Machine Learning 2

DS 4420 - Spring 2020

Self-supervised learning

Byron C Wallace

Today

- Auto-Encoders
- “Self-Supervised” learning as a general paradigm

Auto-Encoders

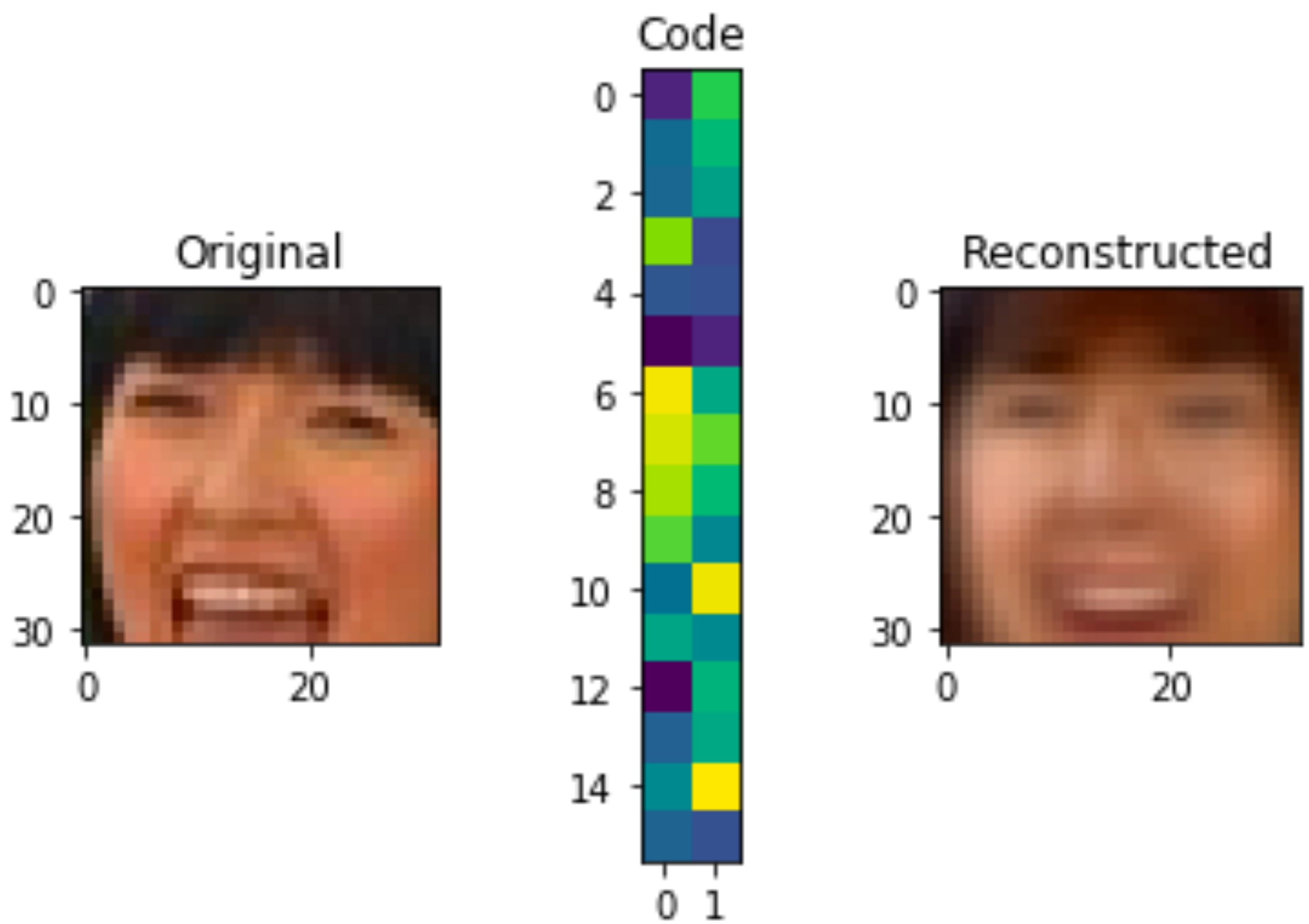
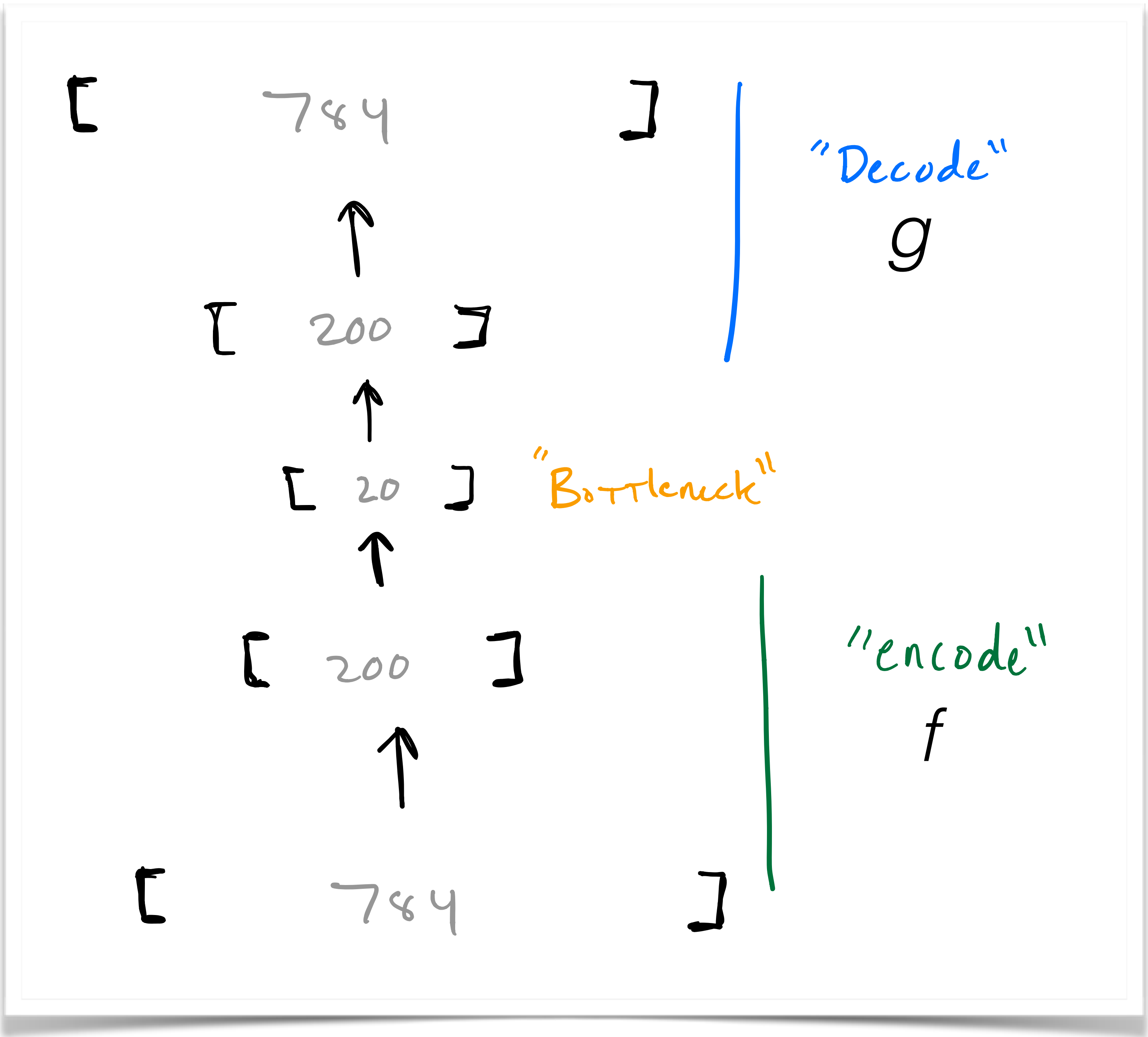


Figure credit: https://stackabuse.com/autoencoders-for-image-reconstruction-in-python-and-keras/



[784]
"Decode"
g
[200]
[20]
"Bottleneck"
[200]
"encode"
f
[784]

Loss

$$L(\boldsymbol{x}, g(f(\boldsymbol{x})))$$

Loss

$$L(\boldsymbol{x}, g(f(\boldsymbol{x})))$$

- Both *f* and *g* are parameterized

Loss

$$L(\boldsymbol{x}, g(f(\boldsymbol{x})))$$

- Both *f* and *g* are parameterized
- If *L* is the MSE and *f*, *g* are linear, then this is PCA

“code”

$$z = f(x)$$

$$\tilde{x} = g(z)$$

“code”

$$z = f(x)$$

$$\tilde{x} = g(z)$$

- Set z to be (much) lower dim than x: **Undercomplete**

Overfitting

- An issue with auto-encoders: Even if h is relatively low-dimensional, if we have a deep auto-encoder (many params) the model might not learn anything particularly useful

Overfitting

- An issue with auto-encoders: Even if *h* is relatively low-dimensional, if we have a deep auto-encoder (many params) the model might not learn anything particularly useful

- Solution: *Regularized* auto-encoders

$$L(x, g(f(x)) + \Omega(z)$$

Note: We can use this function to bake-in other constraints and inductive biases as well

$$\Omega(z)$$

Probabilistic view

- Another means of regularizing *z* involves imposing a prior, similar to PPCA.

$$p(x, z) = p(z)p(x|z)$$

Probabilistic view

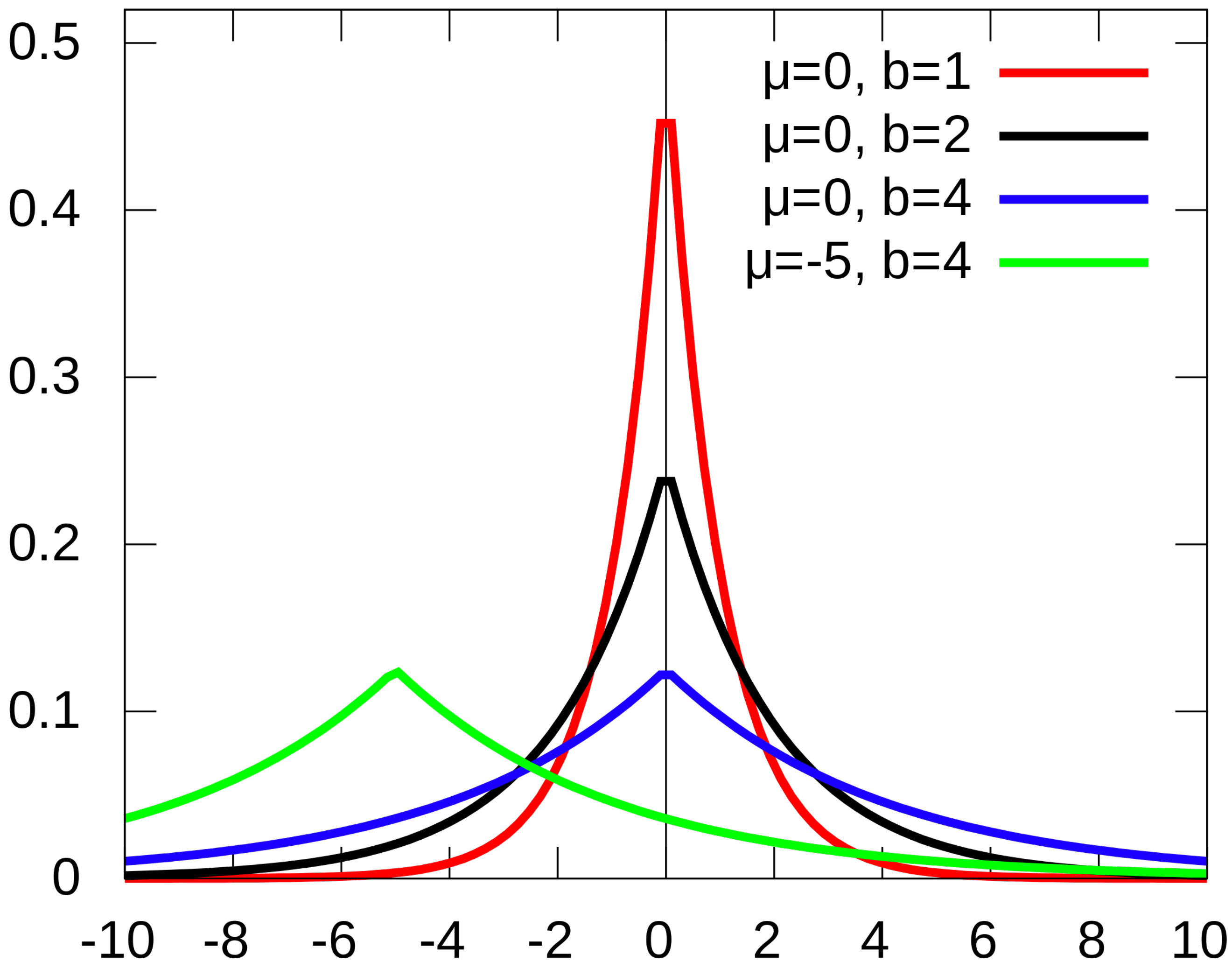
- Another means of regularizing *z* involves imposing a prior, similar to PPCA.

$$p(x, z) = p(z)p(x|z)$$

$$\log(p(x, z)) = \log(p(z)) + \log(p(x|z))$$

Inducing sparsity

Idea: Pick a prior to encourage 0s



Laplacian prior ~ L1 Norm

Inducing sparsity

$$\Omega(z) = \|h\|_1 = \sum_j |z_j|$$

Inducing sparsity

$$\Omega(z) = \|h\|_1 = \sum_j |z_j|$$

Can be combined with a ReLU to get actual 0s

Denoising auto-encoders

Instead of the typical auto-encoder loss:

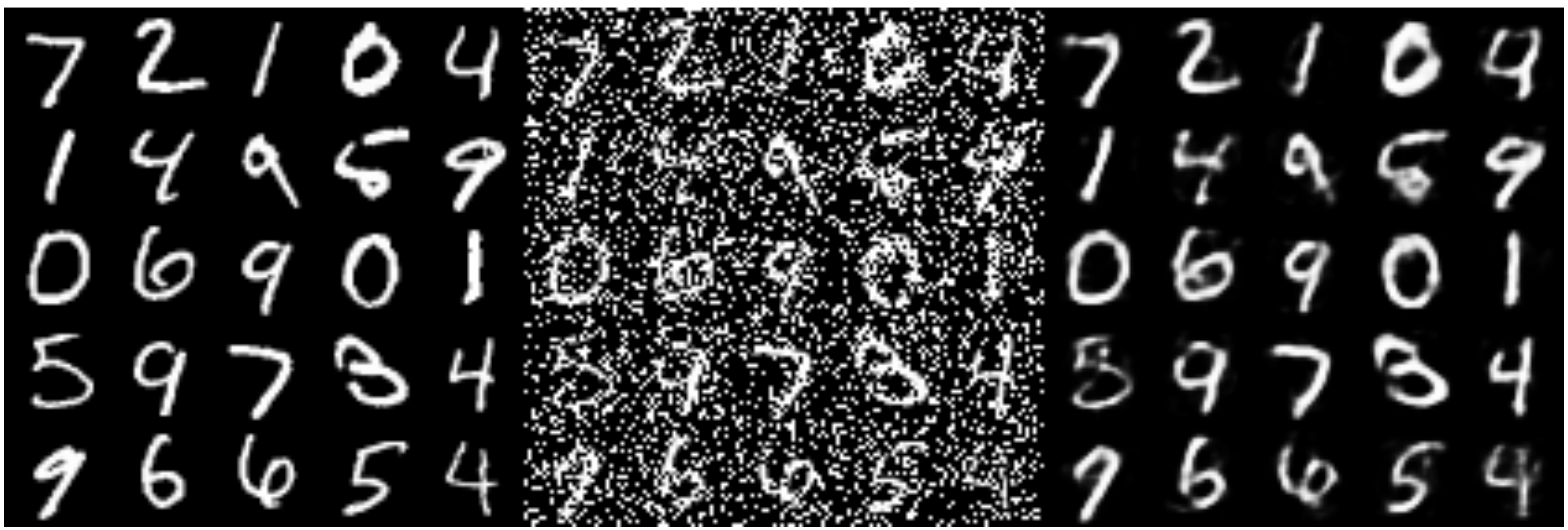
$$L(x, g(f(x)))$$

Attempt to reconstruct the input from a *corrupted* version

$$L(x, g(f(x')))$$

Denoising auto-encoders

x x' $g(f(x))$



$$L(x, g(f(x')))$$

Let’s play around a bit in torch…

[notebook/exercise: get starter from blackboard!]

Variational AEs

(see notes and notebook)

Self-supervision in vision and NLP

"If intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake. We know how to make the icing and the cherry, but we don't know how to make the cake." — Yann LeCun

Self-supervised learning in images

These slides are derived from Andrew Zisserman's materials: https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf, which in turn include content from: Carl Doersch, Ishan Misra, Andrew Owens, Carl Vondrick, Richard Zhang

- Self-supervision: A form of **unsupervised** learning in which the data itself provides the **supervision**

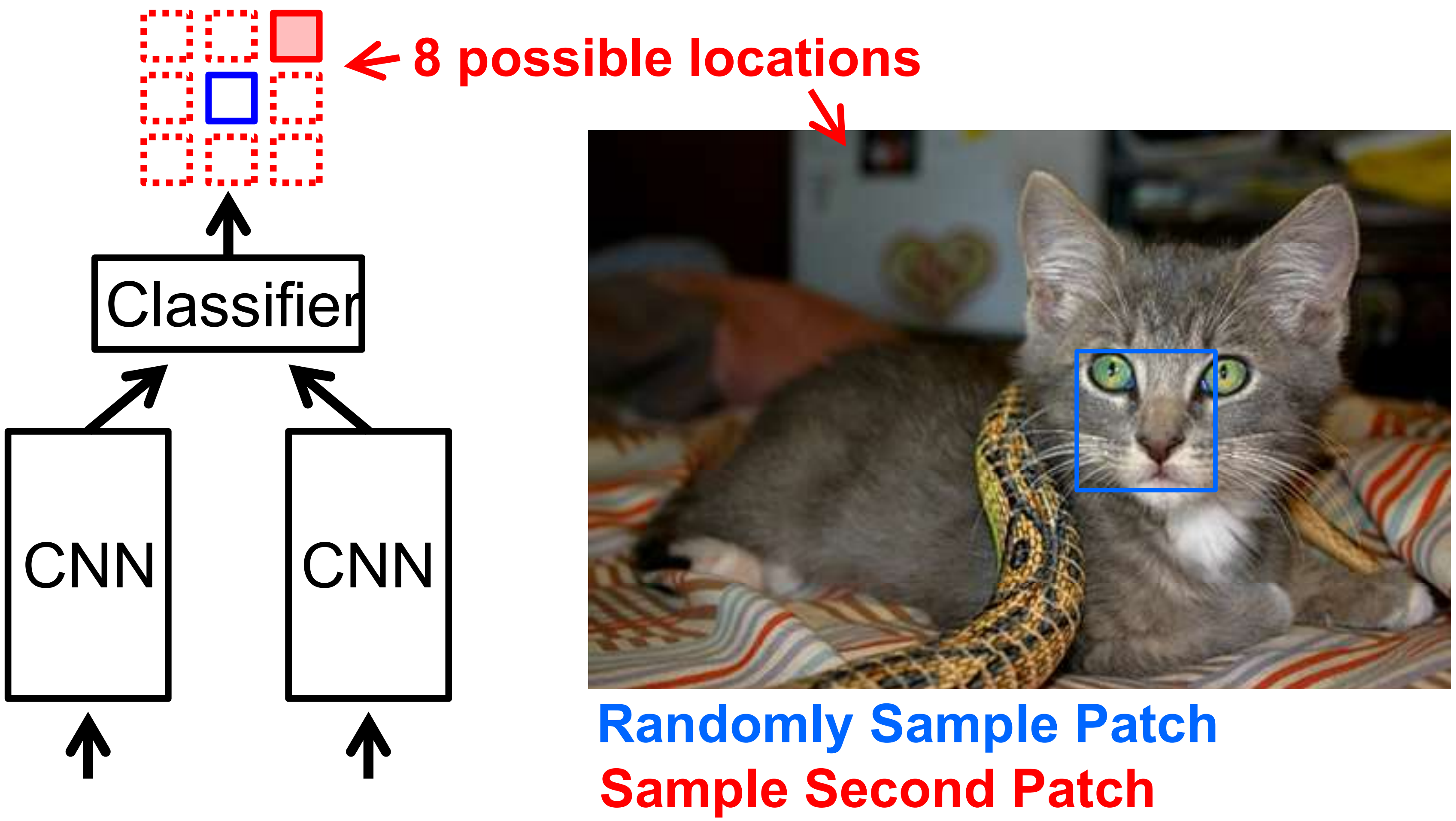
- Self-supervision: A form of **unsupervised** learning in which the data itself provides the **supervision**

- Generally: Hide some aspect of the data, attempt to reconstruct it from the rest

- Self-supervision: A form of **unsupervised** learning in which the data itself provides the **supervision**
- Generally: Hide some aspect of the data, attempt to reconstruct it from the rest
- Formulating “good” self-training objectives is an active area of research!

Example: Relative positioning

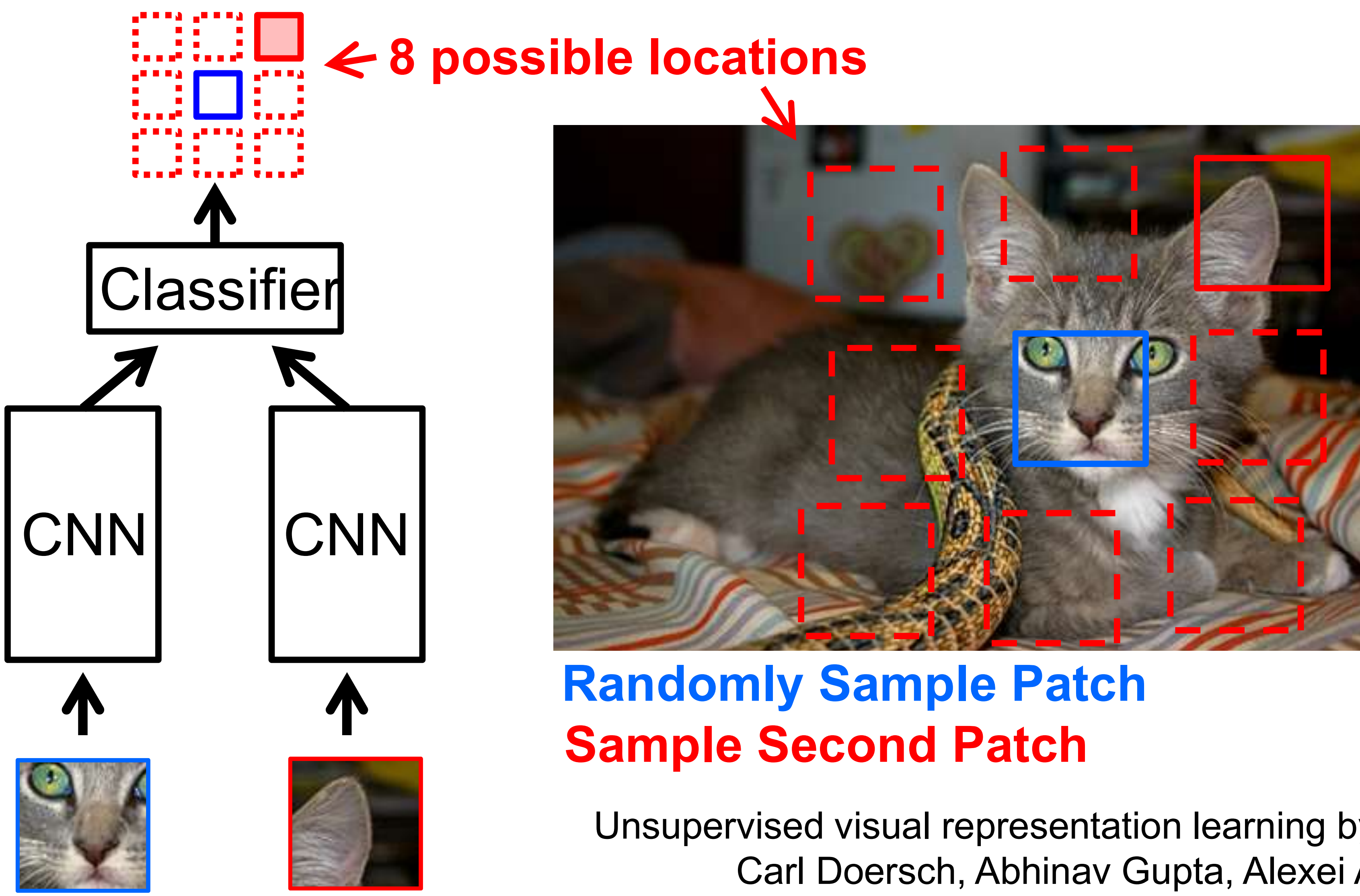
Train network to predict relative position of two regions in the same image



Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Example: Relative positioning

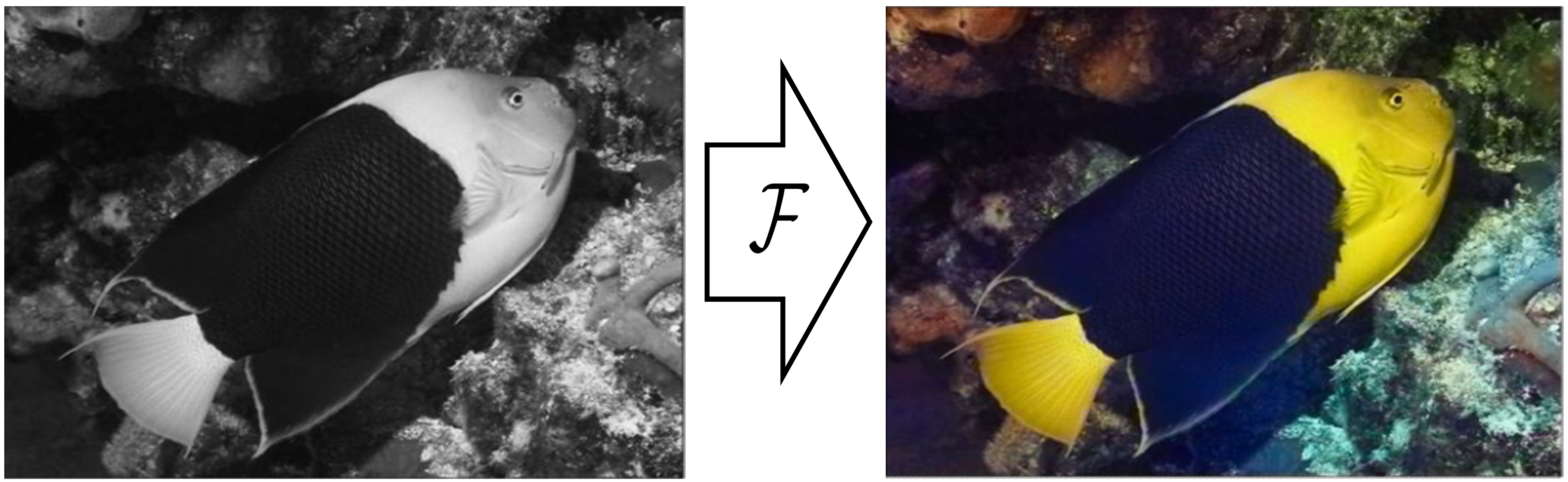
Train network to predict relative position of two regions in the same image



Unsupervised visual representation learning by context prediction,
Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

Example: Colorizing

Train network to predict pixel colour from a monochrome input

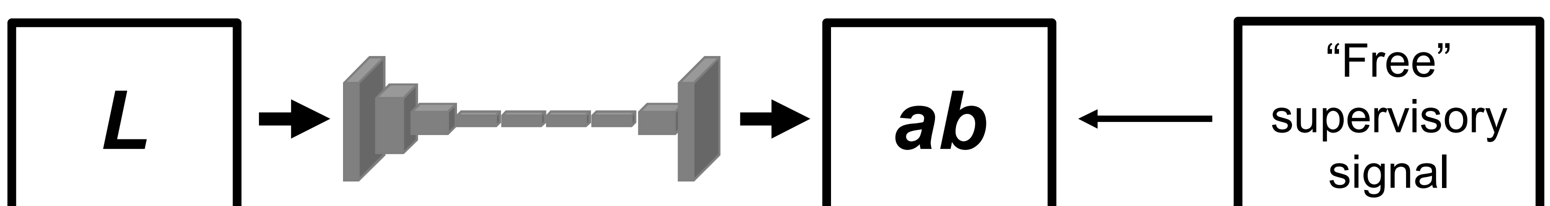


Grayscale image: *L* channel

$$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$$

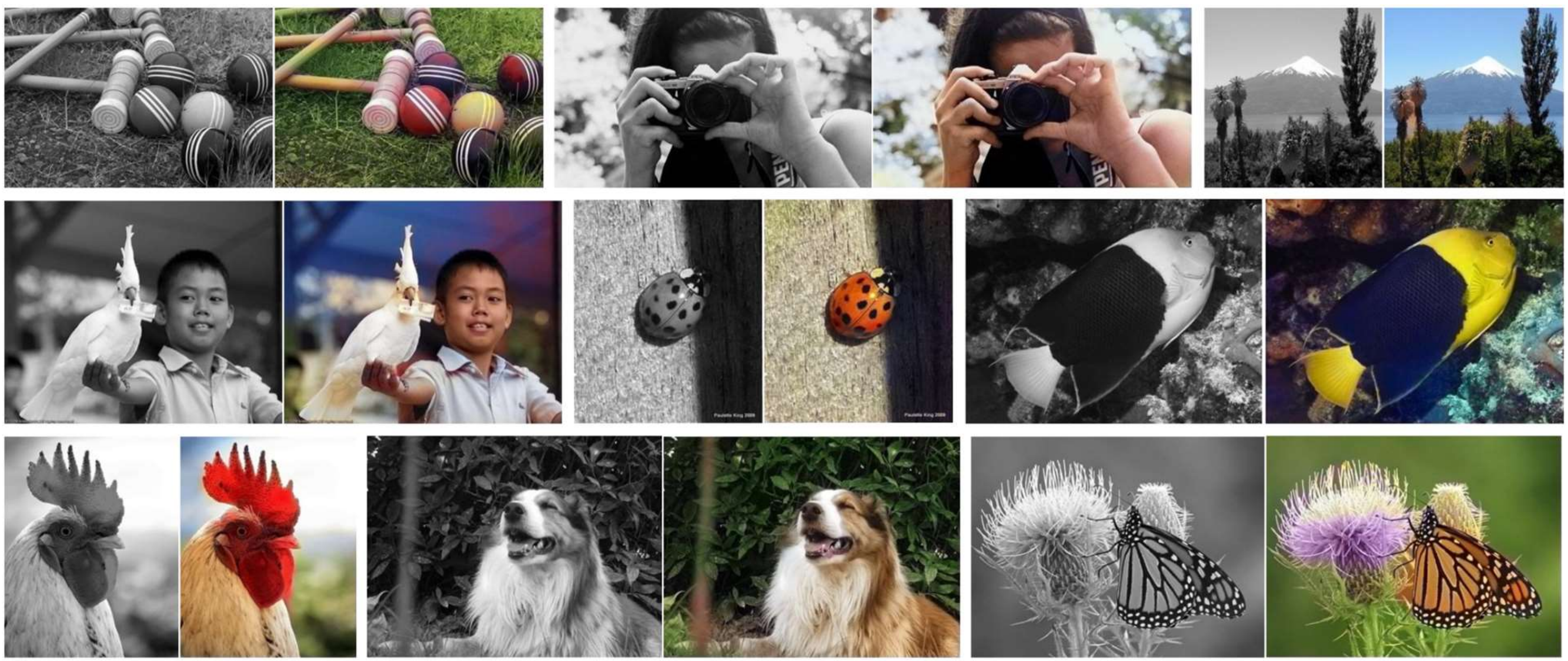
Concatenate (L,ab)

$$(\mathbf{X}, \widehat{\mathbf{Y}})$$



Example: Colorizing

Train network to predict pixel colour from a monochrome input



Example: Rotation

Which image has the correct rotation?

Unsupervised representation learning by predicting image rotations,
Spyros Gidaris, Praveer Singh, Nikos Komodakis, ICLR 2018

Self-supervision in NLP

Learning to embed words

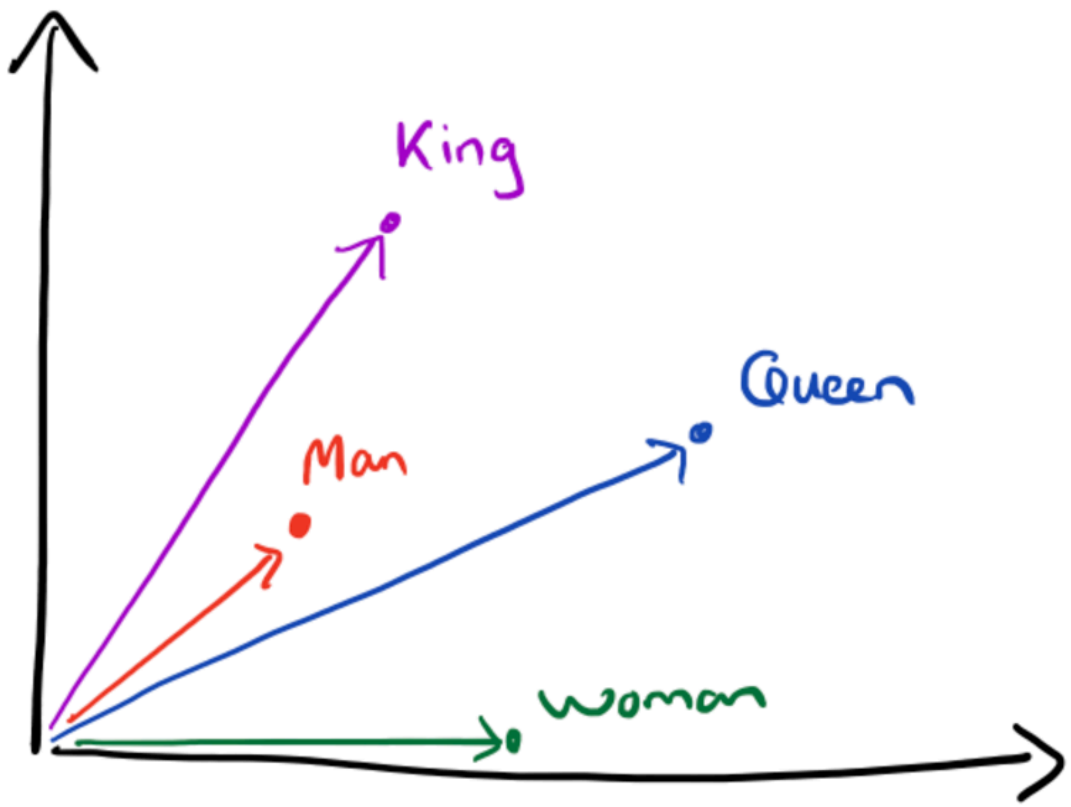
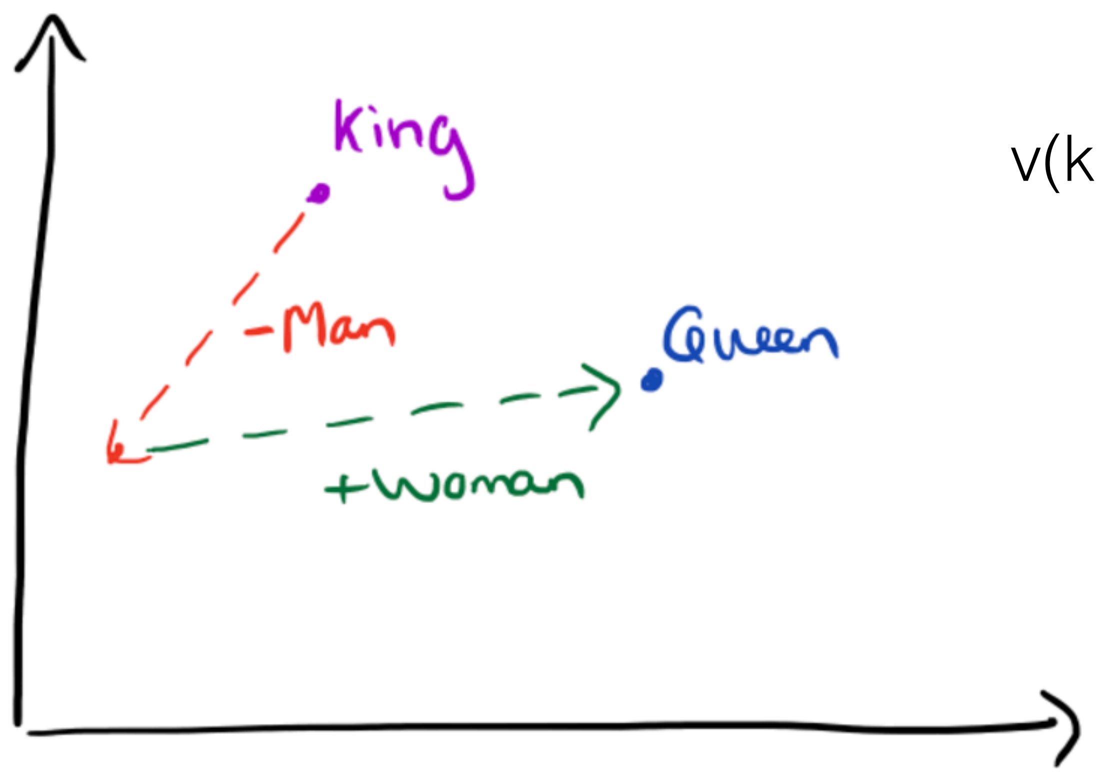


image credit: adrian colyer
https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

Learning to embed words



v(king) - v(woman) =
v(queen)

image credit: adrian colyer
https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

How do we learn these?

How do we learn these?

Distributed representations of words and phrases and their compositionality [PDF] nips.cc

T Mikolov, I Sutskever, K Chen, GS Corrado… - Advances in neural …, 2013 - papers.nips.cc Paperpile

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several improvements that make the Skip-gram model more expressive and enable it to learn higher quality vectors more rapidly. We show that by subsampling frequent words we obtain significant speedup, and also learn higher quality representations as measured by our tasks. We also introduce …

Cited by 17831 Related articles All 47 versions Import into BibTeX

One way: *word2vec*

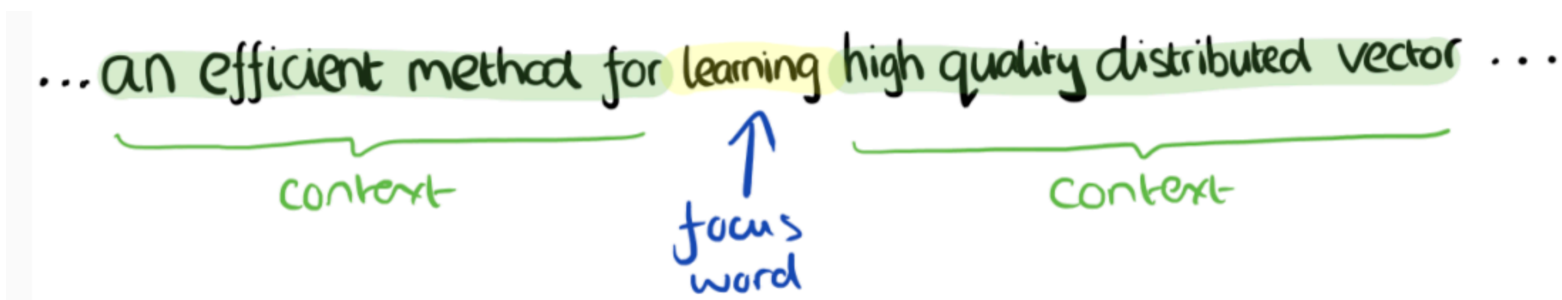


image credit: adrian colyer
https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

Constructing self supervision

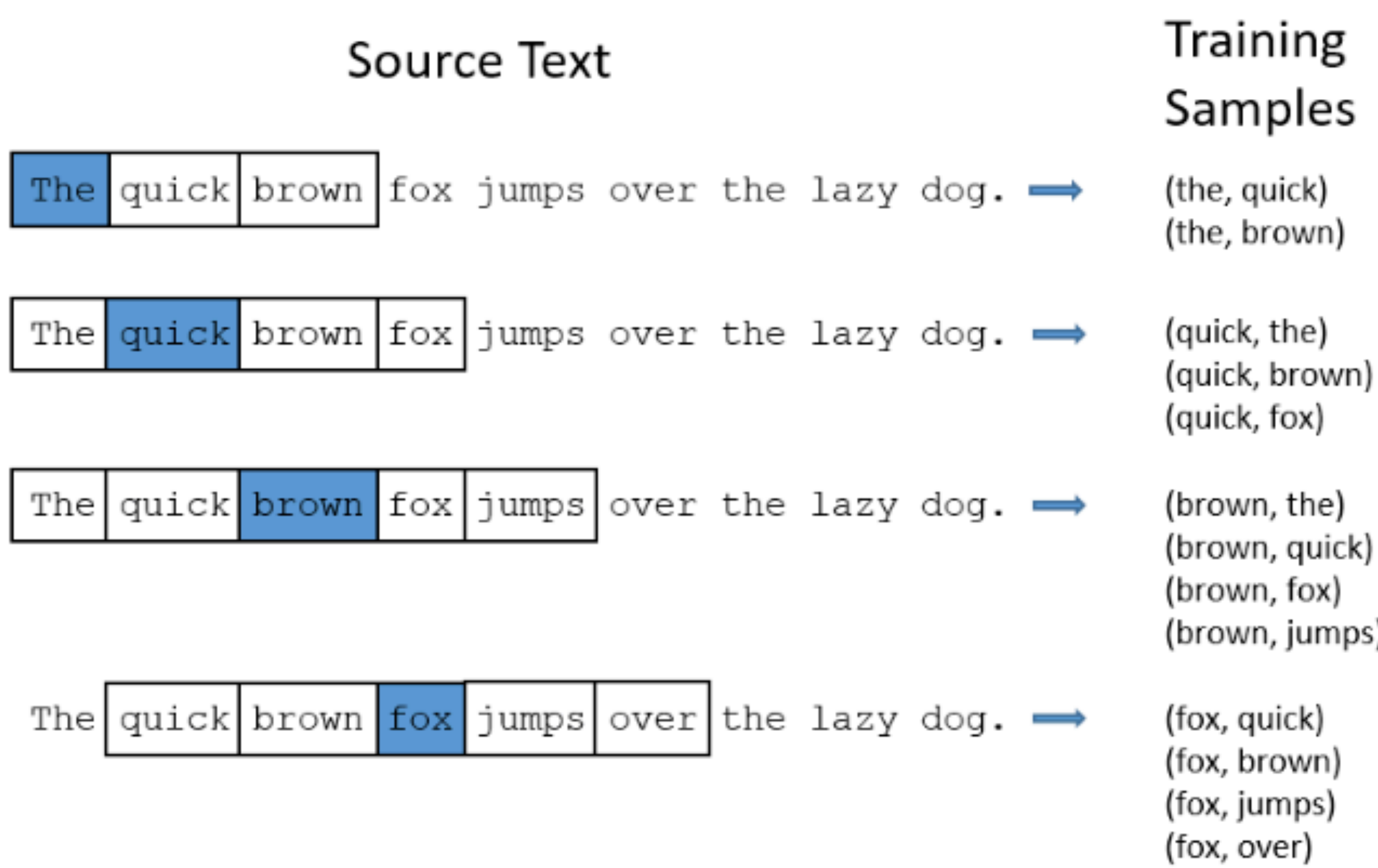


Image credit: Chris McCormick
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

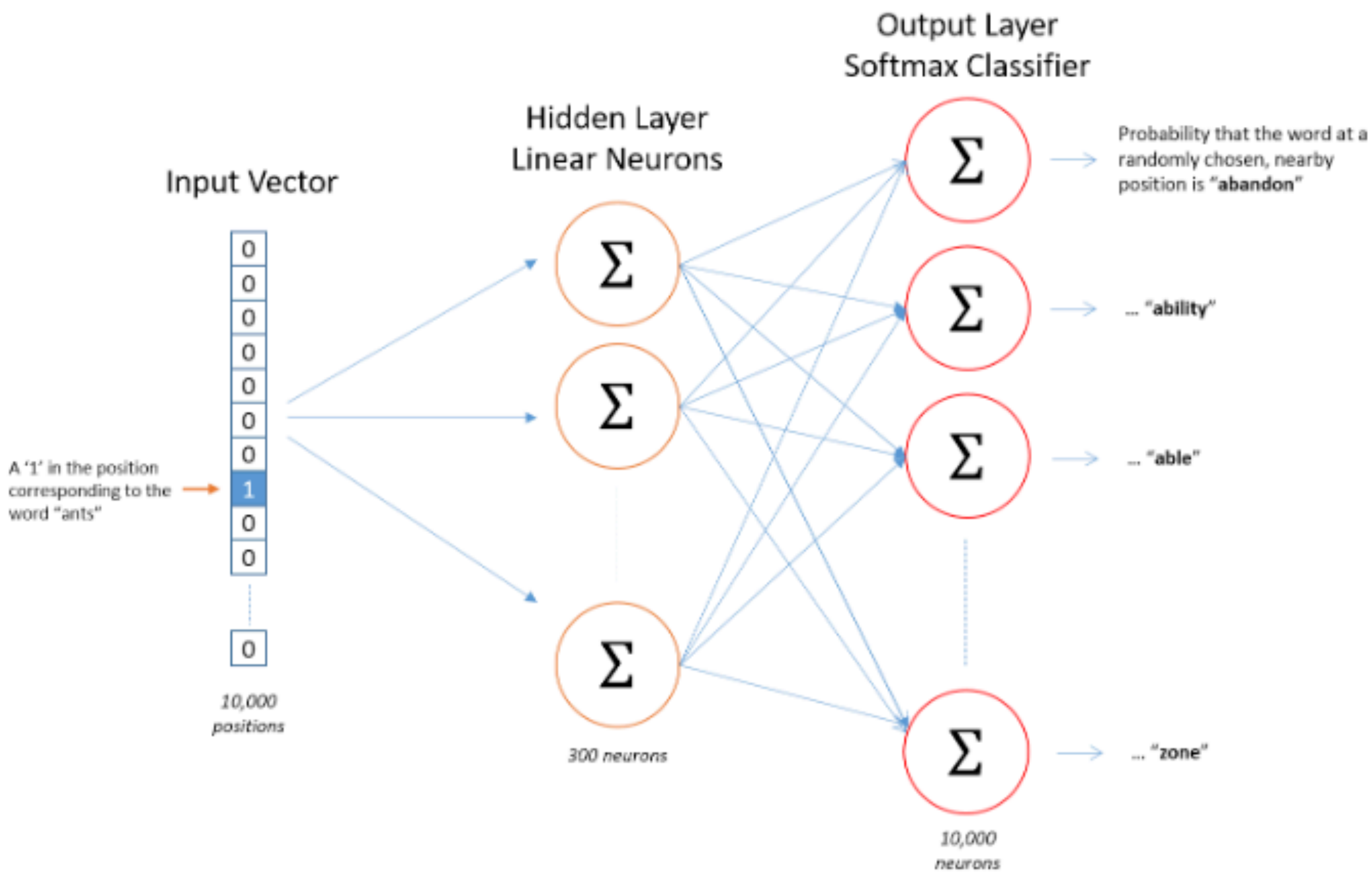


Image credit: Chris McCormick
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

$$\begin{bmatrix} 0 & 0 & 0 & \boxed{1} & 0 \end{bmatrix} \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ \boxed{10 \quad 12 \quad 19} \\ 11 & 18 & 25 \end{bmatrix} = \begin{bmatrix} 10 & 12 & 19 \end{bmatrix}$$

Image credit: Chris McCormick
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

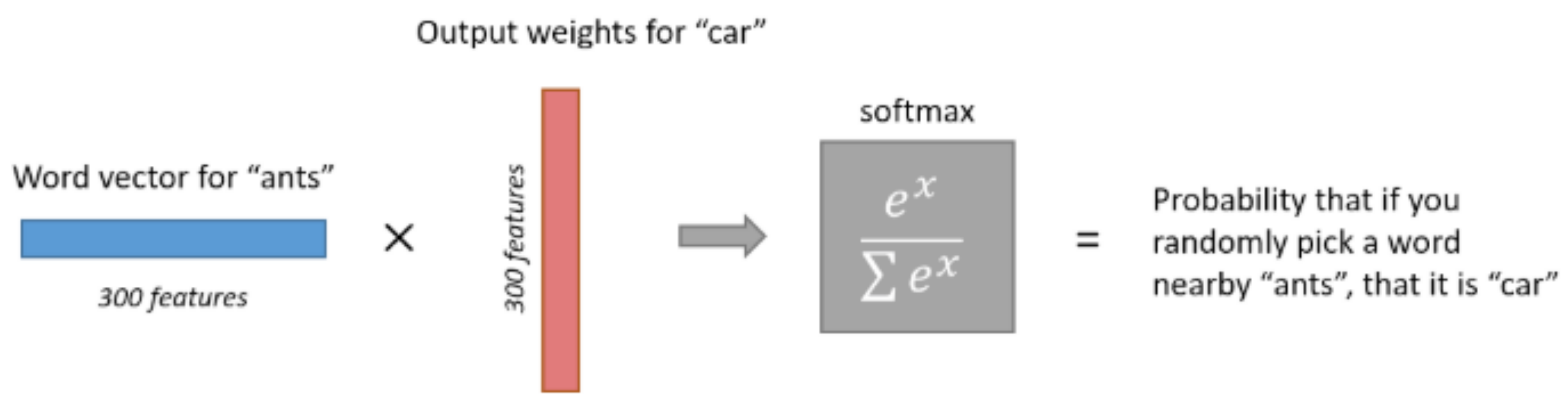


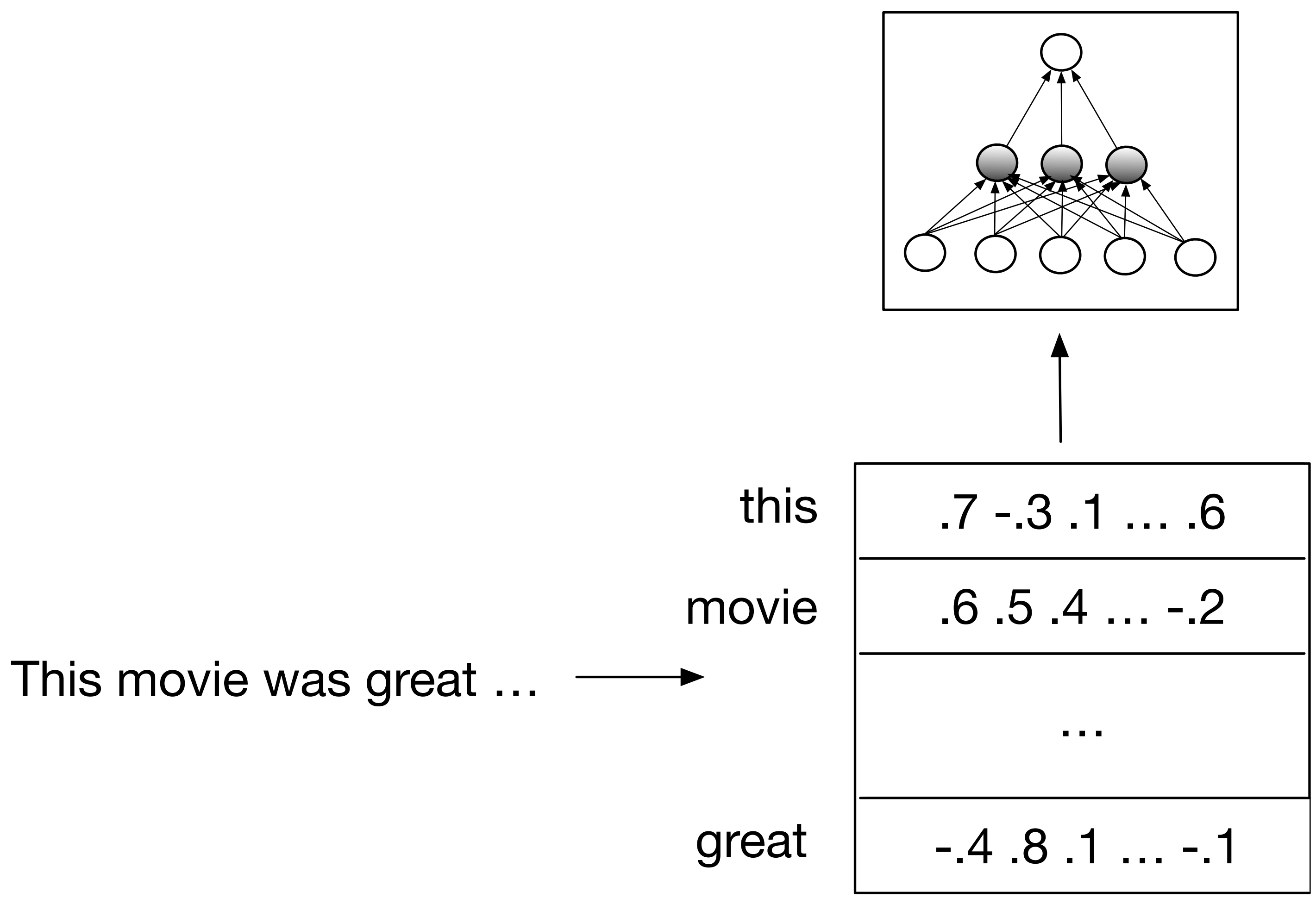
Image credit: Chris McCormick
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Formally

$$\arg\max_{\theta} \sum_{(w,c)\in D} \log p(c|w) = \sum_{(w,c)\in D} (\log e^{v_c \cdot v_w} - \log \sum_{c'} e^{v_{c'} \cdot v_w})$$

Transfer

The advantage of word embeddings is that we can learn them then *transfer* to new target tasks



Practical things

- You can download (static) word embeddings that have been 'pre-trained' — you will often load these as initializations

- Gensim is a nice module for working with these things (https://radimrehurek.com/gensim/models/word2vec.html)

A note of caution

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

A note of caution

Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

A note of caution

We will come back to such issues

Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

	Extreme *she* occupations	
1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

	Extreme *he* occupations	
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Summary

- *Self-supervision* refers, roughly, to unsupervised learning strategies that formulate an objective involving predicting parts of the data itself

Summary

- *Self-supervision* refers, roughly, to unsupervised learning strategies that formulate an objective involving predicting parts of the data itself

- Auto-encoders provide one (popular) family of such methods

Summary

- *Self-supervision* refers, roughly, to unsupervised learning strategies that formulate an objective involving predicting parts of the data itself

- Auto-encoders provide one (popular) family of such methods

- Designing self-supervision strategies is an active area of research in vision, NLP, and other areas