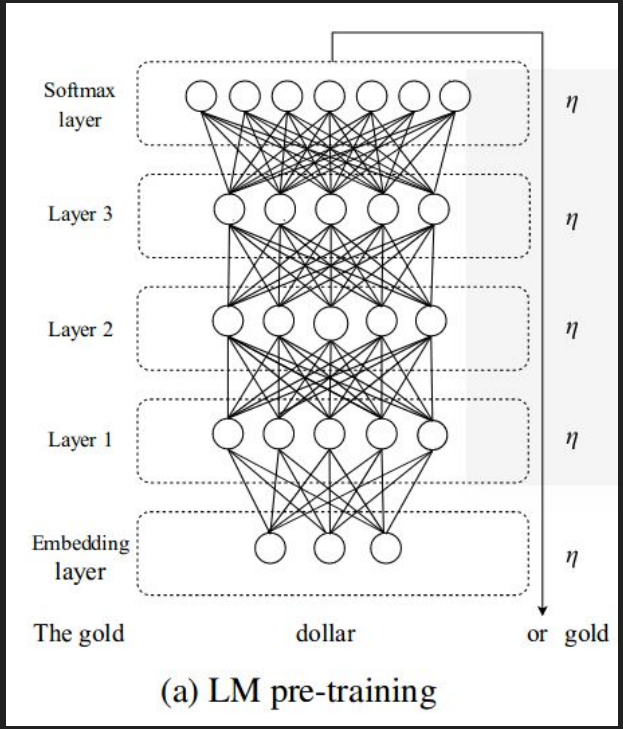
Word embeddings

to też transfer learning
ale da się to zrobić lepiej

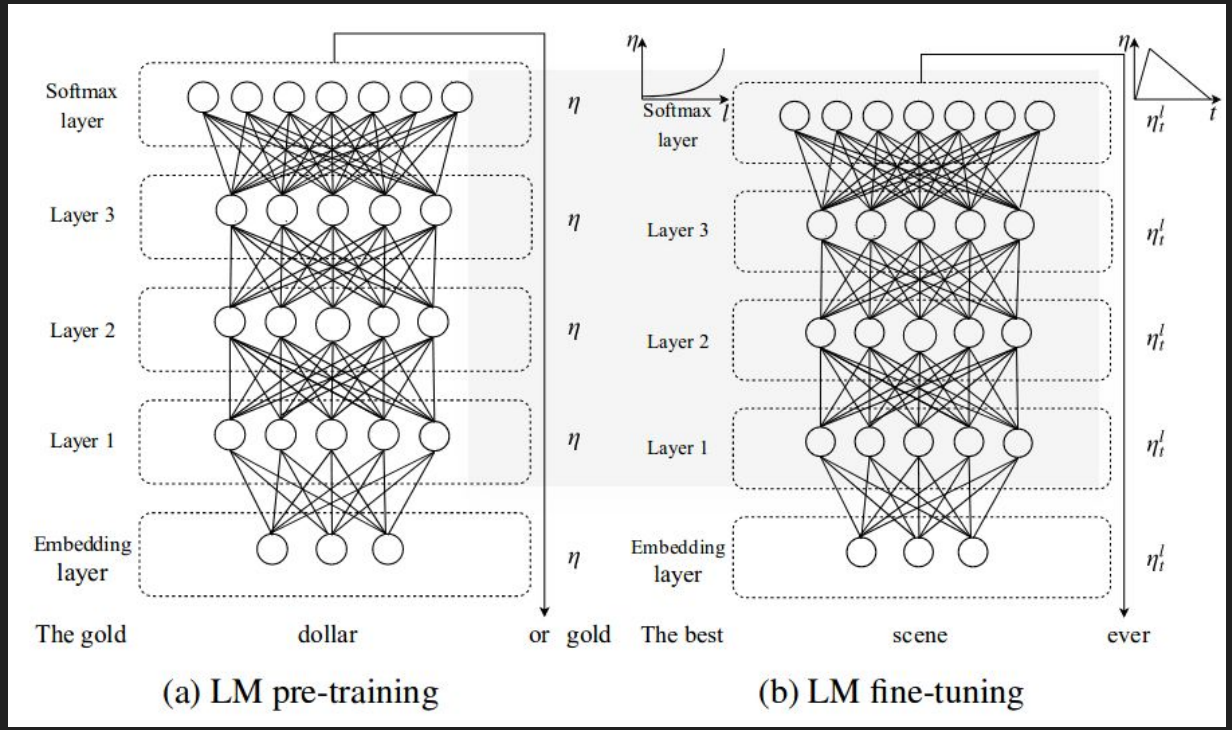
“Całymi dniami gra na komputerze.”

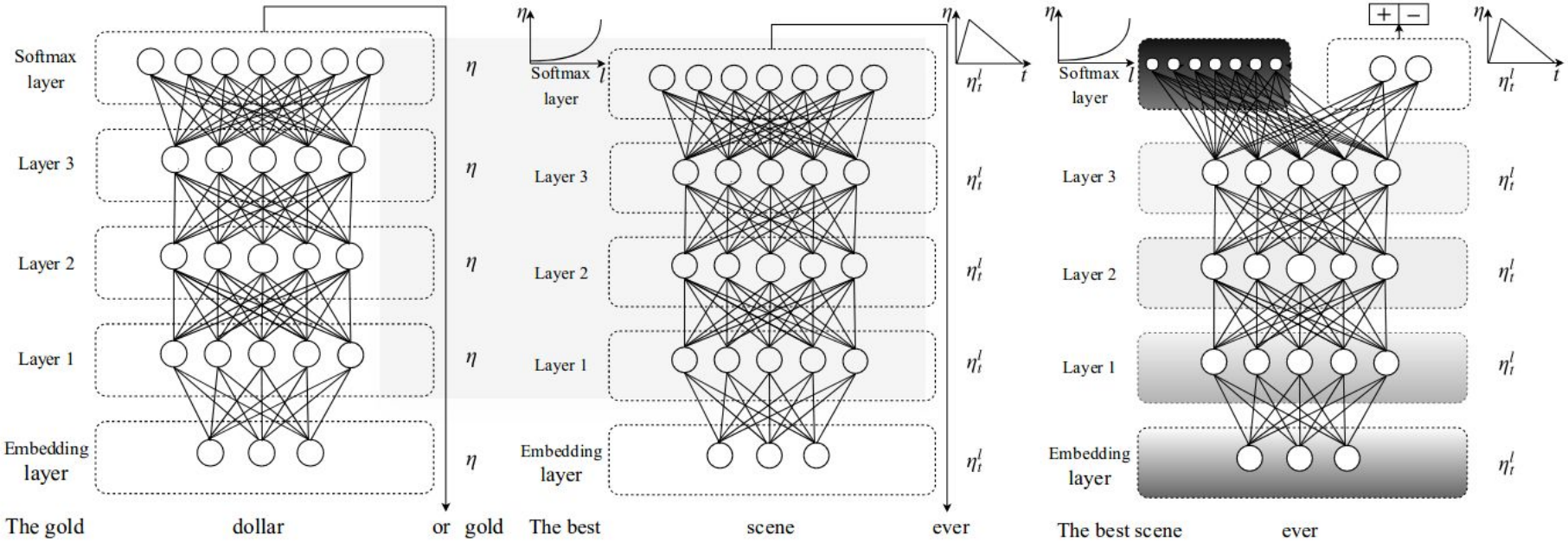
“Ta gra bardzo mi się podoba.”

Universal Language Model Fine-tuning



(a) LM pre-training

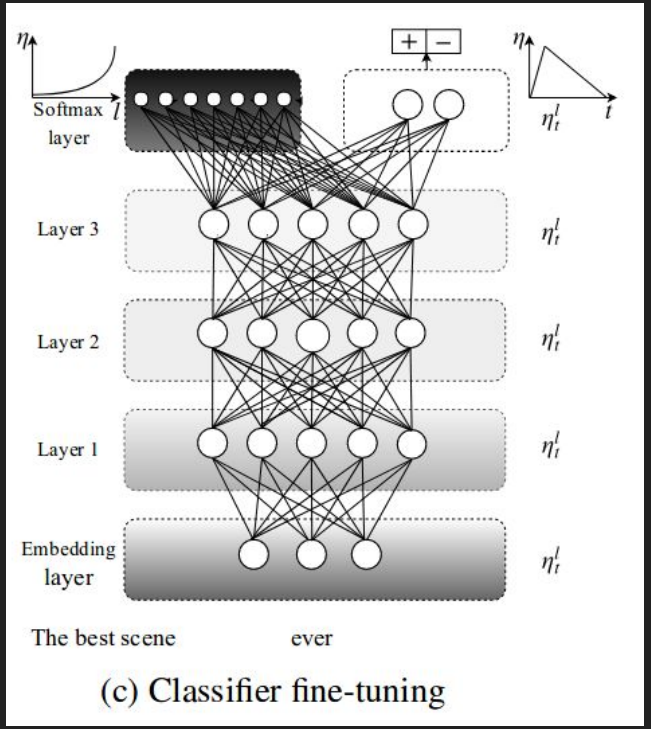

Softmax layer
Layer 3
Layer 2
Layer 1
Embedding layer
η
The gold
dollar
or gold
(a) LM pre-training
Softmax layer
Layer 3
Layer 2
Layer 1
Embedding layer
The best
scene
ever
(b) LM fine-tuning



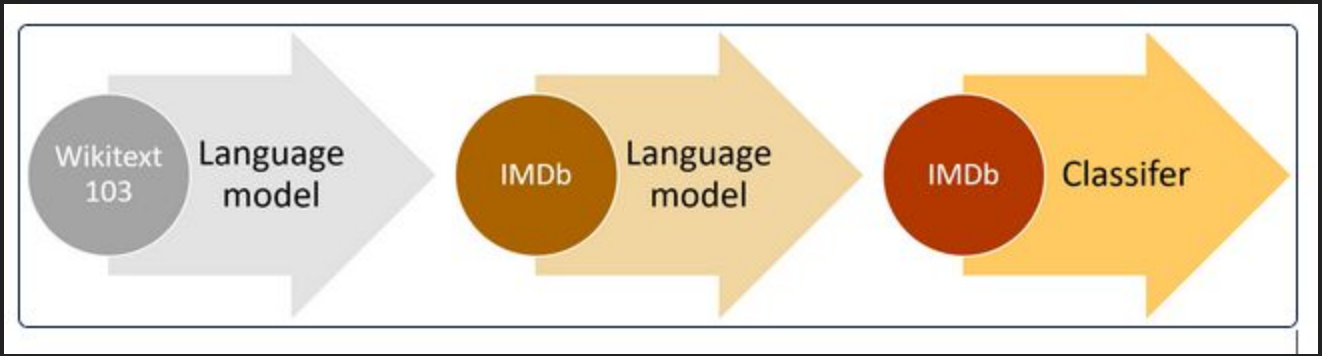
(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

“Całymi dniami gra na komputerze.”

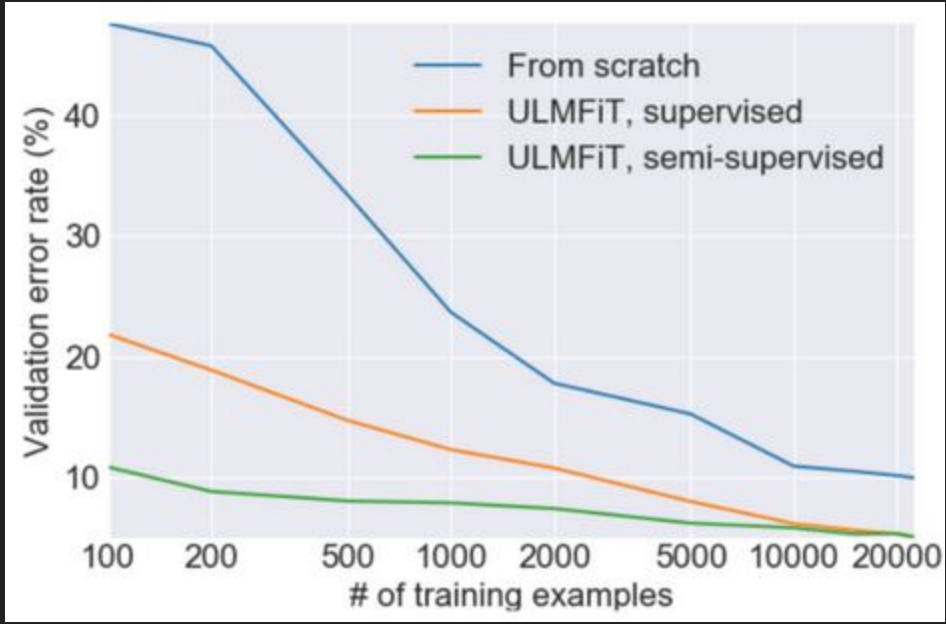
“Ta gra bardzo mi się podoba.”



(c) Classifier fine-tuning


Wikitext
103
Language
model
IMDb
Language
model
IMDb
Classifer

Klasyfikacja recenzji filmów z IMDb:



Używając Tesla V100 (3$ / 1h):

2 dni na wytrenowanie podstawowego language modelu

+

3 - 6h na finetuning

github.com/n-waves/poleval

2018

Protip:
Szybciej = Taniej

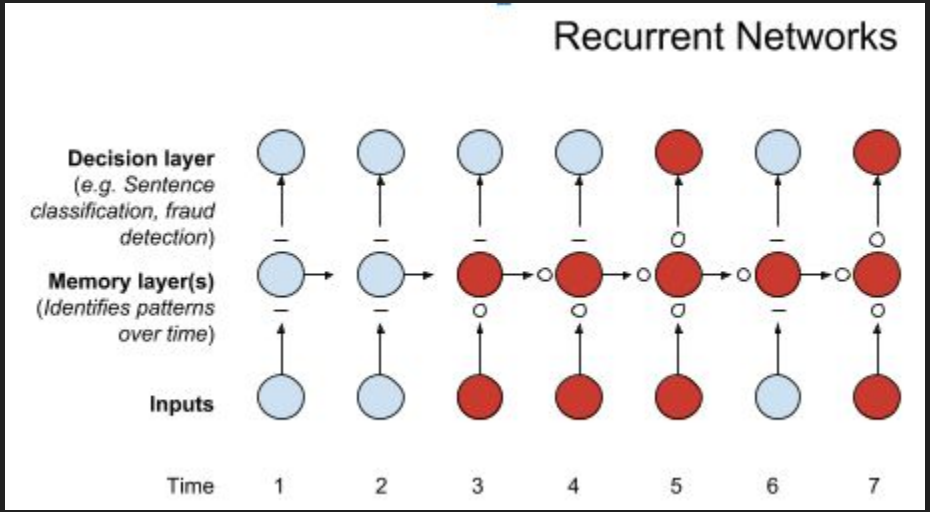
V100 jest 5 x szybsza niż K80
a tylko 3 razy droższa

BERT

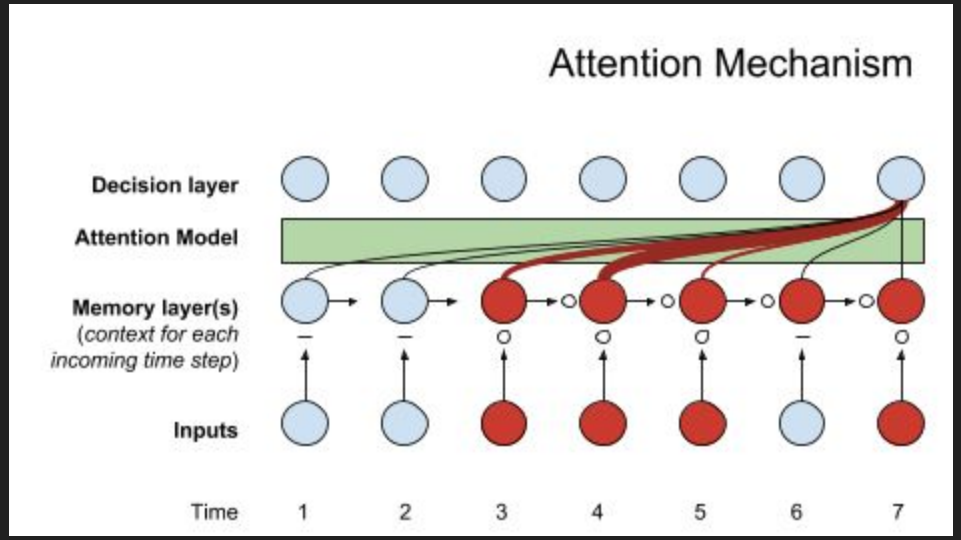
LSTM -> Transformer

Sukces w Tłumaczeniu Maszynowym

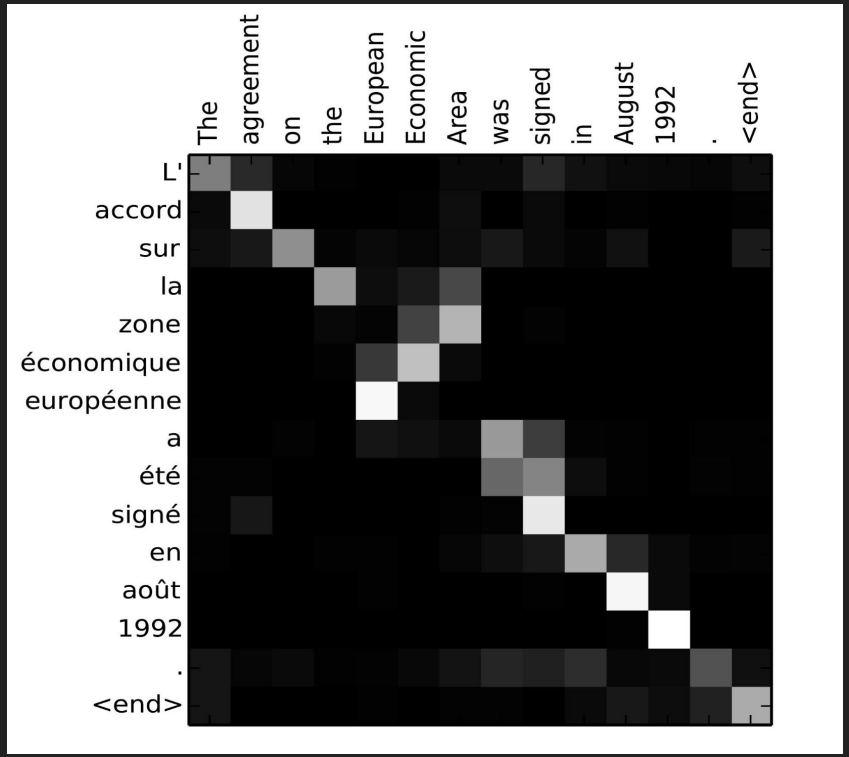
LSTM


Recurrent Networks
Decision layer
(e.g. Sentence classification, fraud detection)
Memory layer(s)
(Identifies patterns over time)
Inputs
Time
1
2
3
4
5
6
7

Attention Models



[European Economic Area] -> [zone e´conomique europe´en]



Tradycyjne Language Modele:

Tradycyjne Language Modele:

"the man went to a store"

Tradycyjne Language Modele:

"the man went to a store"

P(the | <s>)*P(man|<s> the)*P(went|<s> the man)*...

Tradycyjne Language Modele:

"the man went to a store"

P(the | <s>)*P(man|<s> the)*P(went|<s> the man)*...
+
P(store|</s>)*P(a|store </s>)*...

BERT:

Input:
the man [MASK1] to [MASK2] store
Label:
[MASK1] = went; [MASK2] = store

Input:
the man went to the store [SEP] he bought a gallon of milk

Label:
IsNext

Input:
the man went to the store [SEP] penguins are flightless birds

Label:
NotNext

16 x TPU V2

Koszt dzienny:

16 * 24h * 4.5$ = 1728 $

x 4 dni

6912 $

4 x GPU RTX 2080 Ti

24000zł

34 dni

Open source multilingual BERT model

Na razie mały...

github.com/google-research/bert/blob/master/multilingual.md

Tuning: ~pół dnia na Tesla V100
Koszt: ~150zł / 36$

OpenAI GPT-2

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

32 x TPU V3

Koszt dzienny:

32 * 24h * 8$ = 6144 $

x ???

Model nie został upubliczniony

Nie ma modelu w dla języka polskiego

XLNet

On nie był [Mask] [Mask] więc nie mógł wejść do [Mask]

On nie był dobrym człowiekiem więc nie mógł wejść do nieba.

On nie był jeszcze pełnoletni więc nie mógł wejść do klubu.

128 x TPU V3

Koszt dzienny:

128 * 24h * 8$ = 24576 $

x 2.5 dnia

61440 $

Dziękuję