Sieci rekurencyjne.

Analiza tekstu.

Plan

- Sieci rekurencyjne
- Klasyczny model językowy
- Modele językowe za pomocą RNN
- Klasyfikacja tekstu i atencja

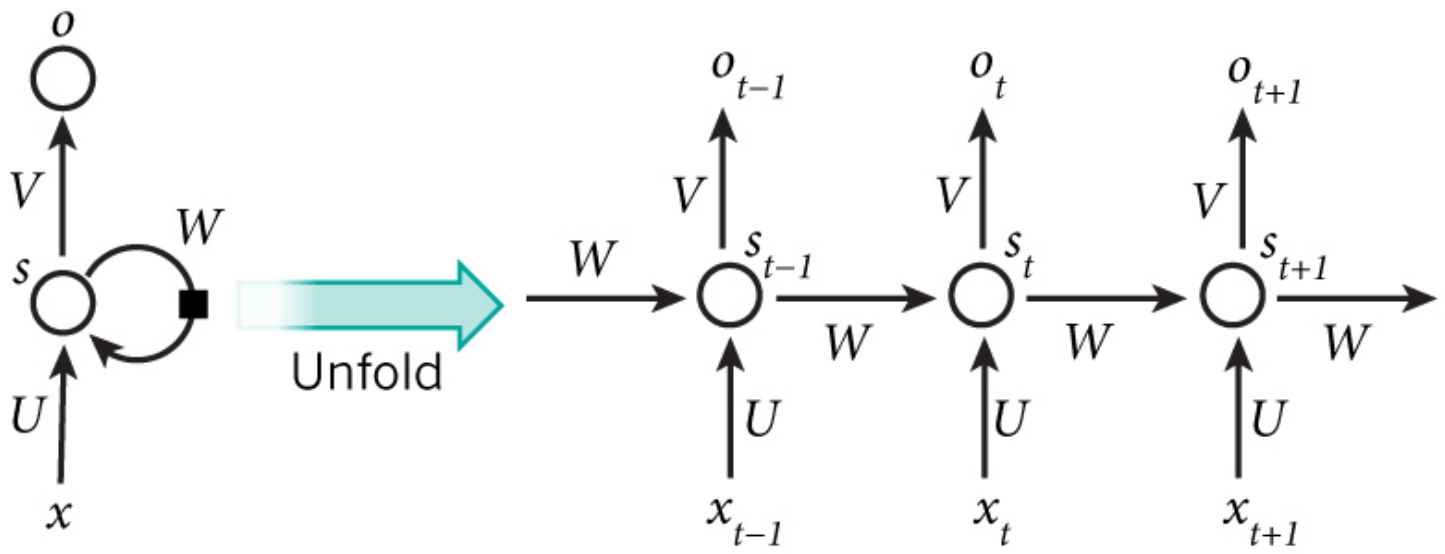
Rekurencyjne sieci neuronowe. Wprowadzenie.

Szczególny typ sieci neuronowych, przeznaczony do analizy sekwencji różnych wartości $x_1, \ldots, x_t$.

W szczególności te sekwencje to:

- Liczby
  - Predykcja szeregów czasowych
- Wyrazy i litery
  - Klasyfikacja tekstu
  - Generowanie tekstu
  - Tłumaczenia języka naturalnego
  - Znakowanie części mowy

RNN. Budowa i rozwijanie („unroll”) sieci.



x_t – zmienna wejściowa w czasie t, np. w postaci wektora one-hot

s_t- ukryty stan (hidden state) w czasie t, który liczony jest : $s_t = f(Ux_t + Ws_{t-1})$.

o_t – wynik modelu w czasie t, rozumiany jako rozkład prawdopodobieństwa na zbiorze

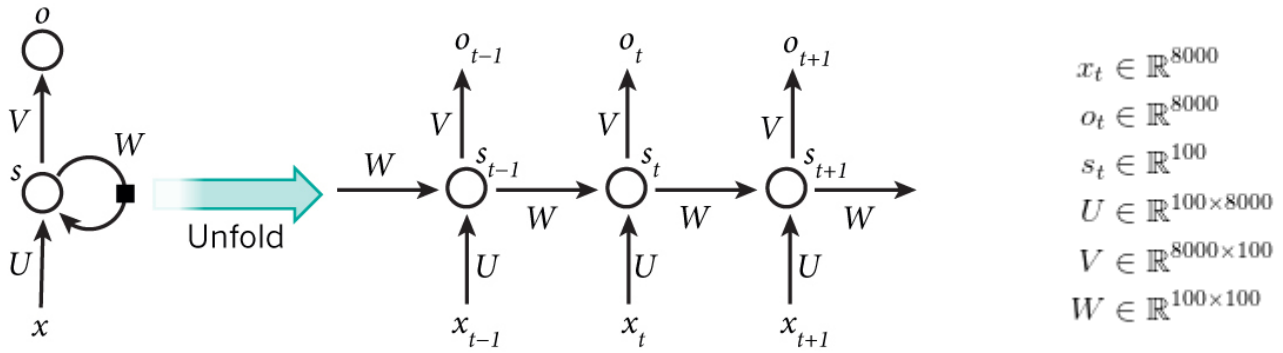
Czemu sieci rekurencyjne są skuteczne w zadaniach sekwencyjnych?

Dzielenie parametrów!

- Możliwe jest analizowanie kontekstu zdania

Przykład:
Pojechałem do **Francji**. Byłem w Paryżu i Nicei. Mówiłem po **francusku.**

- Pozwala na stosowanie modelu do sekwencji o innej długości niż widziana podczas treningu



$x_t \in \mathbb{R}^{8000}$
$o_t \in \mathbb{R}^{8000}$
$s_t \in \mathbb{R}^{100}$
$U \in \mathbb{R}^{100 \times 8000}$
$V \in \mathbb{R}^{8000 \times 100}$
$W \in \mathbb{R}^{100 \times 100}$

Modele językowe (1)

Cel: Obliczenie rozkładu prawdopodobieństwa danej sekwencji wyrazów

Obliczenie prawdopodobieństwa wystąpienia kolejnego słowa $P(w_3|w_1, w_2)$

$$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

P(Ala ma kota kot ma Alę) = P(Ala)*P(ma|Ala)*P(kota|Ala ma)*P(kot|Ala ma kota)
*P(ma|Ala ma kota kot)*P(Alę|Ala ma kota kot ma)

$$P(\text{ma}|\text{Ala ma kota kot}) = \frac{\text{Count(Ala ma kota kot ma)}}{\text{Count(Ala ma kota kot)}}$$

Standardowe metody używają bi, tri-gramów

$$P(\text{ma}|\text{Ala ma kota kot}) = \text{=P(ma|kot)} = \frac{\text{Count(kot ma)}}{\text{Count(kot)}}$$

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

Modele językowe (2)

- N-gramy nie biorą pod uwagę podobieństwa kontekstów
- Liczba analizowanych słów (N) w praktyce jest ograniczona

Perplexity

2^J

$$J = -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{|V|} y_{t,j} \log \hat{y}_{t,j}$$

Recurrent neural network based language model

Tomáš Mikolov[1,2], *Martin Karafiát*[1], *Lukáš Burget*[1], *Jan "Honza" Černocký*[1], *Sanjeev Khudanpur*[2]

[1]Speech@FIT, Brno University of Technology, Czech Republic
[2] Department of Electrical and Computer Engineering, Johns Hopkins University, USA
{imikolov,karafiat,burget,cernocky}@fit.vutbr.cz, khudanpur@jhu.edu

Motywacja:

- Bengio et al. (2003) zbudowali model językowy oparty o prostą sieć neuronową
- Oznacza to, że badany kontekst musi mieć stałą długość (w praktyce od 5 do 10 wyrazów)

Specyfikacja modelu

$$x(t) = w(t) + s(t-1) \quad (1)$$

$$s_j(t) = f\left(\sum_i x_i(t) u_{ji}\right) \quad (2)$$

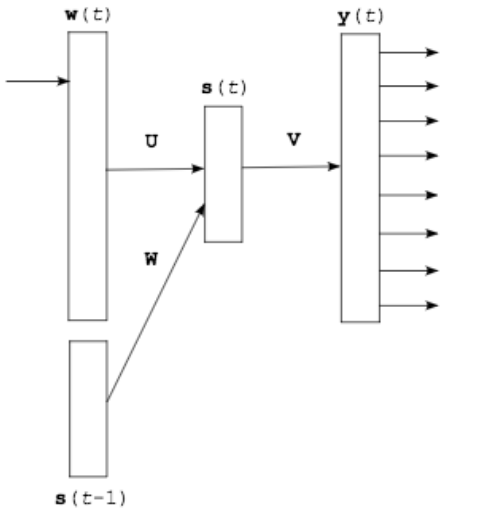
$$y_k(t) = g\left(\sum_j s_j(t) v_{kj}\right) \quad (3)$$

where $f(z)$ is sigmoid activation function:

$$f(z) = \frac{1}{1+e^{-z}} \quad (4)$$

and $g(z)$ is softmax function:

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \quad (5)$$



U - macierz wag pomiędzy wejściem a warstwą ukrytą

V – macierz wag pomiędzy warstwą ukrytą a wyjściem

W – macierz wag przechodząca przez warstwy ukryte

Trenowanie modelu i wyniki

- Trenowanie odbywa się za pomocą Stochastycznego Spadku Gradientu (SGD)
- Do aktualizacji wag U, V, W używany jest algorytm propagacji wstecznej (backpropagation)

Model	# words	PPL	WER
KN5 LM	200K	336	16.4
KN5 LM + RNN 90/2	200K	271	15.4
KN5 LM	1M	287	15.1
KN5 LM + RNN 90/2	1M	225	14.0
KN5 LM	6.4M	221	13.5
KN5 LM + RNN 250/5	6.4M	156	11.7

Table 2: *Comparison of various configurations of RNN LMs and combinations with backoff models while using 6.4M words in training data (WSJ DEV).*

	PPL		WER	
Model	RNN	RNN+KN	RNN	RNN+KN
KN5 - baseline	-	221	-	13.5
RNN 60/20	229	186	13.2	12.6
RNN 90/10	202	173	12.8	12.2
RNN 250/5	173	155	12.3	11.7
RNN 250/2	176	156	12.0	11.9
RNN 400/10	171	152	12.5	12.1
3xRNN static	151	143	11.6	11.3
3xRNN dynamic	128	121	11.3	11.1

Wygenerowany tekst

THANKS FOR COMING IN NEXT IN A COUPLE OF MINUTES
WHEN WE TAKE A LOOK AT OUR ACCOMPANYING STORY IMAGE
GUIDE WHY ARE ANY OF THOSE DETAILS BEING HEARD IN LONDON
BUT DEFENSE ATTORNEYS SAY THEY THOUGHT THE CONTACT WAS
NOT AIMED DAMAGING AT ANY SUSPECTS
THE UNITED NATIONS SECURITY COUNCIL IS NAMED TO WITHIN
TWO MOST OF IRAQI ELECTION OFFICIALS
IT IS THE MINIMUM TIME A TOTAL OF ONE DETERMINED TO
APPLY LIMITS TO THE FOREIGN MINISTERS WHO HAD MORE POWER
AND NOW THAN ANY MAN WOULD NAME A CABINET ORAL
FIND OUT HOW IMPORTANT HIS DIFFERENT RECOMMENDATION IS
TO MAKE WHAT THIS WHITE HOUSE WILL WILL TO BE ADDRESSED
ELAINE MATHEWS IS A POLITICAL CORRESPONDENT FOR THE

OR STUDENT'S IS FROM TEETH PROSECUTORS DO FILLED WITH
HER SOME BACKGROUND ON WHAT WAS GOING ON HERE
ALUMINUM CANS OF PEACE
PIPER SWEAT COLONEL SAYING HAVE ALREADY MADE LAW THAT
WOULD PREVENT THE BACTERIA
DOWN FOR THE MOST OF IT IN NINETEEN SEVENTY EIGHT WHICH
WAS ONE OF A NUMBER OF ISSUES INCLUDING CIVIL SUIT BY
THIS TIME NEXT YEAR
CRYSTAL
FIRMLY AS A HERO OF MINE A PREVIEW THAT
THOMAS SEVENTY BODIES AND ASKING QUESTIONS MAYBE
ATTORNEY'S OFFICE THEATERS CUT ACROSS THE ELEVENTH AND
SUPPORT THEM WITH ELLEN WISEST PULLING DATA GATHERING IN
RESPONSE TO AN UNMITIGATED DISPOSITION CONTRACTORS AND

Hierarchical Attention Networks for Document Classification

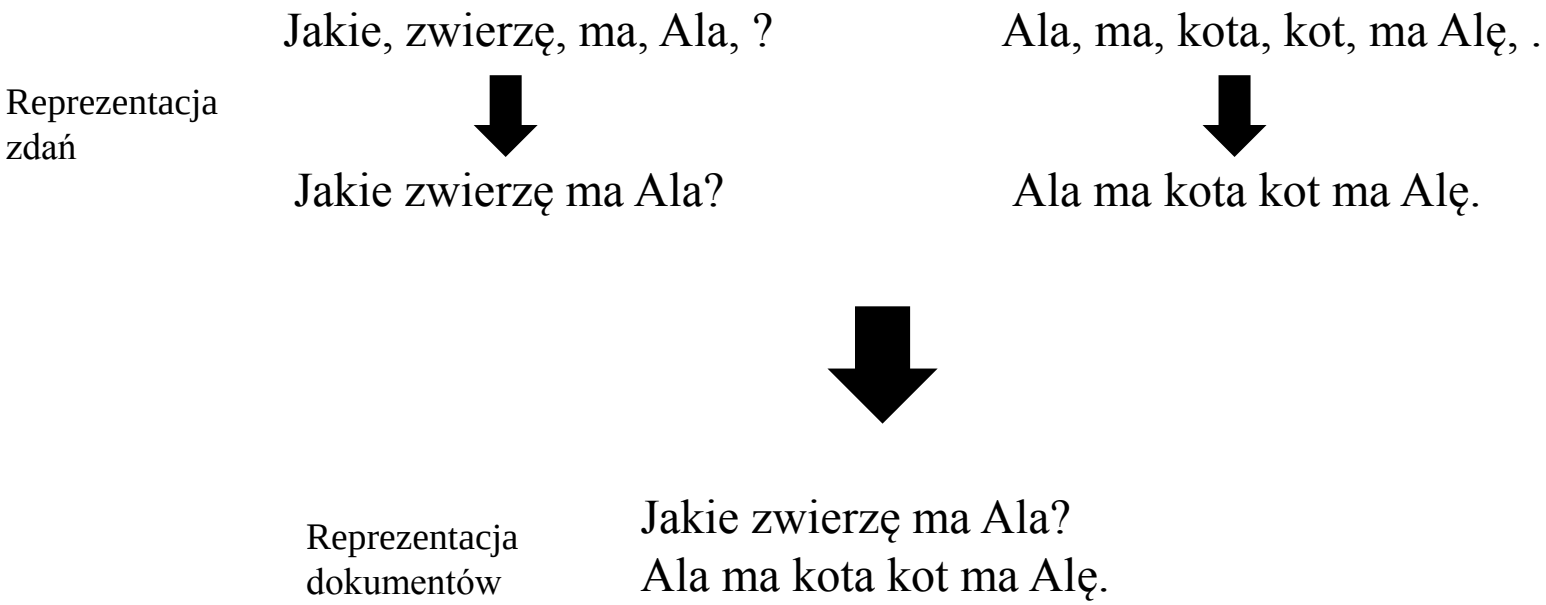
Zichao Yang[1], **Diyi Yang**[1], **Chris Dyer**[1], **Xiaodong He**[2], **Alex Smola**[1], **Eduard Hovy**[1]
[1]Carnegie Mellon University, [2]Microsoft Research, Redmond
{zichaoy, diyiy, cdyer, hovy}@cs.cmu.edu
xiaohe@microsoft.com alex@smola.org

Motywacja:

- Modele klasyfikujące tekst opierają się na zmiennych zbudowanych za pomocą metody TF-IDF
- Dotychczasowe modele klasyfikacji tekstu za pomocą sieci neuronowych osiągały zadowalające a nawet bardzo dobre wyniki
- Hierarchiczna Sieć Atencji (Hierarchical Attention Network) dzięki włączeniu struktury dokumentu do modelu pozwala zbudować lepszą reprezentację
- Nie wszystkie fragmenty dokumentu są tak samo istotne
- Atencja!

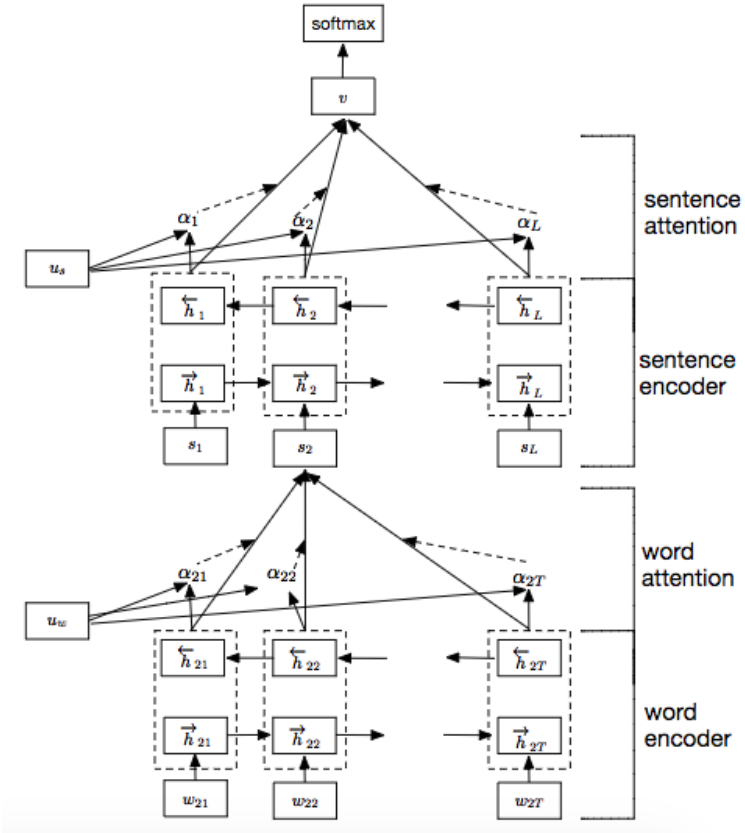
Logika modelu (1)

Hierarchiczna struktura modelu wyrażona za pomocą rozbicia na poszczególne komponenty.



Logika modelu (2)

Słowa w dokumencie mają różny poziom informatywności, który często zależy od kontekstu.



$$
\begin{aligned}
x_{it} &= W_e w_{it}, t \in [1, T], \\
\overrightarrow{h}_{it} &= \overrightarrow{\mathrm{GRU}}(x_{it}), t \in [1, T], \\
\overleftarrow{h}_{it} &= \overleftarrow{\mathrm{GRU}}(x_{it}), t \in [T, 1].
\end{aligned}
$$

$$
\begin{aligned}
u_{it} &= \tanh(W_w h_{it} + b_w) \\
\alpha_{it} &= \frac{\exp(u_{it}^\top u_w)}{\sum_t \exp(u_{it}^\top u_w)} \\
s_i &= \sum_t \alpha_{it} h_{it}.
\end{aligned}
$$

Wyniki modelu

	Methods	Yelp'13	Yelp'14	Yelp'15	IMDB	Yahoo Answer	Amazon
Zhang et al., 2015	BoW	-	-	58.0	-	68.9	54.4
	BoW TFIDF	-	-	59.9	-	71.0	55.3
	ngrams	-	-	56.3	-	68.5	54.3
	ngrams TFIDF	-	-	54.8	-	68.5	52.4
	Bag-of-means	-	-	52.5	-	60.5	44.1
Tang et al., 2015	Majority	35.6	36.1	36.9	17.9	-	-
	SVM + Unigrams	58.9	60.0	61.1	39.9	-	-
	SVM + Bigrams	57.6	61.6	62.4	40.9	-	-
	SVM + TextFeatures	59.8	61.8	62.4	40.5	-	-
	SVM + AverageSG	54.3	55.7	56.8	31.9	-	-
	SVM + SSWE	53.5	54.3	55.4	26.2	-	-
Zhang et al., 2015	LSTM	-	-	58.2	-	70.8	59.4
	CNN-char	-	-	62.0	-	71.2	59.6
	CNN-word	-	-	60.5	-	71.2	57.6
Tang et al., 2015	Paragraph Vector	57.7	59.2	60.5	34.1	-	-
	CNN-word	59.7	61.0	61.5	37.6	-	-
	Conv-GRNN	63.7	65.5	66.0	42.5	-	-
	LSTM-GRNN	65.1	67.1	67.6	45.3	-	-
This paper	HN-AVE	67.0	69.3	69.9	47.8	75.2	62.9
	HN-MAX	66.9	69.3	70.1	48.2	75.2	62.9
	HN-ATT	**68.2**	**70.5**	**71.0**	**49.4**	**75.8**	**63.6**

Table 2: Document Classification, in percentage

Atencja

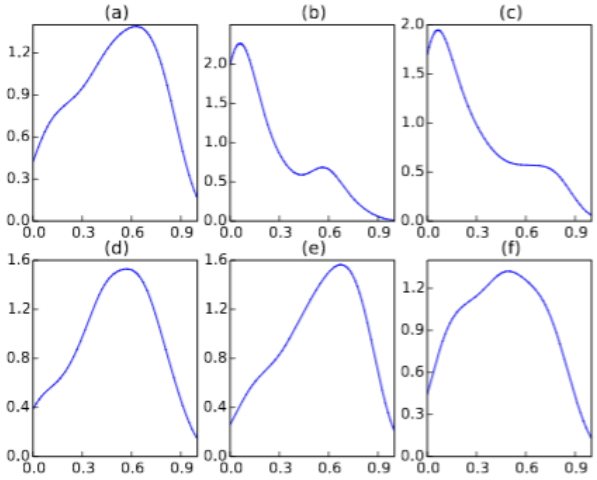
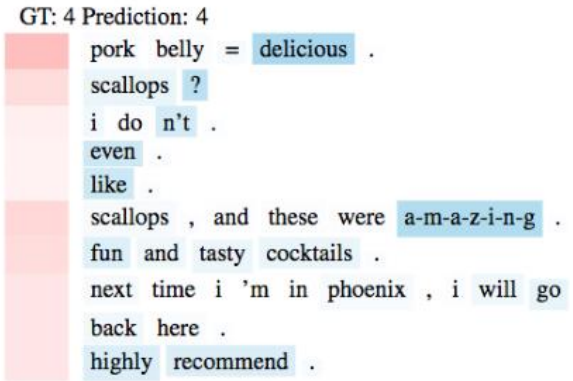
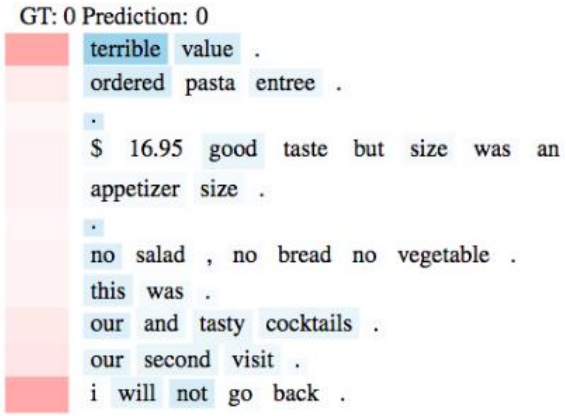


Figure 3: Attention weight distribution of good. (a) — aggregate distribution on the test split; (b)-(f) stratified for reviews with ratings 1-5 respectively. We can see that the weight distribution shifts to *higher* end as the rating goes higher.





Atencja ma zastosowanie również w przetwarzaniu obrazu i generowaniu opisów.

„Show, Attend and Tell: Neural Image Caption Generation with Visual Attention” (Xu et al., 2016)

- Hierarchical Attention Networks for Document Classification Yang et al.
http://www.cs.cmu.edu/~./hovy/papers/16HLT-hierarchical-attention-networks.pdf

- Recurrent neural network based language model, Mikolov et al.
http://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf

- Speech and language processing, Daniel Jurfasky i James Martin
Language modeling with N-grams
https://web.stanford.edu/~jurafsky/slp3/4.pdf