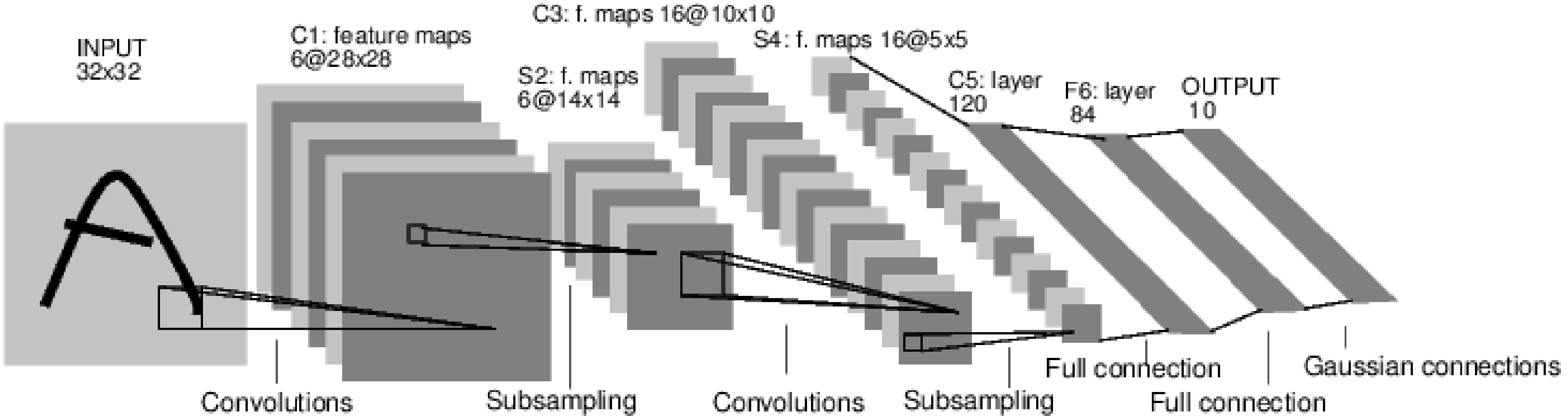
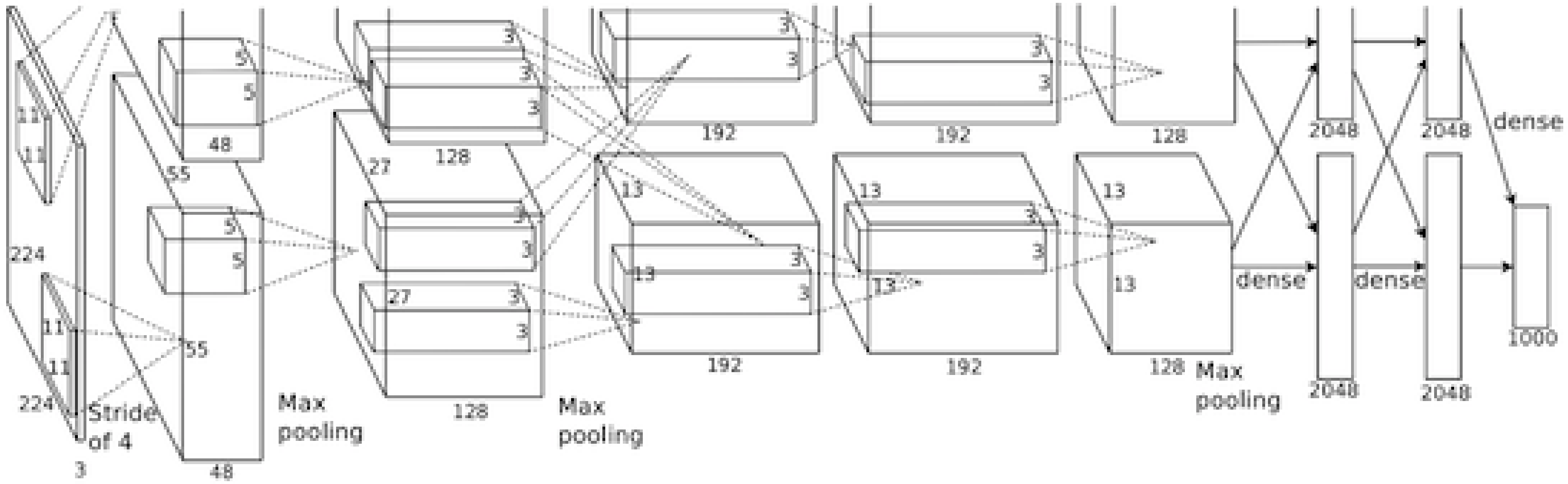
Deep learning

LeNet-5 (1998)



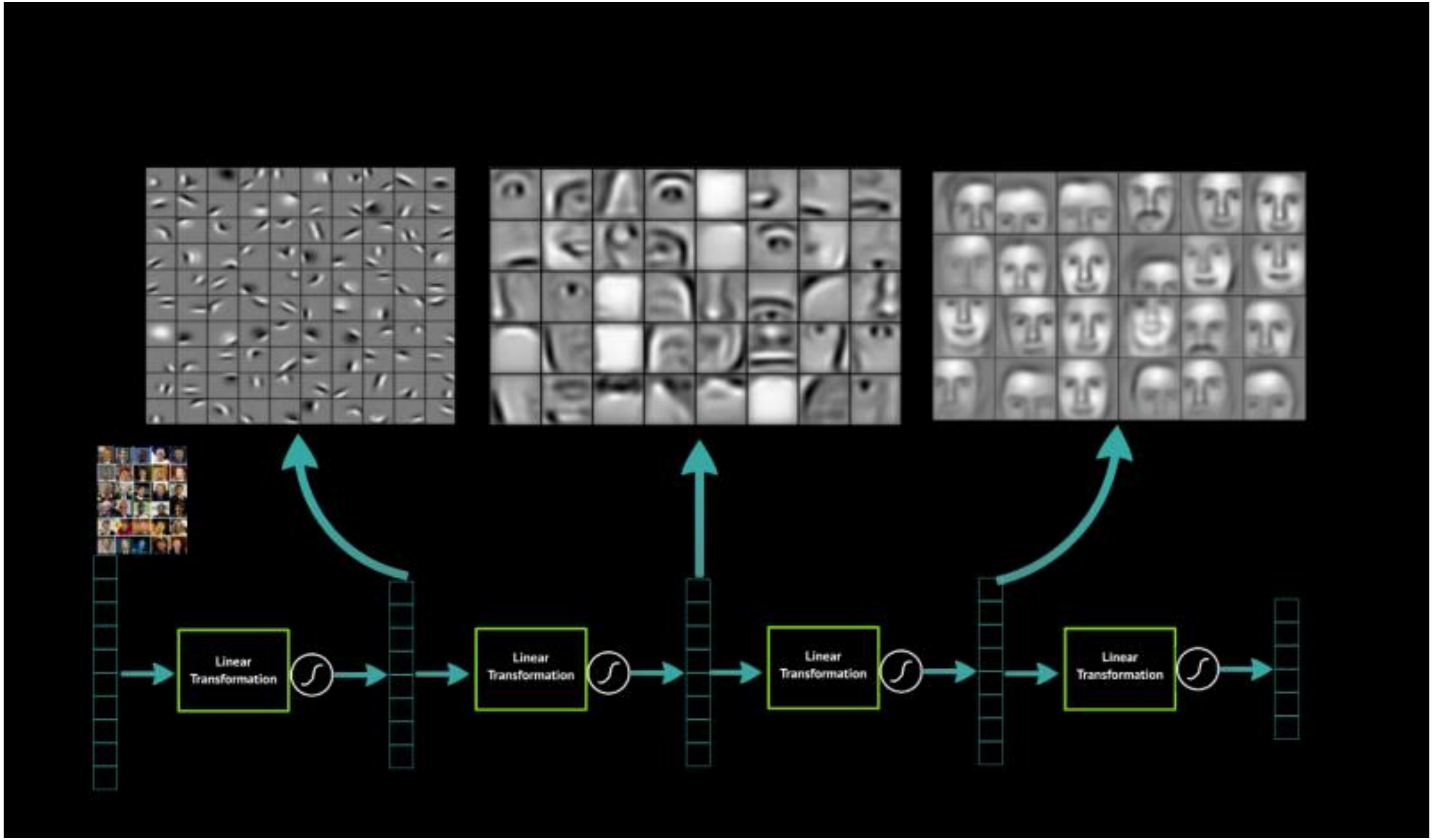
(5x5)x6 + (5x5x6)x16 + (16x5x5)x120 + 120x84 + 84x10 = 61470

AlexNet (2012)



(11x11x3)x48 + (5x5x48)x128 + (3x3x128)x192x2 + (3x3x192)x192 + (3x3x192)x128 + (13x13x128)x2048x2 + 2048x2048x2 + 2048x1000 = 100 207 632

Convolution layers



Deep Learning for NLP
(Natural Language Processing)

word2vec (Google 2013)

The Idea : Let’s maximize cooccurrence probability for cooccurring words (don’t care about order – bag-of-words).

Quick brown fox jumps over the lazy dog

w_t w_t w_c w_t w_t

word2vec (Google 2013)

The Idea : Let’s maximize cooccurrence probability for cooccurring words (don’t care about order – bag-of-words).

Quick brown fox jumps over the lazy dog

w_t w_t w_c w_t w_t

$$P(w_t|w_c) = \frac{\exp\big(f(w_t, w_c)\big)}{\sum_{w_i \in Dict} \exp(f(w_i, w_c))}$$

$$Loss\ function\ J = \frac{1}{T}\sum J_t, where\ J_t = -logP(w_t|w_c) = -f(w_t, w_c) + \log\left(\sum_{w_i \in Dict} \exp\big(f(w_i, w_c)\big)\right)$$

word2vec (Google 2013)

The Idea : Let's maximize cooccurrence probability for cooccurring words (don't care about order – bag-of-words).

Quick	brown	fox	jumps	over	the lazy dog
w_t	w_t	w_c	w_t	w_t	

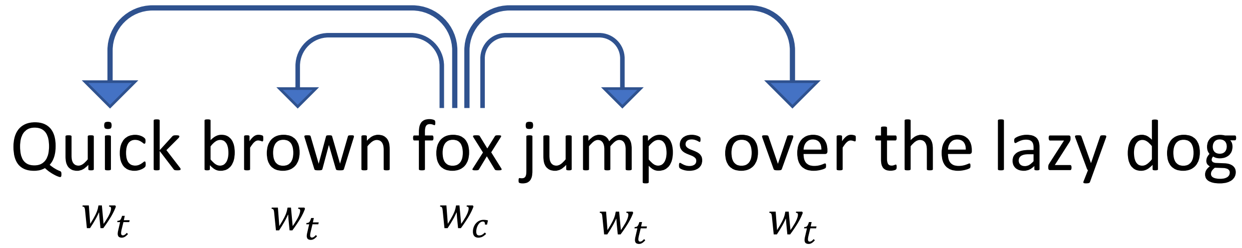
$$P(w_t|w_c) = \frac{\exp\big(f(w_t, w_c)\big)}{\sum_{w_i \in Dict} \exp(f(w_i, w_c))}$$

$$Loss\ function\ J = \frac{1}{T}\sum J_t, where\ J_t = -logP(w_t|w_c) = -f(w_t, w_c) + \log\left(\sum_{w_i \in Dict} \exp\big(f(w_i, w_c)\big)\right)$$

The simplest (but effective approach): $f(w_t, w_c) = u_{w_t}^T v_{w_c}$, where u and v are two-vector word embeddings.

word2vec (Google 2013)

The Idea : Let's maximize cooccurrence probability for cooccurring words (don't care about order – bag-of-words).



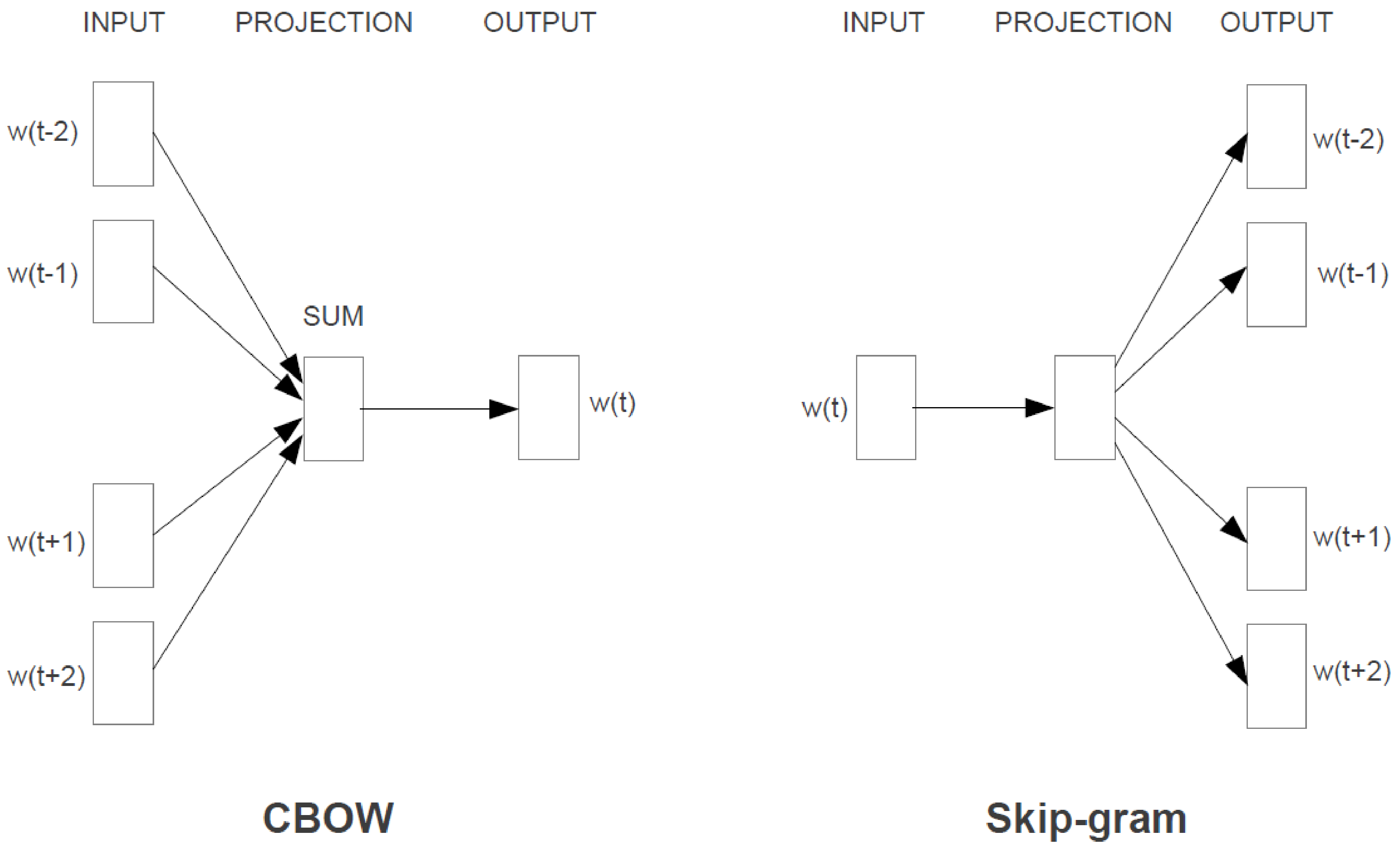
$$P(w_t|w_c) = \frac{\exp\left(f(w_t, w_c)\right)}{\sum_{w_i \in Dict} \exp(f(w_i, w_c))}$$

$$Loss\ function\ J = \frac{1}{T}\sum J_t, where\ J_t = -logP(w_t|w_c) = -f(w_t, w_c) + \log\left(\sum_{w_i \in Dict} \exp\left(f(w_i, w_c)\right)\right)$$

The simplest (but effective approach): $f(w_t, w_c) = u_{w_t}^T v_{w_c}$, where u and v are two-vector word embeddings.

We can use gradient descent.

CBOW & Skip-gram

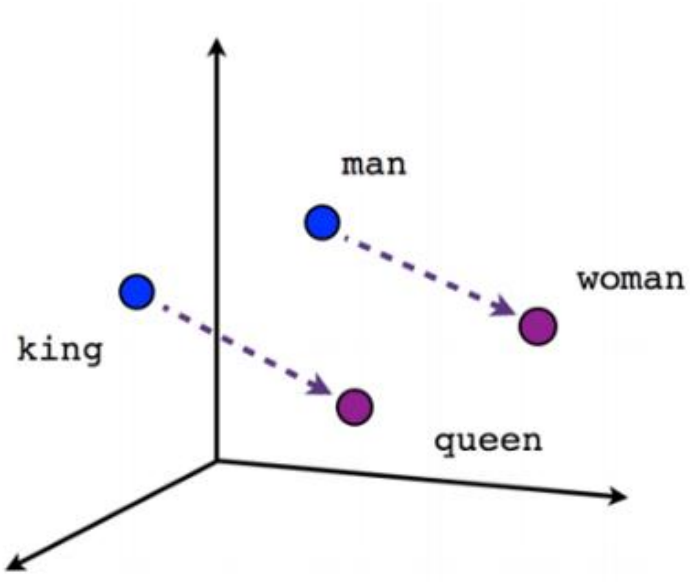


Fasttext

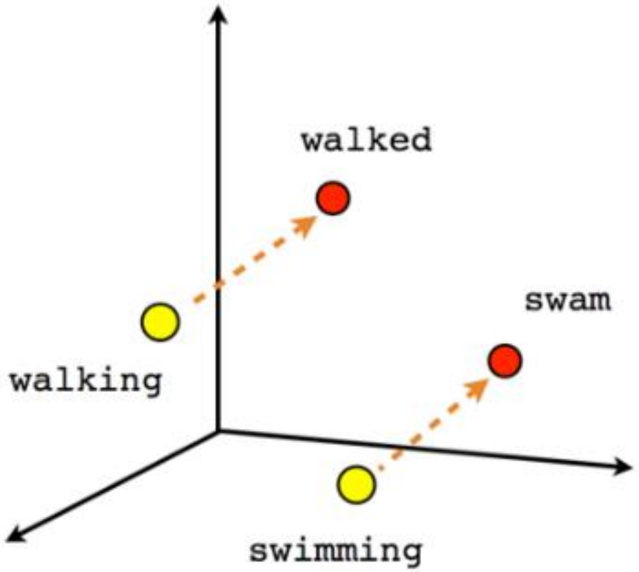
Adding n-grams to the dictionary:

where → <where> + <wh + whe + her + ere + re>

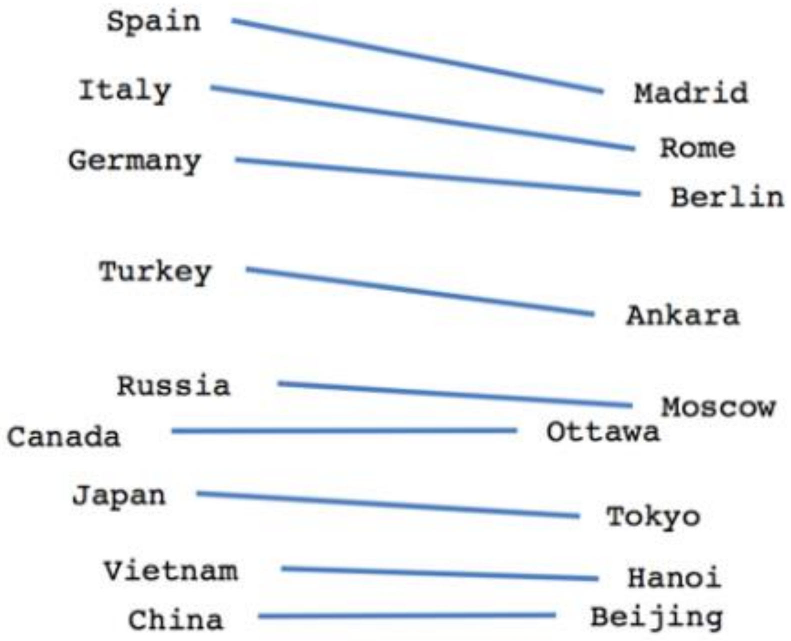
Word Embeddings



Male-Female



Verb tense



Country-Capital

Semantic and syntactic relations

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

skip-gram pair results

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Metrics comparison

Model	Dim.	Dataset	Semantic	Syntactic	Average
PCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	**67.5**	**54.3**	**60.3**
SG	300	1.6B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
GloVe	300	1.6B	**80.8**	**61.5**	**70.3**
SVD	300	6B	6.3	8.1	7.3
CBOW	300	6B	63.6	**67.4**	65.7
GloVe	300	6B	**77.4**	67.0	**71.7**
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD	300	42B	38.4	58.2	49.2
GloVe	300	42B	**81.9**	**69.3**	**75.0**

Model	ACE	MUC7
SVD	77.3	73.7
PCA	81.7	80.7
CBOW	82.2	81.1
GloVe	**82.9**	**82.2**

BioVectors

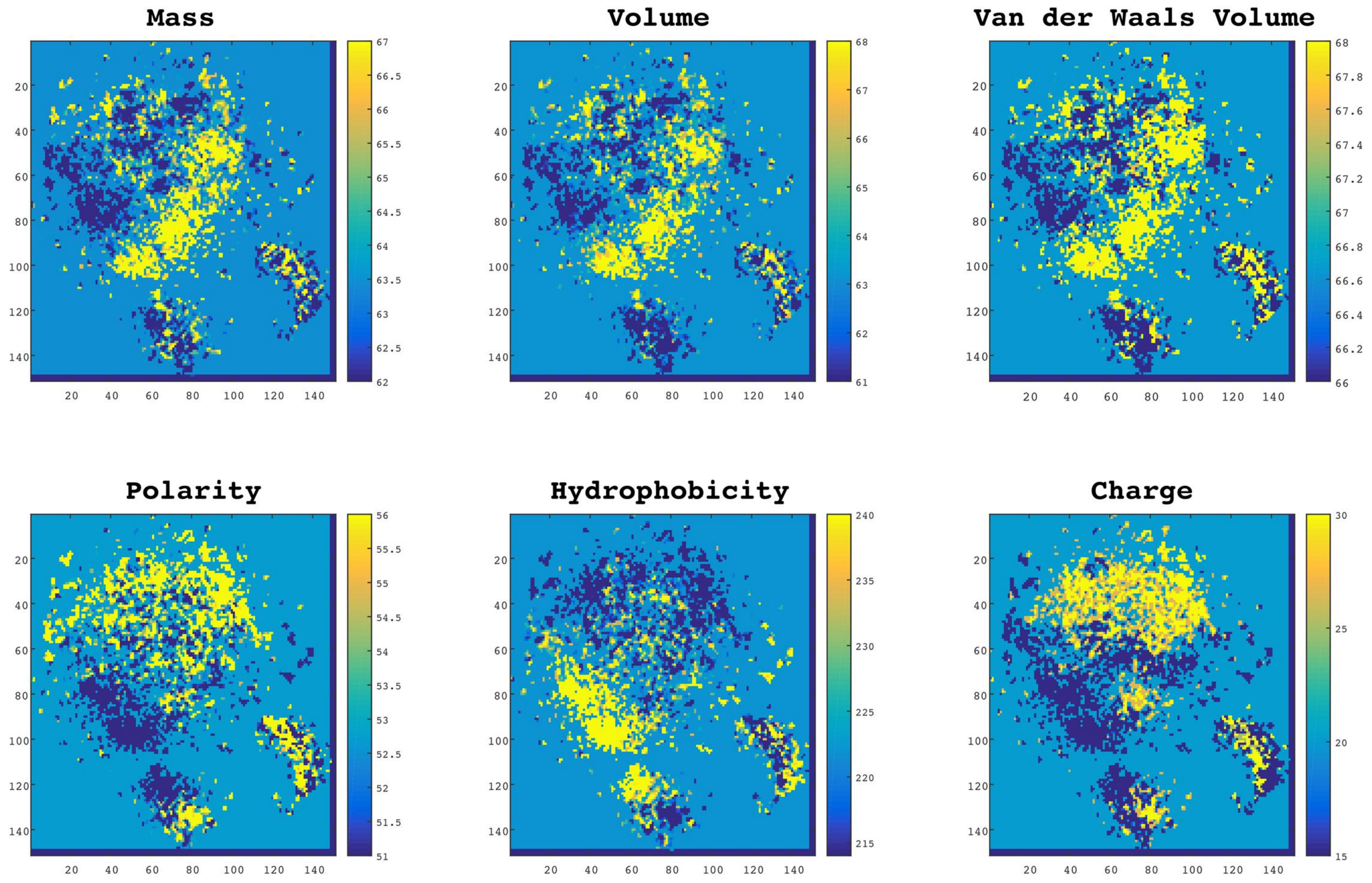
Original Sequence

$$^{(1)}\overrightarrow{M}\,^{(2)}\overrightarrow{A}\,^{(3)}\overrightarrow{F}SAEDVLKEYDRRRRRMEAL..$$

Splittings

$$\begin{cases} 1) & \text{MAF, SAE, DVL, KEY, DRR, RRM, ..} \\ 2) & \text{AFS, AED, VLK, EYD, RRR, RME, ..} \\ 3) & \text{FSA ,EDV, LKE, YDR, RRR, MEA, ..} \end{cases}$$

BioVecs

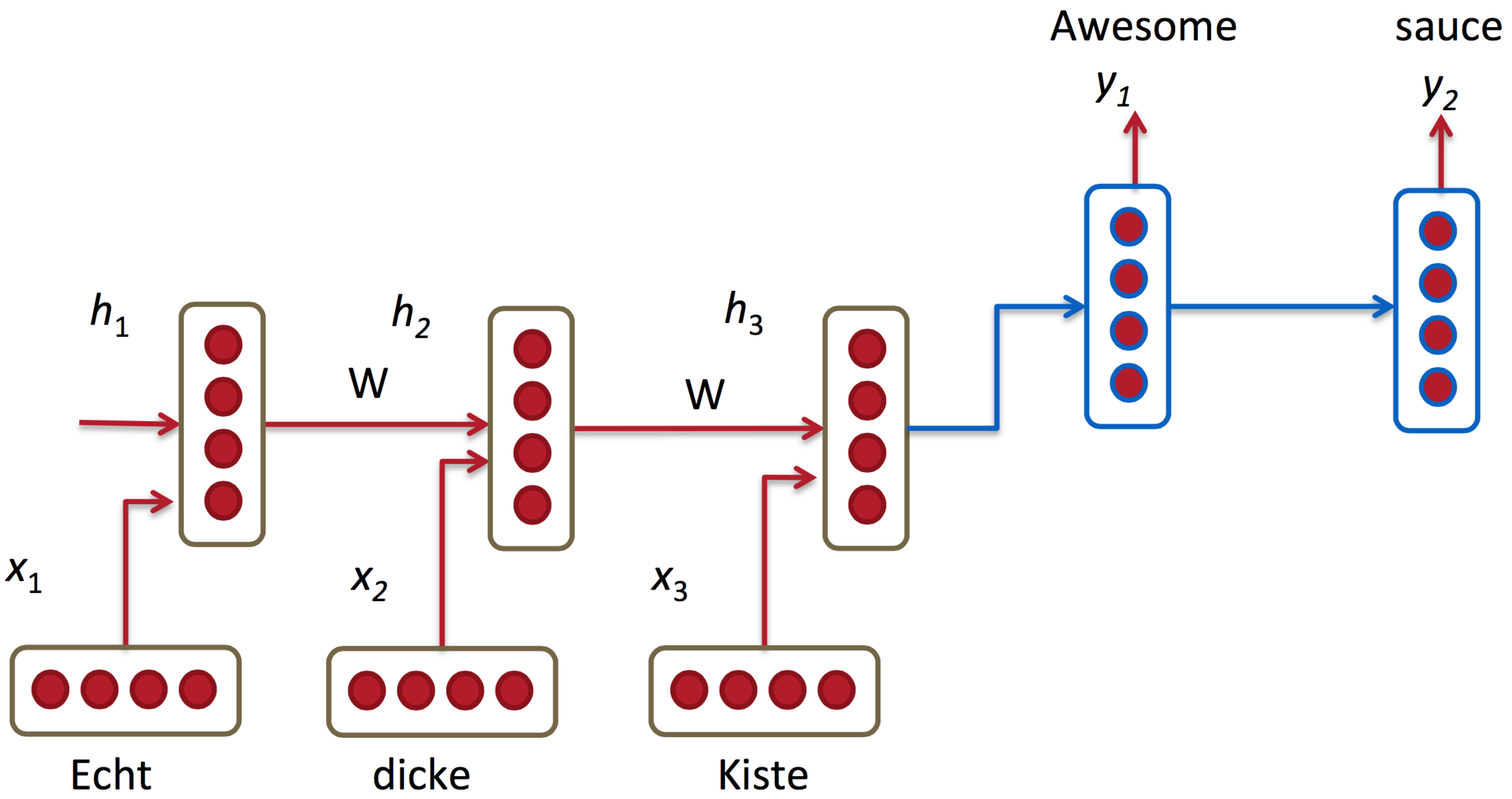


Mass
20 40 60 80 100 120 140
20 40 60 80 100 120 140
62 62.5 63 63.5 64 64.5 65 65.5 66 66.5 67
Volume
20 40 60 80 100 120 140
20 40 60 80 100 120 140
61 62 63 64 65 66 67 68
Van der Waals Volume
20 40 60 80 100 120 140
20 40 60 80 100 120 140
66 66.2 66.4 66.6 66.8 67 67.2 67.4 67.6 67.8 68
Polarity
20 40 60 80 100 120 140
20 40 60 80 100 120 140
51 51.5 52 52.5 53 53.5 54 54.5 55 55.5 56
Hydrophobicity
20 40 60 80 100 120 140
20 40 60 80 100 120 140
215 220 225 230 235 240
Charge
20 40 60 80 100 120 140
20 40 60 80 100 120 140
15 20 25 30

Protein family classification

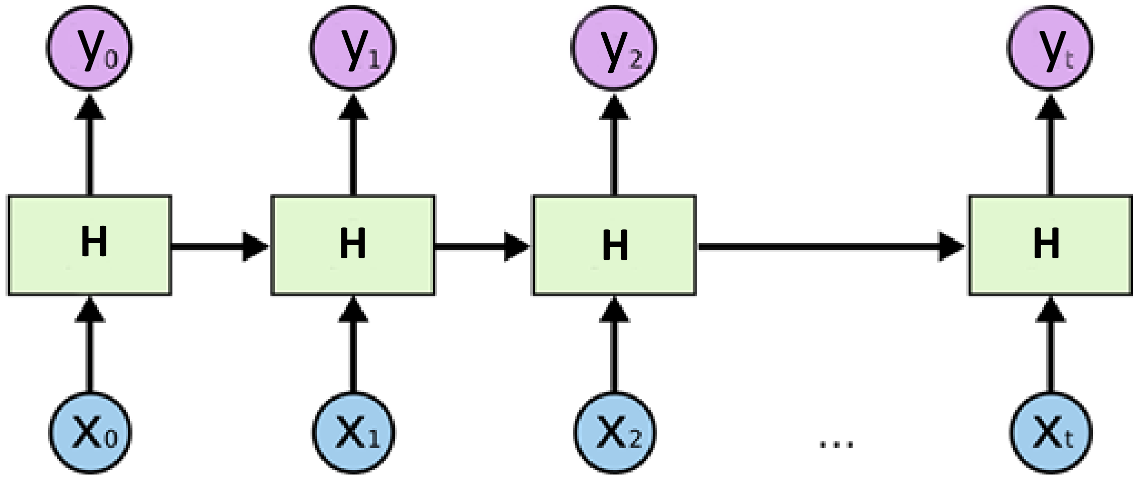
Family name	Training instances		Classification Result		
	# of positive sequences	# of negative sequences	Specificity	Sensitivity	Accuracy
50S ribosome-binding GTPase	3,084	3,084	0.95	0.93	0.94
Helicase conserved C-terminal domain	2,518	2,518	0.83	0.80	0.82
ATP synthase alpha-beta family, nucleotide-binding domain	2,387	2,387	0.98	0.97	0.97
7 transmembrane receptor (rhodopsin family)	1,820	1,820	0.95	0.96	0.95
Amino acid kinase family	1,750	1,750	0.91	0.92	0.91
ATPase family associated with various cellular activities (AAA)	1711	1711	0.92	0.90	0.91
tRNA synthetases class I (I, L, M and V)	1,634	1,634	0.97	0.97	0.97
tRNA synthetases class II (D, K and N)	1,419	1,419	0.88	0.83	0.85
Major Facilitator Superfamily	1,303	1,303	0.95	0.97	0.96
Hsp70 protein	1,272	1,272	0.97	0.97	0.97
NADH-Ubiquinone-plastoquinone (complex I), various chains	1,251	1,251	0.97	0.97	0.97
Histidine biosynthesis protein	1,248	1,248	0.96	0.97	0.97
TCP-1-cpn60 chaperonin family	1,246	1,246	0.95	0.96	0.95

Deep Learning for
Natural Language Processing

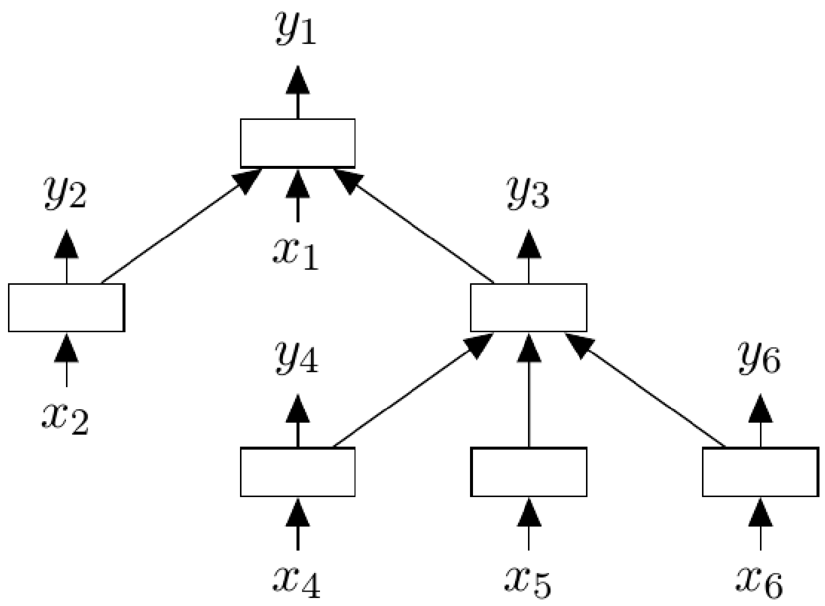

Awesome
sauce
y_1
y_2
h_1
h_2
h_3
W
W
x_1
x_2
x_3
Echt
dicke
Kiste

RNN

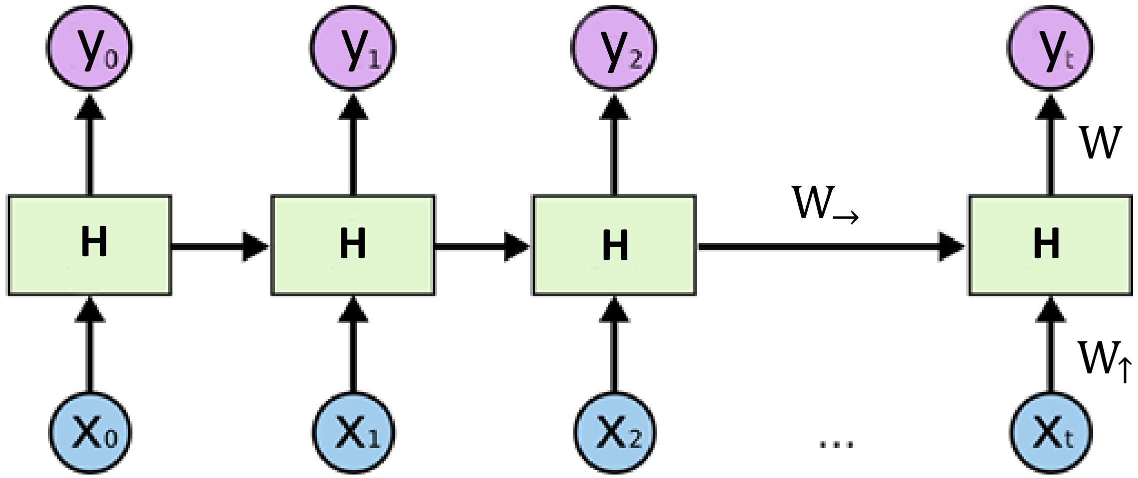
Recurrent



Recursive



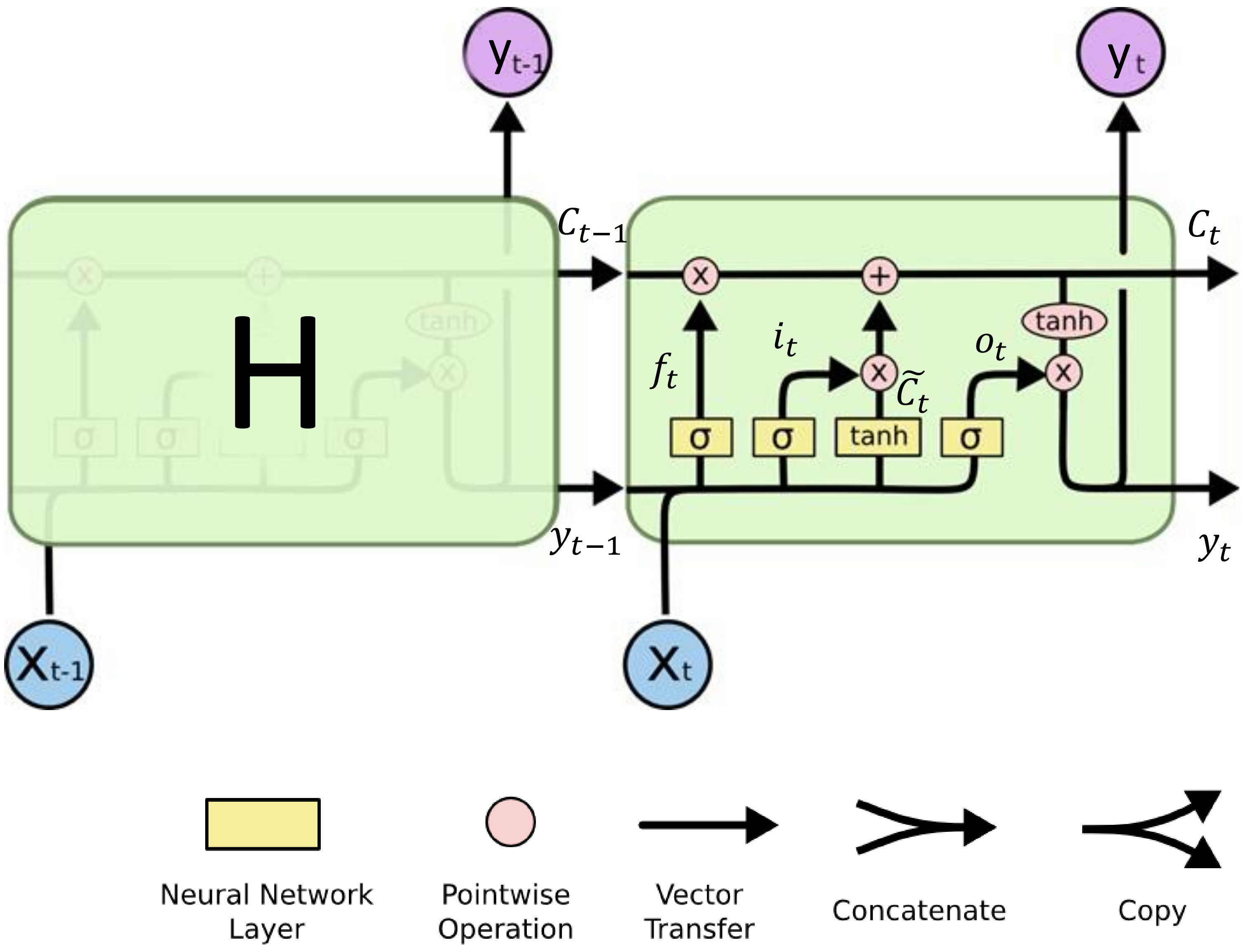
Recurrent Neural Network



$$\mathrm{H}_t = \mathrm{W}_{\rightarrow}\sigma(\mathrm{H}_{t-1}) + \mathrm{W}_{\uparrow}(\mathrm{x}_t)$$

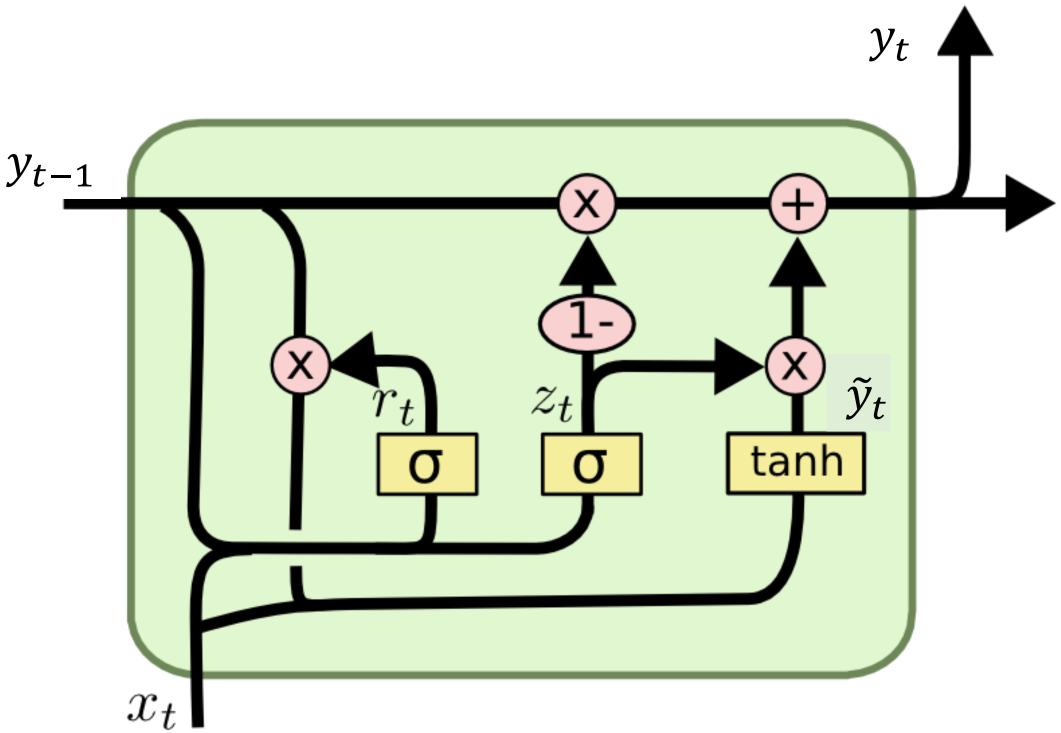
$$\hat{\mathrm{y}}_t = \mathrm{W}\sigma(\mathrm{H}_t)$$

LSTM (Long short-term memory)



$$\boldsymbol{forget\ gate}\ - f_t = \sigma\left(W_f \cdot [y_{t-1}, \mathrm{x}_t] + b_f\right)$$
$$\boldsymbol{input\ gate}\ - \ i_t = \sigma(W_i \cdot [y_{t-1}, \mathrm{x}_t] + b_i)$$
$$\widetilde{C}_t = tanh(W_c \cdot [y_{t-1}, \mathrm{x}_t] + b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t$$
$$\boldsymbol{output\ gate}\ - o_t = \sigma(W_o \cdot [y_{t-1}, \mathrm{x}_t] + b_o)$$
$$y_t = o_t * \tanh(C_t)$$

GRU (Gated Recurrent Unit)



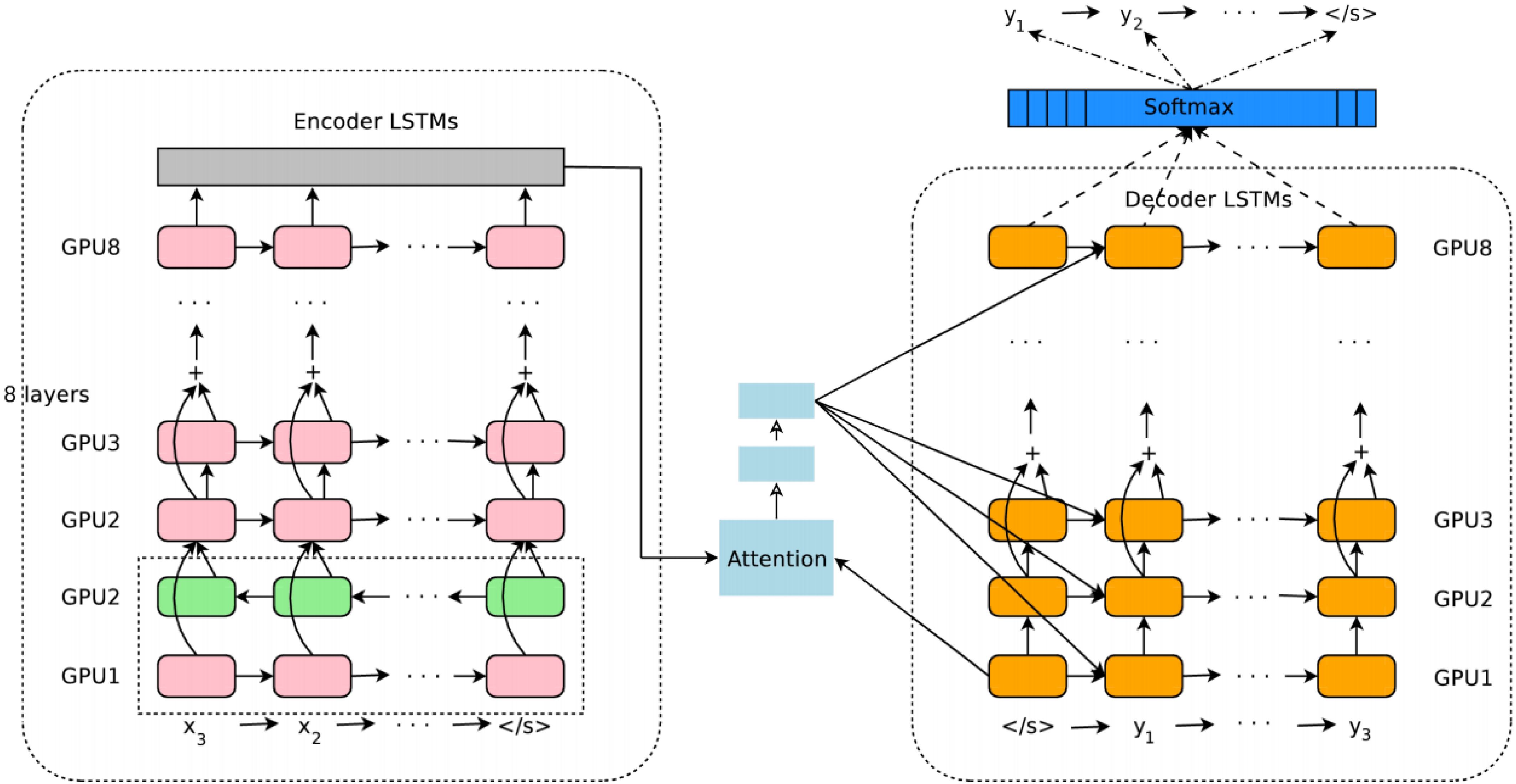
$$update\ gate\ - z_t = \sigma(\mathrm{W}_z \cdot [y_{t-1}, \mathrm{x}_t])$$

$$reset\ gate\ - r_t = \sigma(\mathrm{W}_r \cdot [y_{t-1}, \mathrm{x}_t])$$

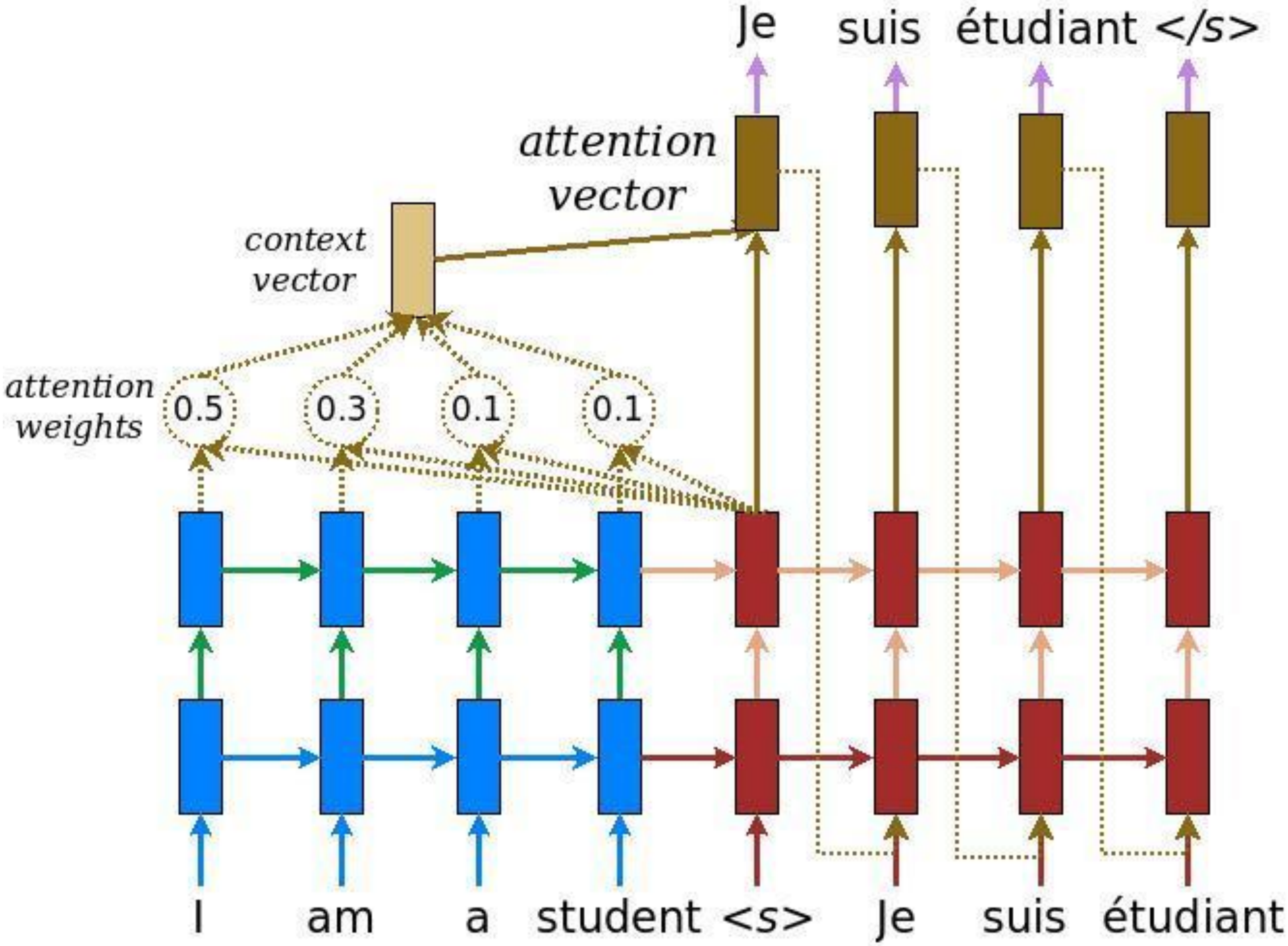
$$\tilde{y}_t = tanh(\mathrm{W} \cdot [r_t * y_{t-1}, \mathrm{x}_t])$$

$$y_t = (1 - z_t) * y_{t-1} + \mathrm{z_t} * \tilde{y}_t$$

Google Neural Machine Translation

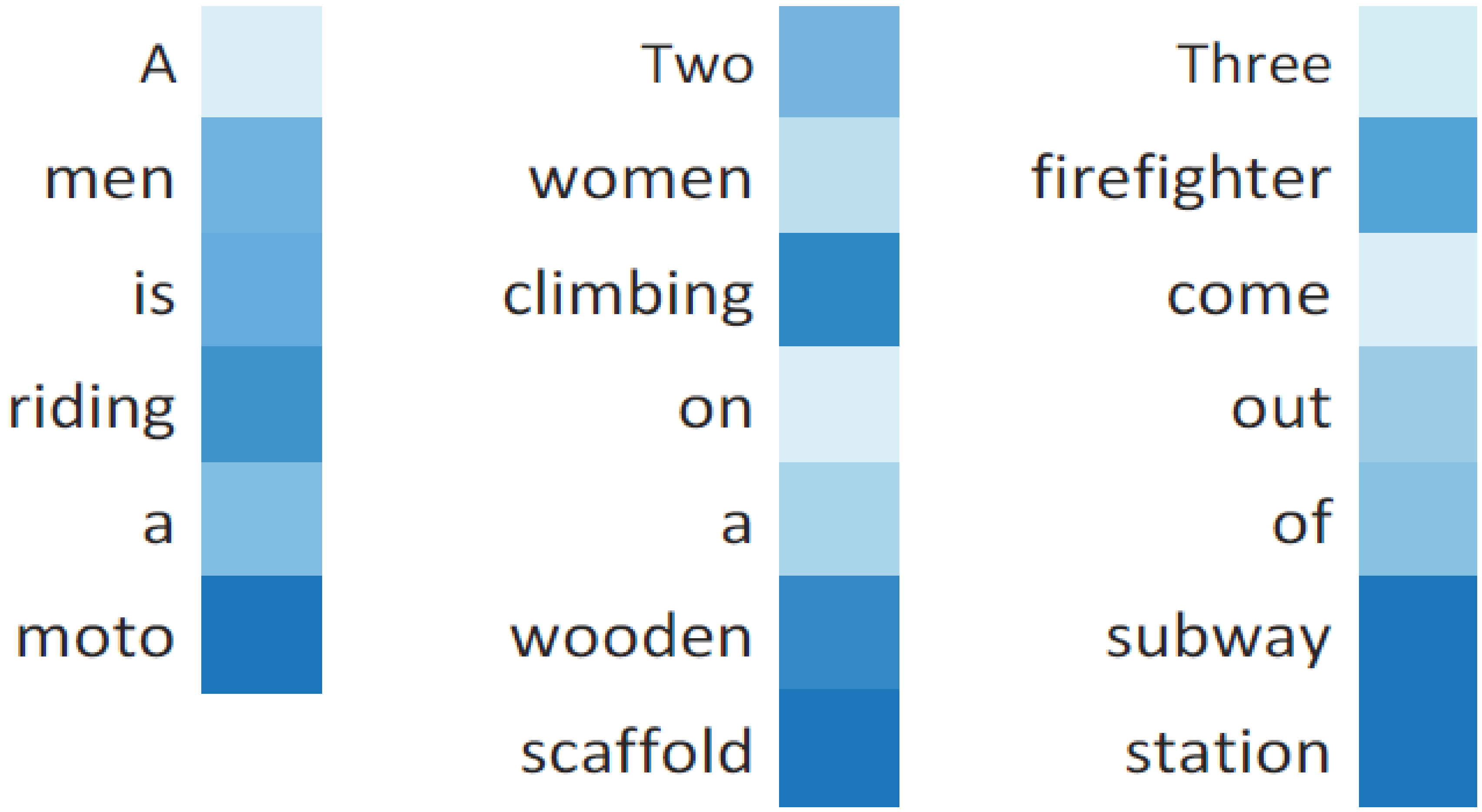


Attention

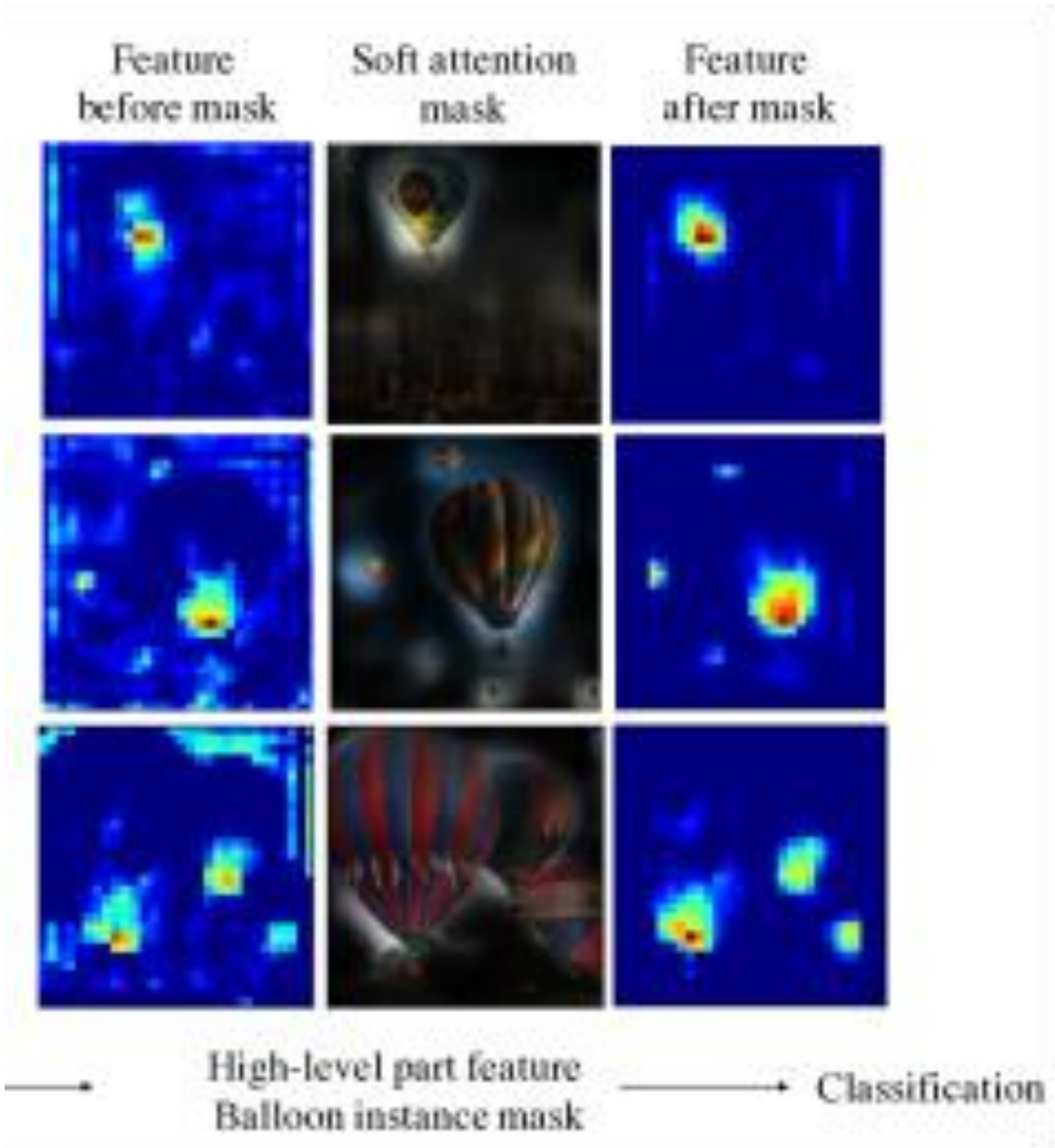


Je
suis
étudiant
</s>
attention
vector
context
vector
attention
weights
0.5
0.3
0.1
0.1
I
am
a
student
<s>
Je
suis
étudiant

Attention


A
men
is
riding
a
moto
Two
women
climbing
on
a
wooden
scaffold
Three
firefighter
come
out
of
subway
station

Visual attention



SQuAD1

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 13, 2018	nlnet (single model) *Microsoft Research Asia*	**74.238**	**77.022**
2 Sep 17, 2018	Unet (ensemble) *Fudan University & Liulishuo Lab*	71.553	75.011
2 Aug 15, 2018	Reinforced Mnemonic Reader + Answer Verifier (single model) *NUDT* https://arxiv.org/abs/1808.05759	71.699	74.238
2 Aug 28, 2018	SLQA+ (single model) *Alibaba DAMO NLP* http://www.aclweb.org/anthology/P18-1158	71.451	74.422
3 Sep 14, 2018	SAN (ensemble model) *Microsoft Business Applications Research Group* https://arxiv.org/abs/1712.03556	71.282	73.658

SQuAD2

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**88.592**	**90.859**
2 Jul 19, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk*	88.050	90.645
3 Jul 23, 2019	XLNet + SG-Net Verifier (single model) *Shanghai Jiao Tong University & CloudWalk*	87.046	89.899
3 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	87.147	89.474
3 Jul 20, 2019	RoBERTa (single model) *Facebook AI*	86.820	89.795
4 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
5 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147