كلية الهندسة
قسم هندسة الروبوت و الأنظمة الذكية

الشبكات العصبية

محاضرة 5

NLP and Word Embeddings

د. فادي متوج

Word2Vec

Skip-grams

I want a glass of orange juice to go along with my cereal.

Context	Target
orange	juice
orange	glass
orange	my

[Mikolov et. al., 2013. Efficient estimation of word representations in vector space.]

Model

Vocab size = 10,000k

$x \longrightarrow y$

Content c ("orange") 6257 $\longrightarrow$ Target t ("juice") 4834

$O_c \to E \to e_c \to O \to \hat{y}$ (Softmax)

$e_c = E O_c$

Softmax: $p(t|c) = \dfrac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$

θ_t = parameter associated with output t

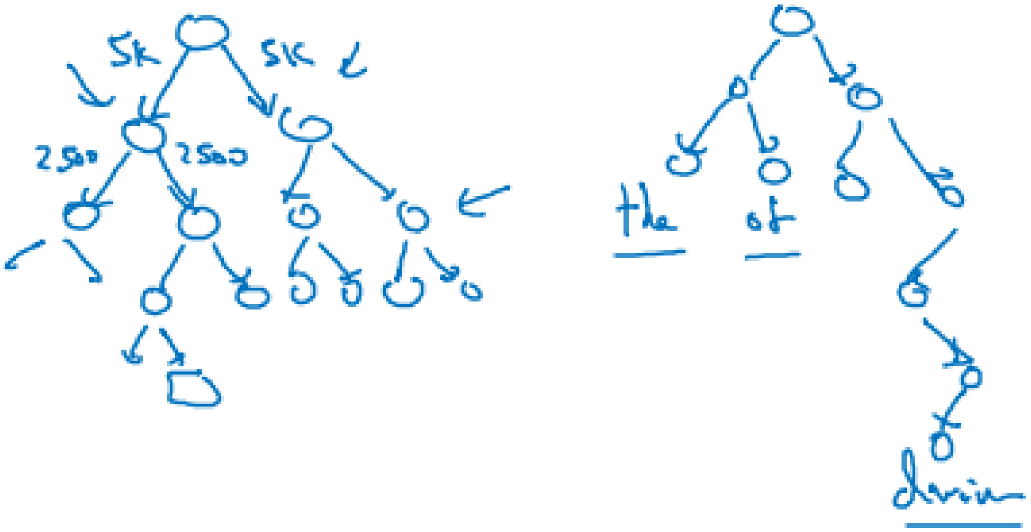
$\rightarrow \mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{10,000} y_i \log \hat{y}_i$

$y = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow 4834$

Problems with softmax classification

$$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$$

Hierarchical Softmax

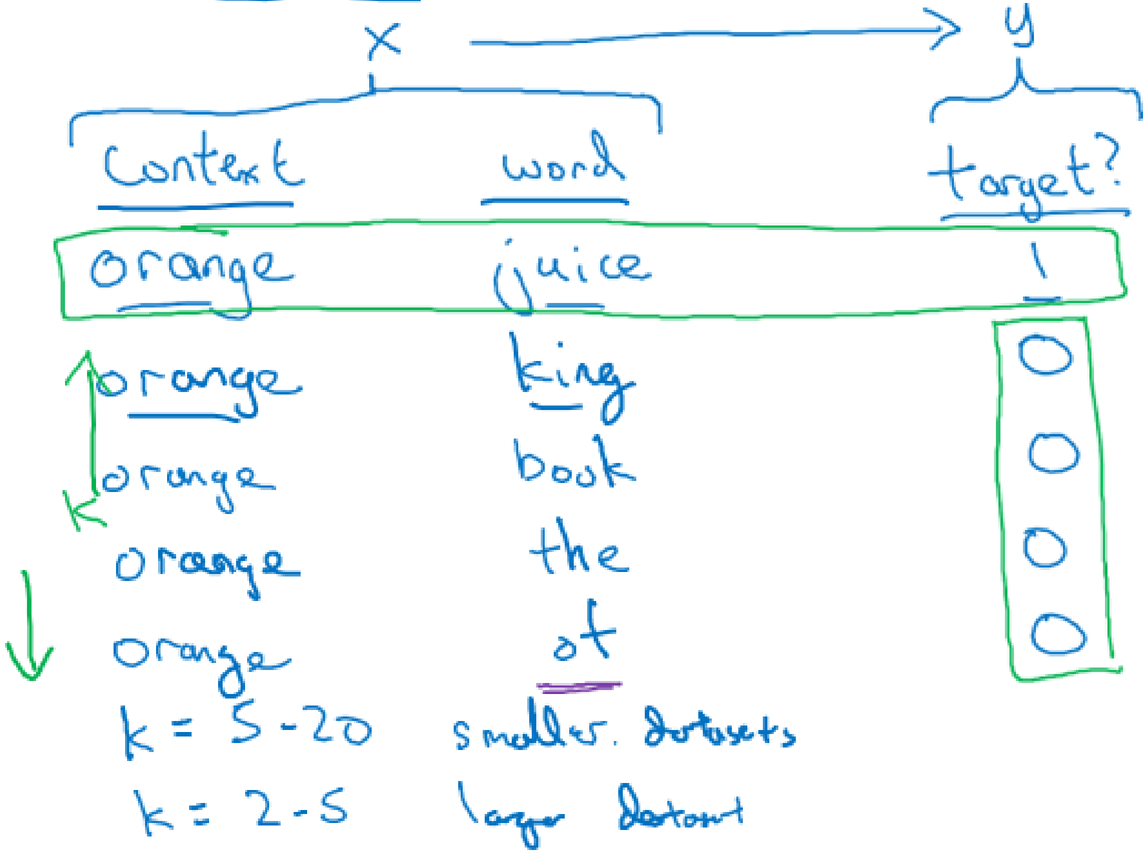


Negative sampling

Defining a new learning problem

جَامعة
المَنارة

I want a glass of orange juice to go along with my cereal.



[Mikolov et. al., 2013. Distributed representation of words and phrases and their compositionality]

Model

جامعة المنارة

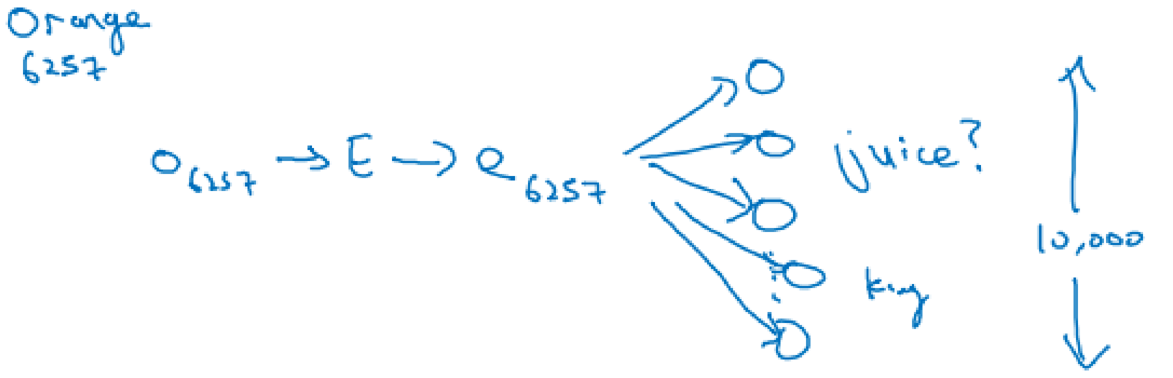
Softmax: $$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$$

10,000-way softmax

$$P(y=1 \mid c,t) = \sigma(\theta_t^T e_c)$$

context	word	target?
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

x = (context, word), y = target?; c, t, y; k



10,000 binary classification problem

k+1

GloVe word vectors

GloVe (global vectors for word representation)

I want a glass of orange juice to go along with my cereal

c, t

X_{ij} = # times i appears in context of j.

($i \uparrow t$, $j \uparrow c$)

$X_{ij} = X_{ji}$ ←

[Pennington et. al., 2014. GloVe: Global vectors for word representation]

Model

minimize $\sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(X_{ij}) \left(\theta_i^T e_j + b_i + b'_j - \log X_{ij} \right)^2$

t c

$0?$

"$\theta_t^T e_c$"

weighting term

$f(X_{ij}) = 0$ if $X_{ij} = 0$.

"$0 \log 0$" $= 0$

this, is, of, a, ...

durian

Sentiment classification

Sentiment classification problem

جامعة المنارة

$x \longrightarrow y$

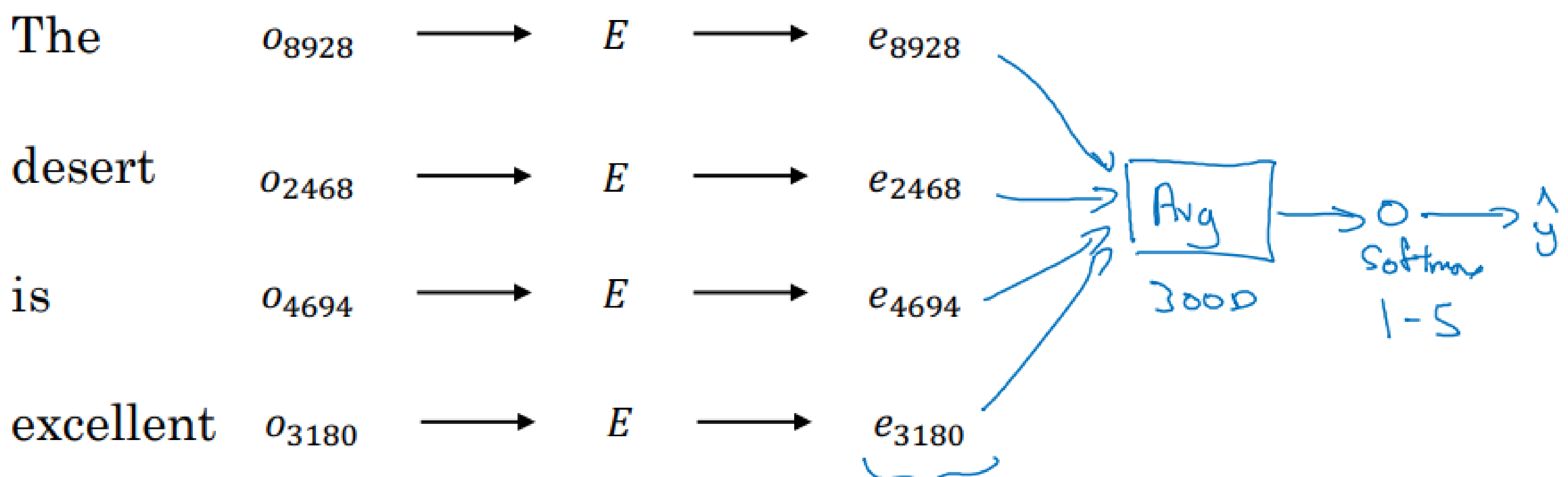
x	y
The dessert is excellent.	★★★★☆
Service was quite slow.	★★☆☆☆
Good for a quick meal, but nothing special.	★★★☆☆
Completely lacking in good taste, good service, and good ambience.	★☆☆☆☆

10,000 → 100,000 words

Simple sentiment classification model

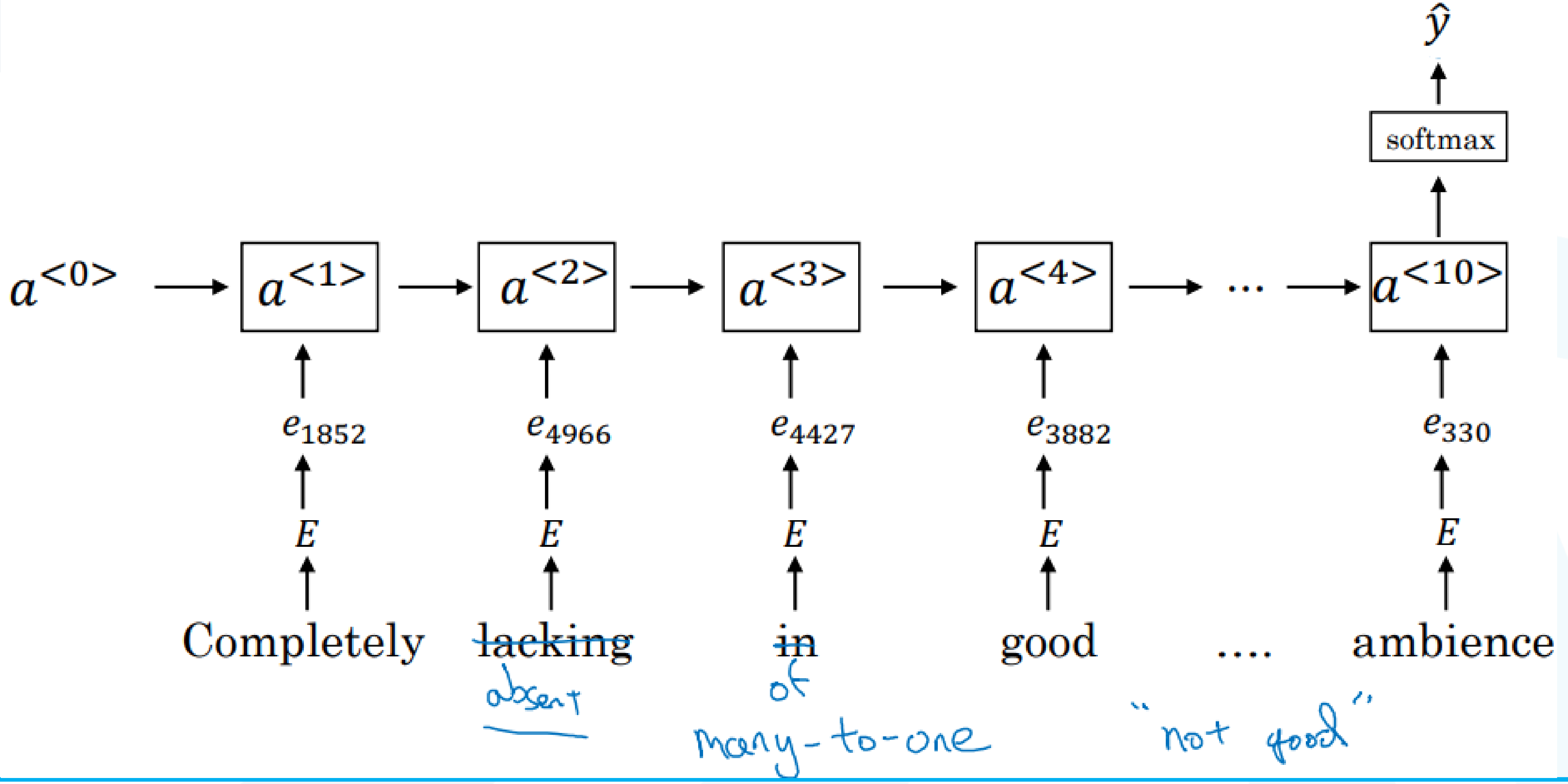
The dessert is excellent ★★★★☆

8928 2468 4694 3180



"Completely lacking in good taste, good service, and good ambience."

RNN for sentiment classification



Debiasing word embeddings

The problem of bias in word embeddings

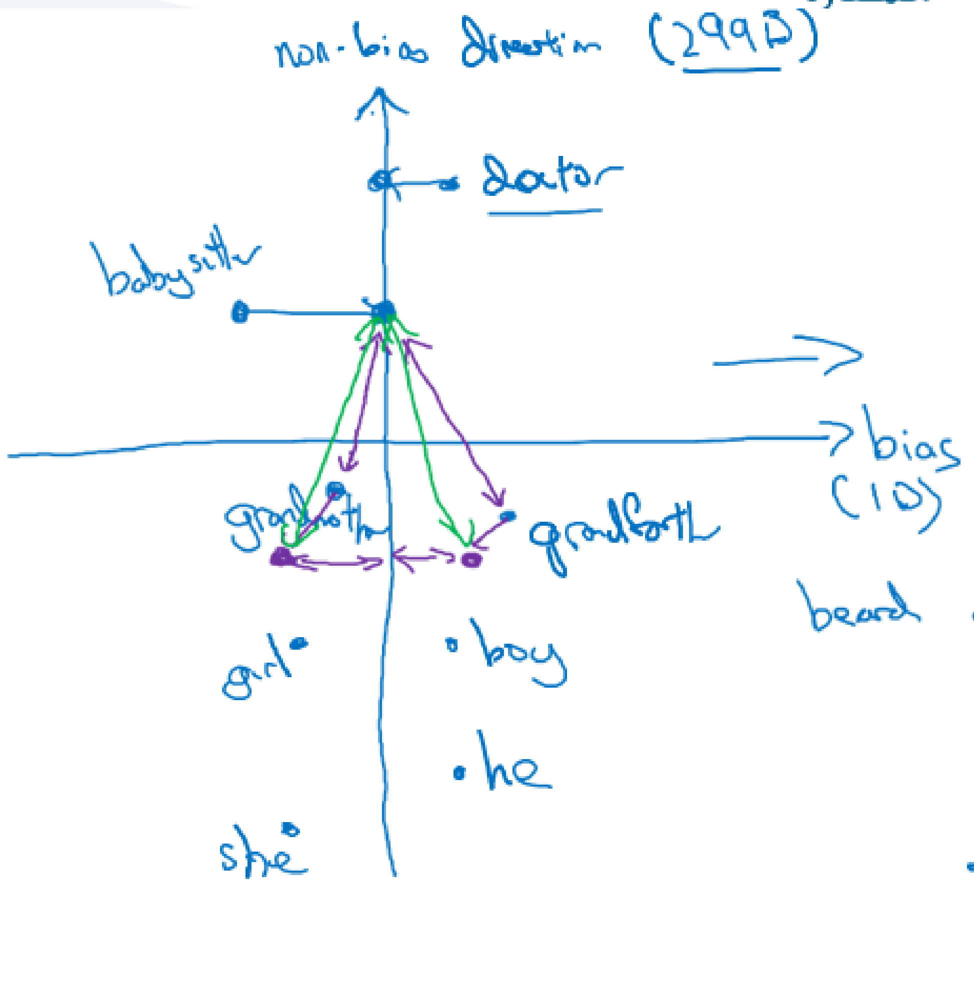
Man:Woman as King:Queen

Man:Computer_Programmer as Woman:

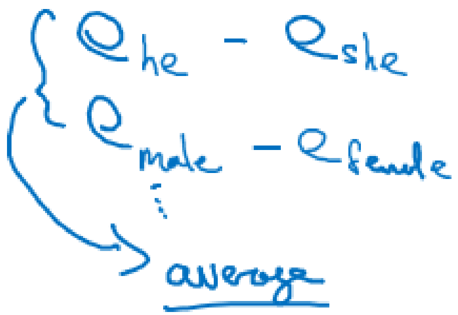
Father:Doctor as Mother:

Word embeddings can reflect gender, ethnicity, age, and other biases of the text used to train the model.

Addressing bias in word embeddings

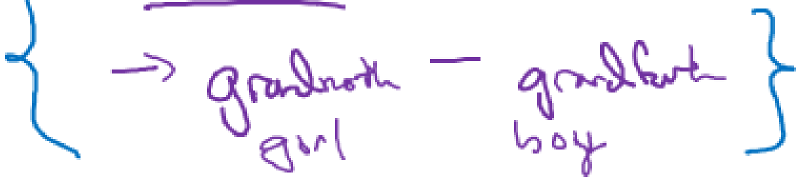


1. Identify bias direction.



2. Neutralize: For every word that is not definitional, project to get rid of bias.

3. Equalize pairs.



[Bolukbasi et. al., 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings]

Sequence to sequence models

Basic models

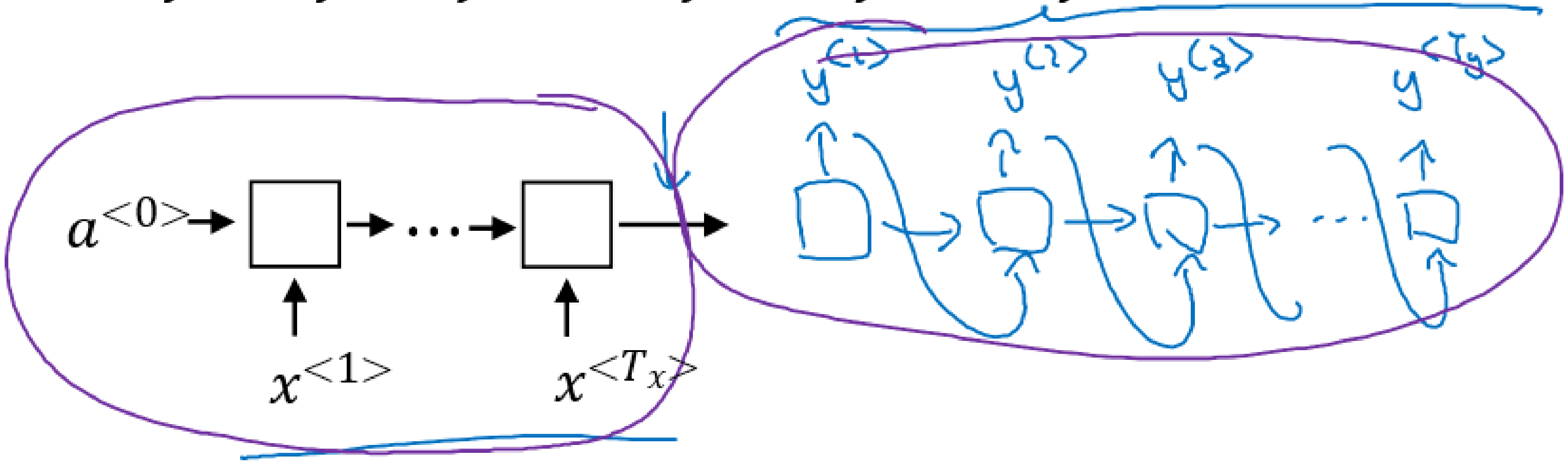
Sequence to sequence model

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ $x^{<4>}$ $x^{<5>}$

Jane visite l'Afrique en septembre

→ Jane is visiting Africa in September.

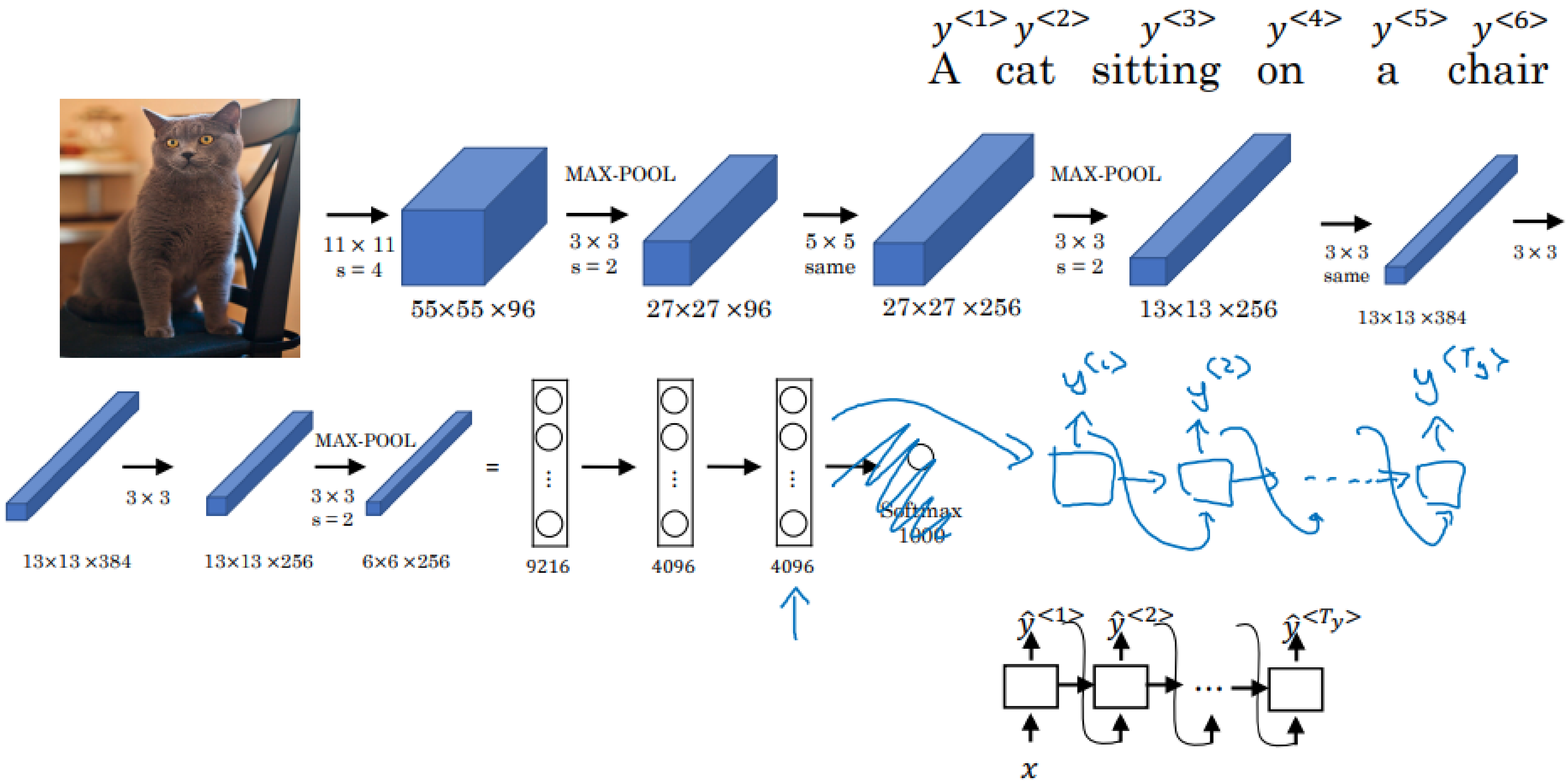
$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$



[Sutskever et al., 2014. Sequence to sequence learning with neural networks]

[Cho et al., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation]

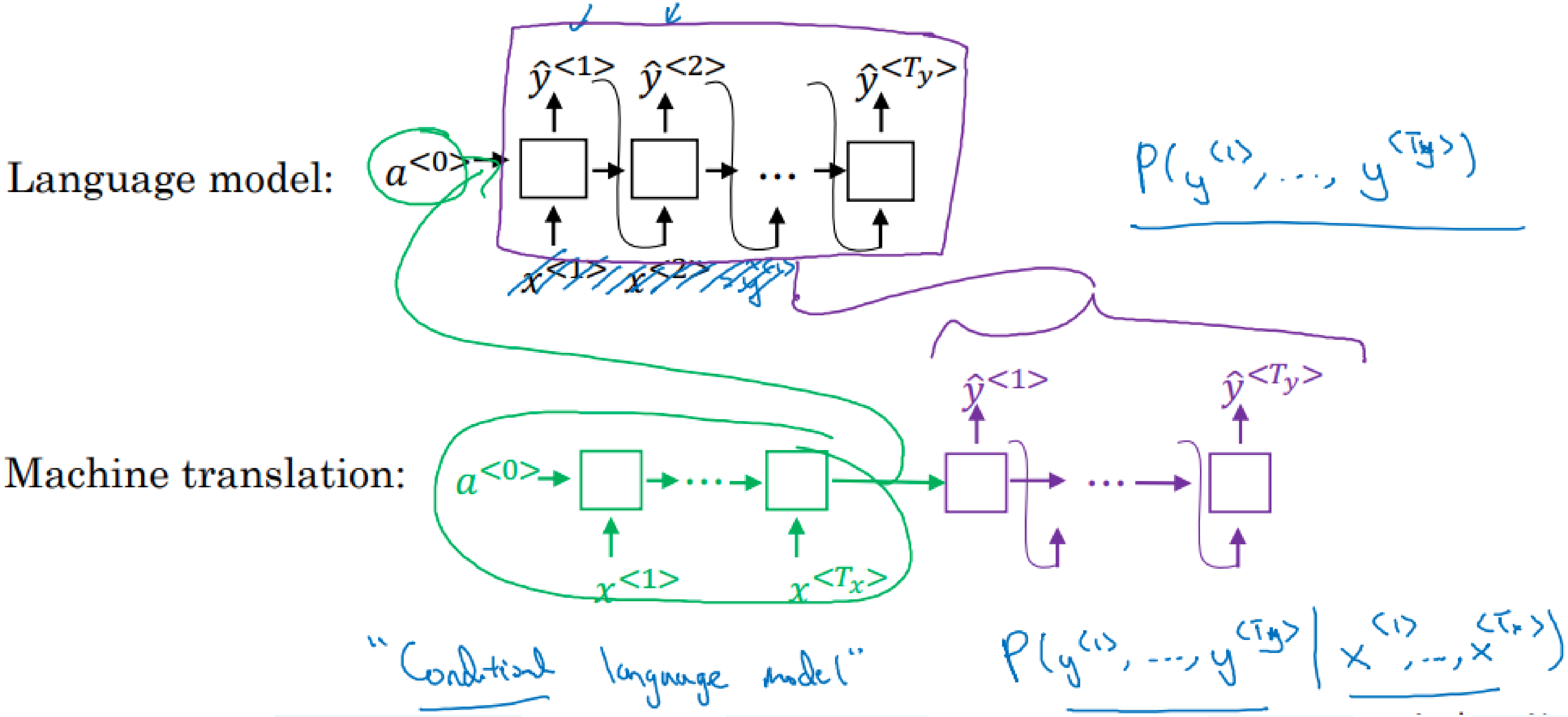
Image captioning



[Mao et. al., 2014. Deep captioning with multimodal recurrent neural networks]
[Vinyals et. al., 2014. Show and tell: Neural image caption generator]
[Karpathy and Li, 2015. Deep visual-semantic alignments for generating image descriptions]

Picking the most Likely sentence

Machine translation as building a conditional language mode



Finding the most likely translation

Jane visite l'Afrique en septembre.

$$P(y^{<1>}, \ldots, y^{<T_y>} \mid x)$$

→ Jane is visiting Africa in September.

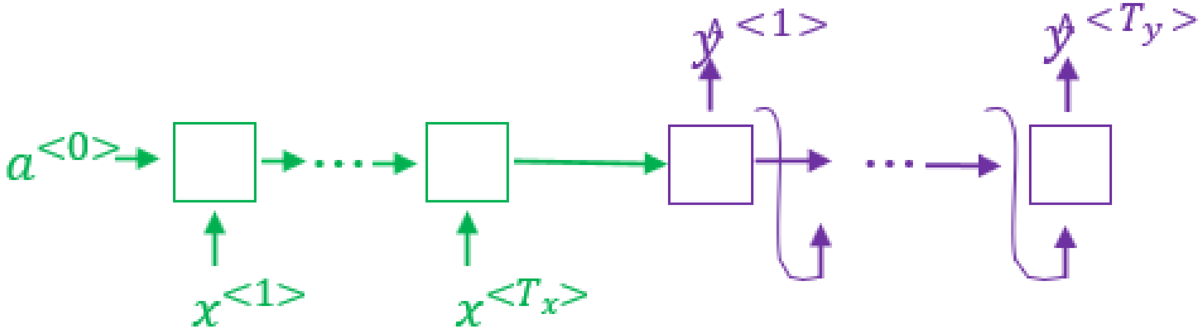
→ Jane is going to be visiting Africa in September.

→ In September, Jane will visit Africa.

→ Her African friend welcomed Jane in September.

$$\underset{y^{<1>},\ldots,y^{<T_y>}}{\arg\max} \; P(y^{<1>}, \ldots, y^{<T_y>} \mid x)$$

Why not a greedy search?



$$\arg\max_{y} P(\hat{y}^{<1>}, \hat{y}^{<2>}, \ldots, \hat{y}^{<T_y>} \mid x)$$

→ Jane is visiting Africa in September.

→ Jane is going to be visiting Africa in September.

$$P(\text{Jane is going} \mid x) > P(\text{Jane is visiting} \mid x)$$