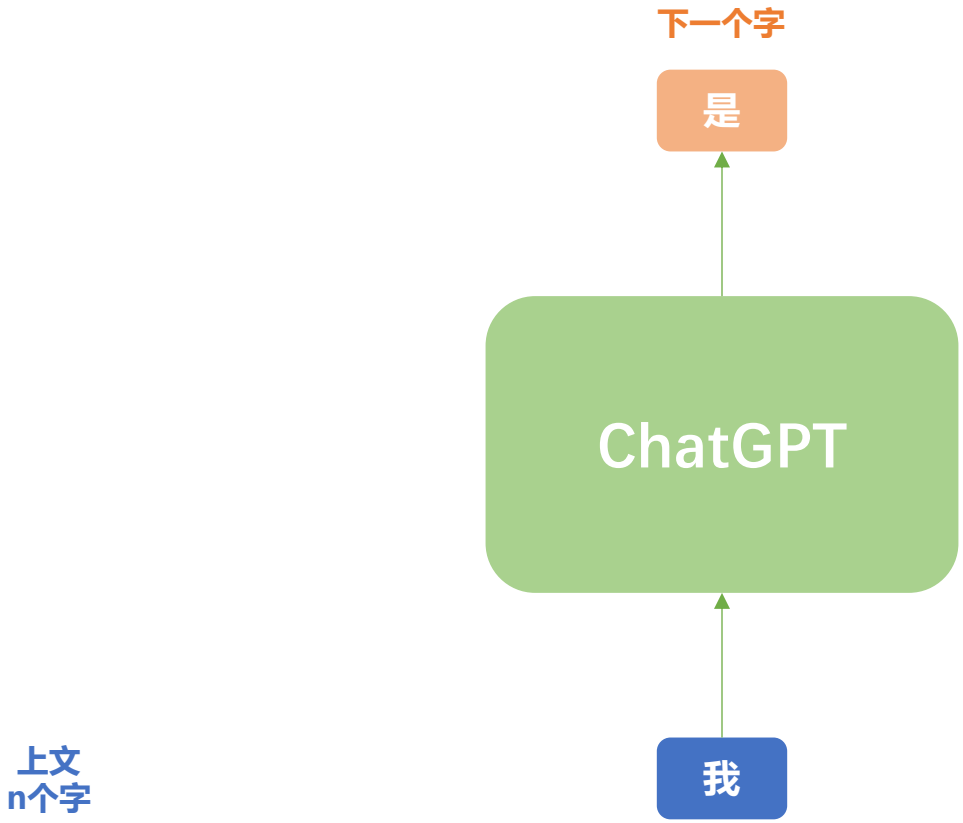
Word Embedding
是什么?

史轩宇

2024.02.24

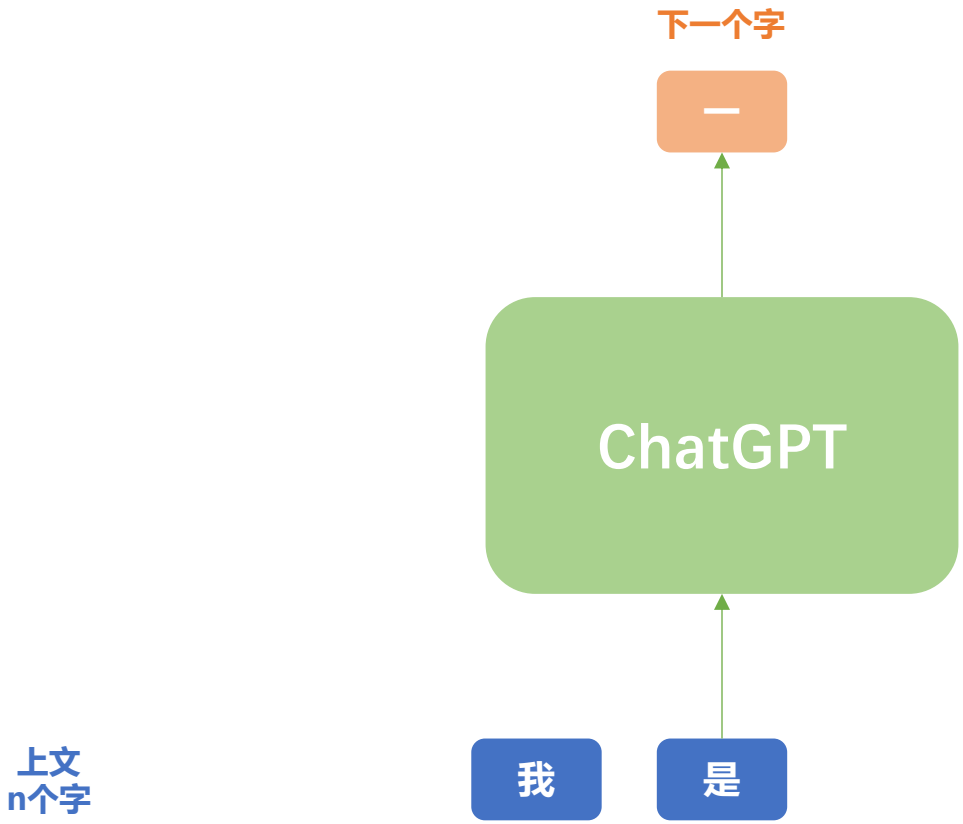
vicshi94

ChatGPT 的基础原理：单字接龙



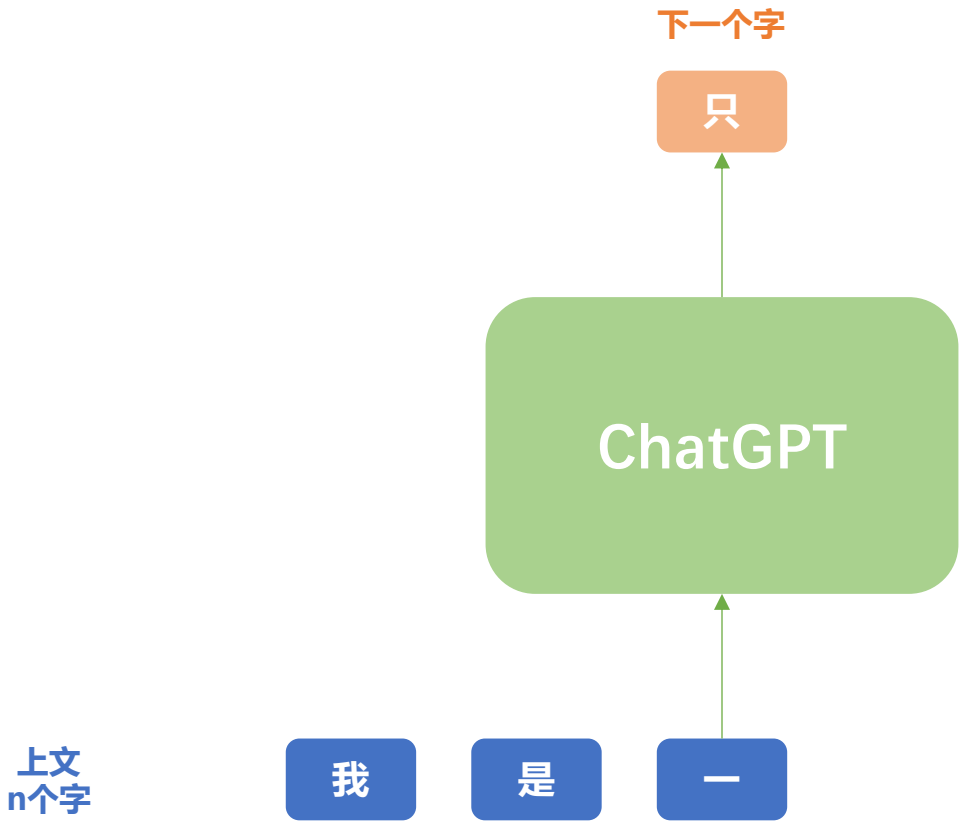
此部分内容来源于 于建国老师视频《【渐构】万字科普GPT4为何会颠覆现有工作流；为何你要关注微软Copilot、文心一言等大模型》
https://www.bilibili.com/video/BV1MY4y1R7EN/

ChatGPT 的基础原理：单字接龙



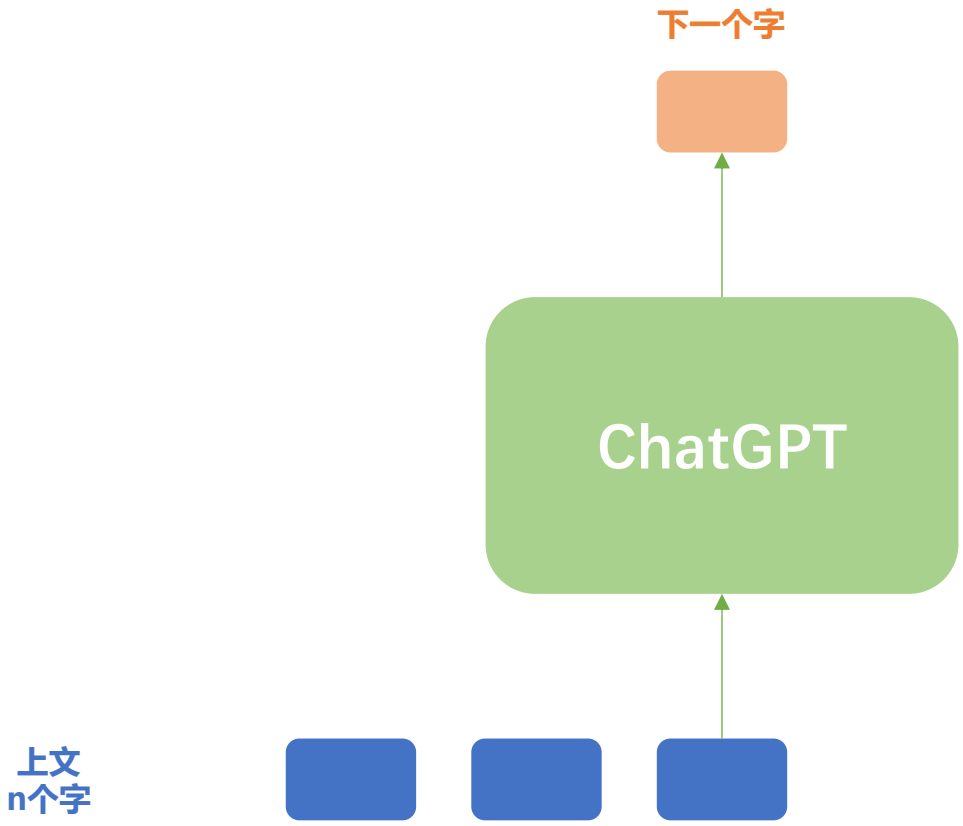
此部分内容来源于 于建国老师视频《【渐构】万字科普GPT4为何会颠覆现有工作流；为何你要关注微软Copilot、文心一言等大模型》
https://www.bilibili.com/video/BV1MY4y1R7EN/

ChatGPT 的基础原理：单字接龙



此部分内容来源于 于建国老师视频《【渐构】万字科普GPT4为何会颠覆现有工作流；为何你要关注微软Copilot、文心一言等大模型》
https://www.bilibili.com/video/BV1MY4y1R7EN/

ChatGPT 的基础原理：单字接龙



“不同的上文，同一个模型”，

产生对应的回答

此部分内容来源于 于建国老师视频《【渐构】万字科普GPT4为何会颠覆现有工作流；为何你要关注微软Copilot、文心一言等大模型》
https://www.bilibili.com/video/BV1MY4y1R7EN/

Text as data?

现在，我们有一个简单模型还有一句话：

“Stitch is cute”

One-hot Encoding

word	encoding
Stitch	1
is	2
cute	3

An ordinal relationship in numbers!

我们的模型会误以为 “cute” 大于 “is”

	Stitch	is	cute
Stitch	1	0	0
is	0	1	0
cute	0	0	1

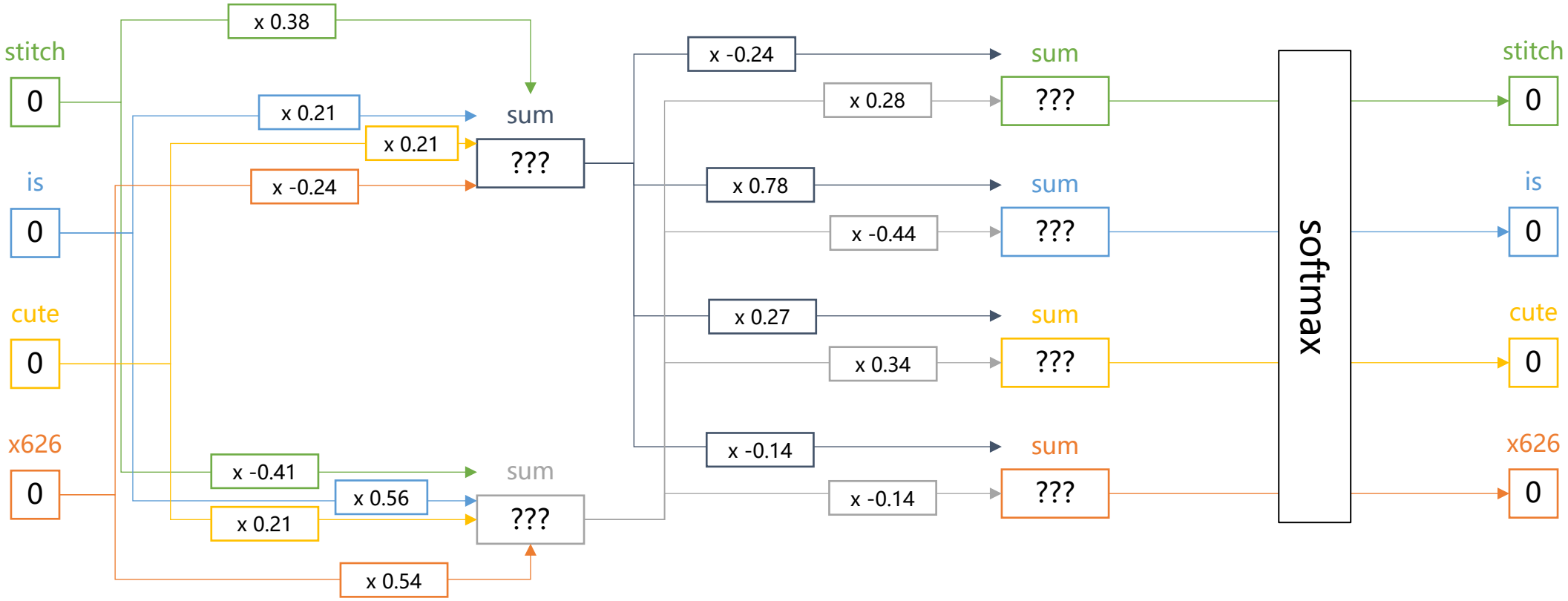
word	one-hot encoding
Stitch	(1,0,0)
is	(0,1,0)
cute	(0,0,1)

这样就没有大小关系了，而且每个单词都被映射为向量空间中距离原点 （0,0,0) 的点，且距离原点的欧式距离都是单位距离1。

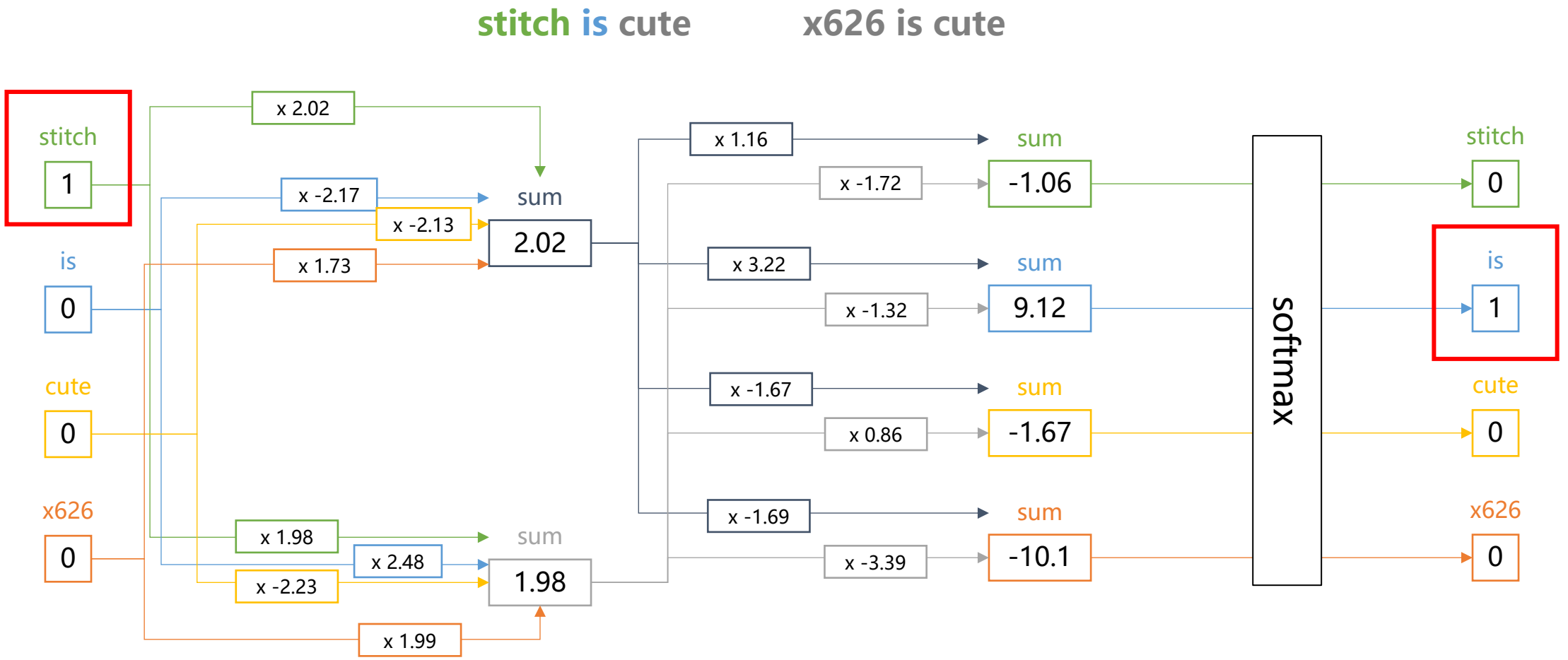
完美！开整！

开始训练超级简单神经网络模型！

我们的training data是两句话：“stitch is cute”，“x626 is cute”

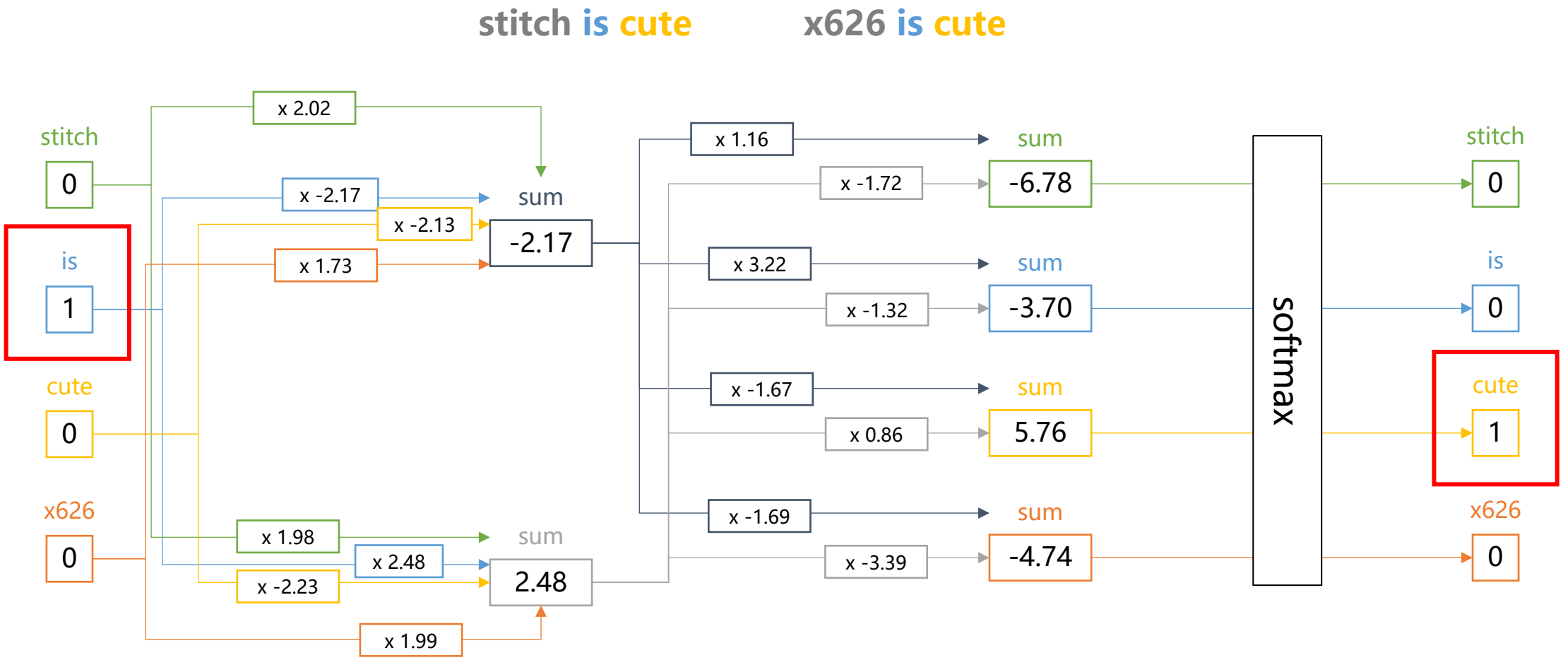


超简单模经过超多轮训练后



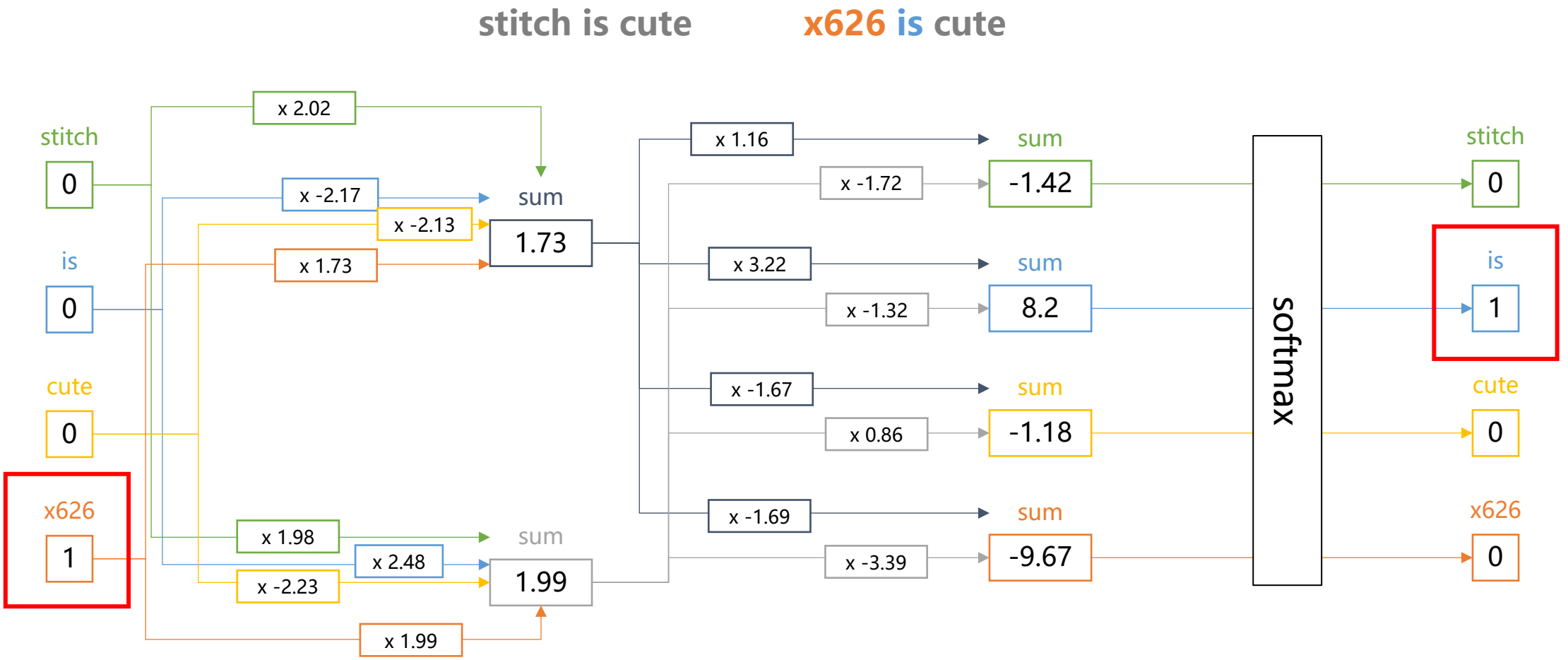
此部分参考了StatQuest with Josh Starmer 的视频《Word Embedding and Word2Vec, Clearly Explained!!!》
https://www.youtube.com/watch?v=viZrOnJclY0

超简单模经过超多轮训练后



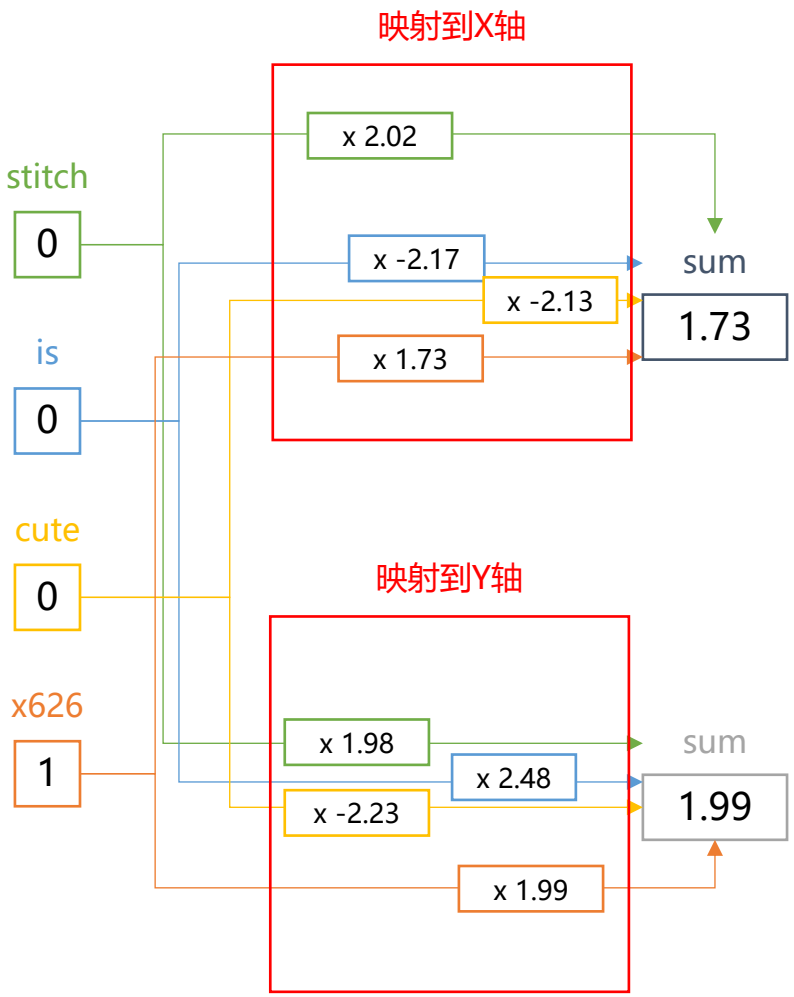
此部分参考了StatQuest with Josh Starmer 的视频《Word Embedding and Word2Vec, Clearly Explained!!!》
https://www.youtube.com/watch?v=viZrOnJclY0

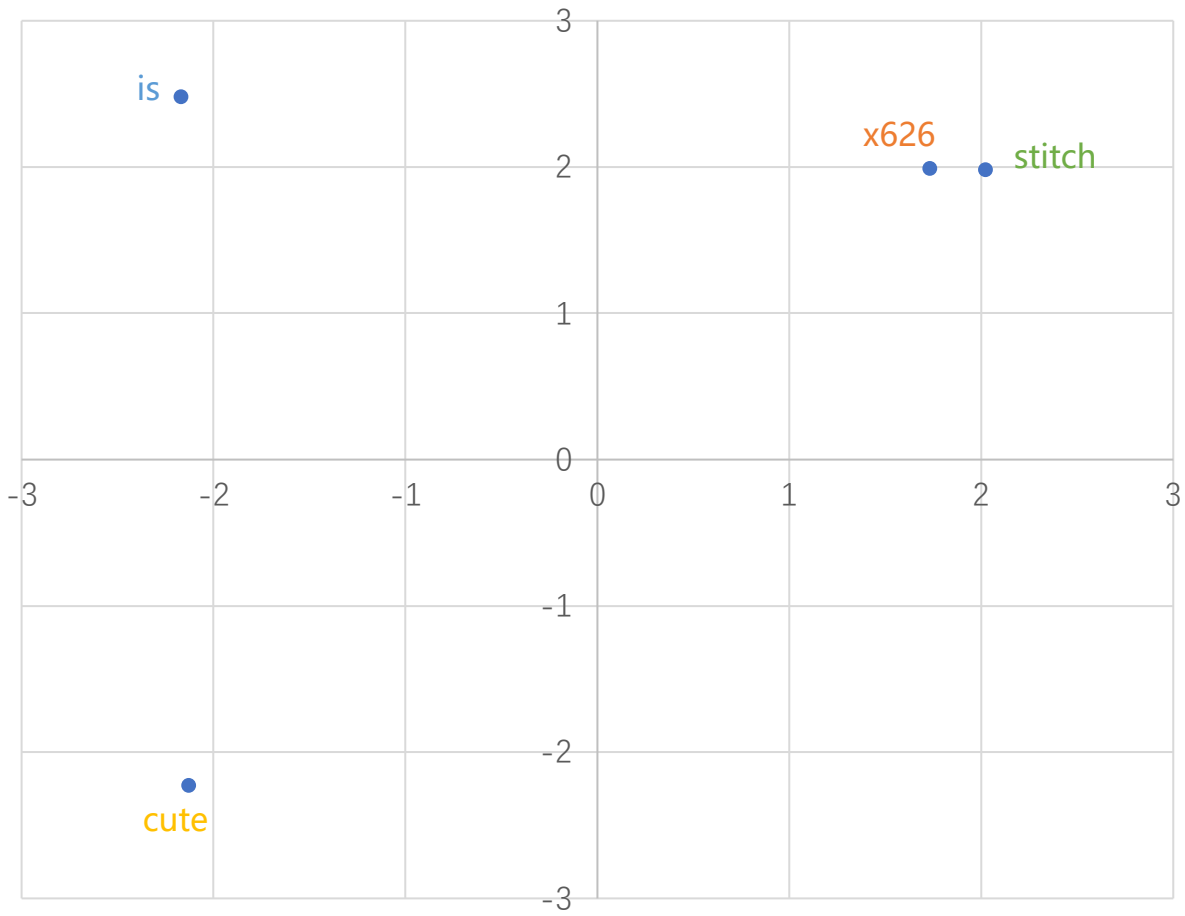
超简单模经过超多轮训练后



此部分参考了StatQuest with Josh Starmer 的视频《Word Embedding and Word2Vec, Clearly Explained!!!》
https://www.youtube.com/watch?v=viZrOnJclY0

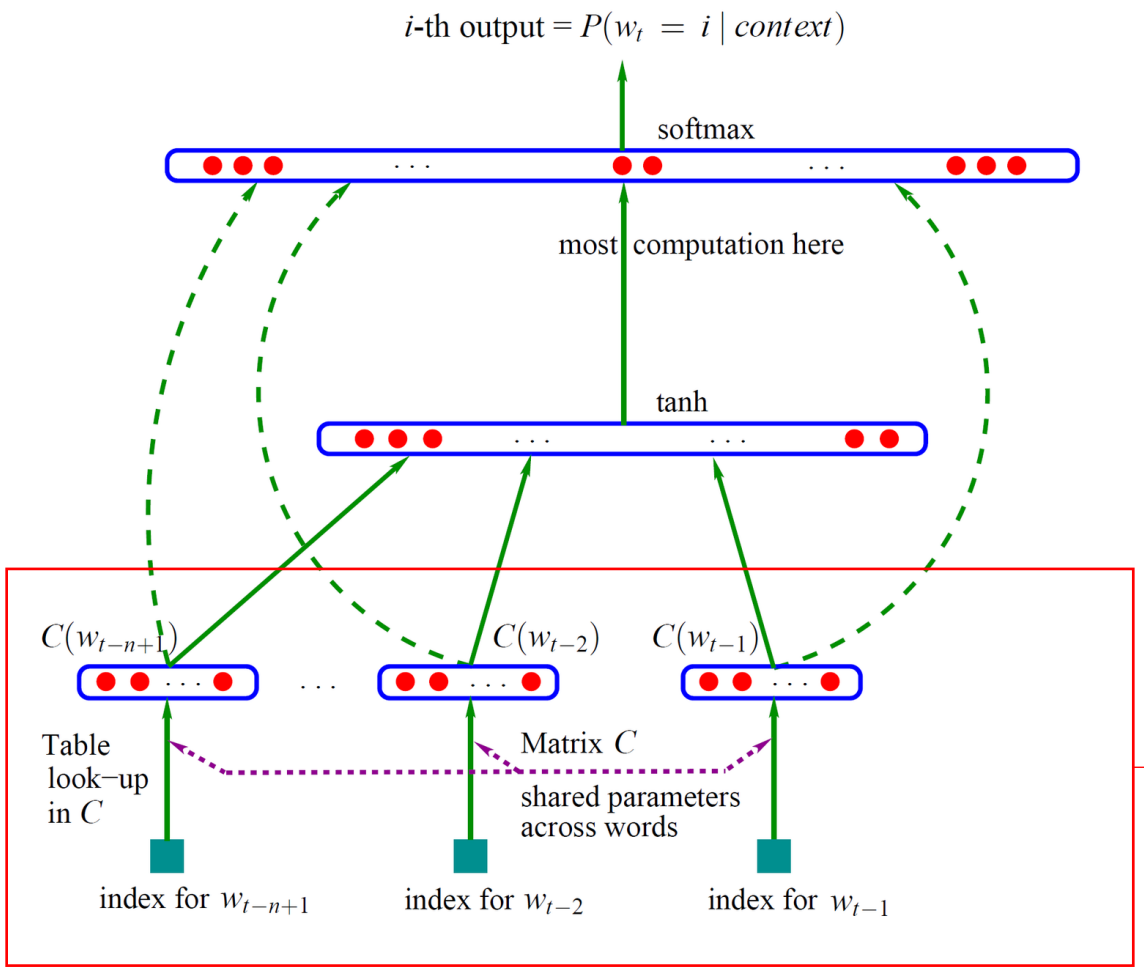
超简单模经过超多轮训练后





此部分参考了StatQuest with Josh Starmer 的视频《Word Embedding and Word2Vec, Clearly Explained!!!》
https://www.youtube.com/watch?v=viZrOnJclY0

Vectorization



$$w \times C = c$$

$$\underbrace{\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix}}_{\text{One-hot vector}} \times \underbrace{\begin{bmatrix} 8 & 2 & 1 & 9 \\ 6 & 5 & 4 & 0 \\ 7 & 1 & 6 & 2 \\ 1 & 3 & 5 & 8 \\ 0 & 4 & 9 & 1 \end{bmatrix}}_{\text{Embedding Weight Matrix}} = \underbrace{\begin{bmatrix} 1 & 3 & 5 & 8 \end{bmatrix}}_{\text{Hidden layer output}}$$

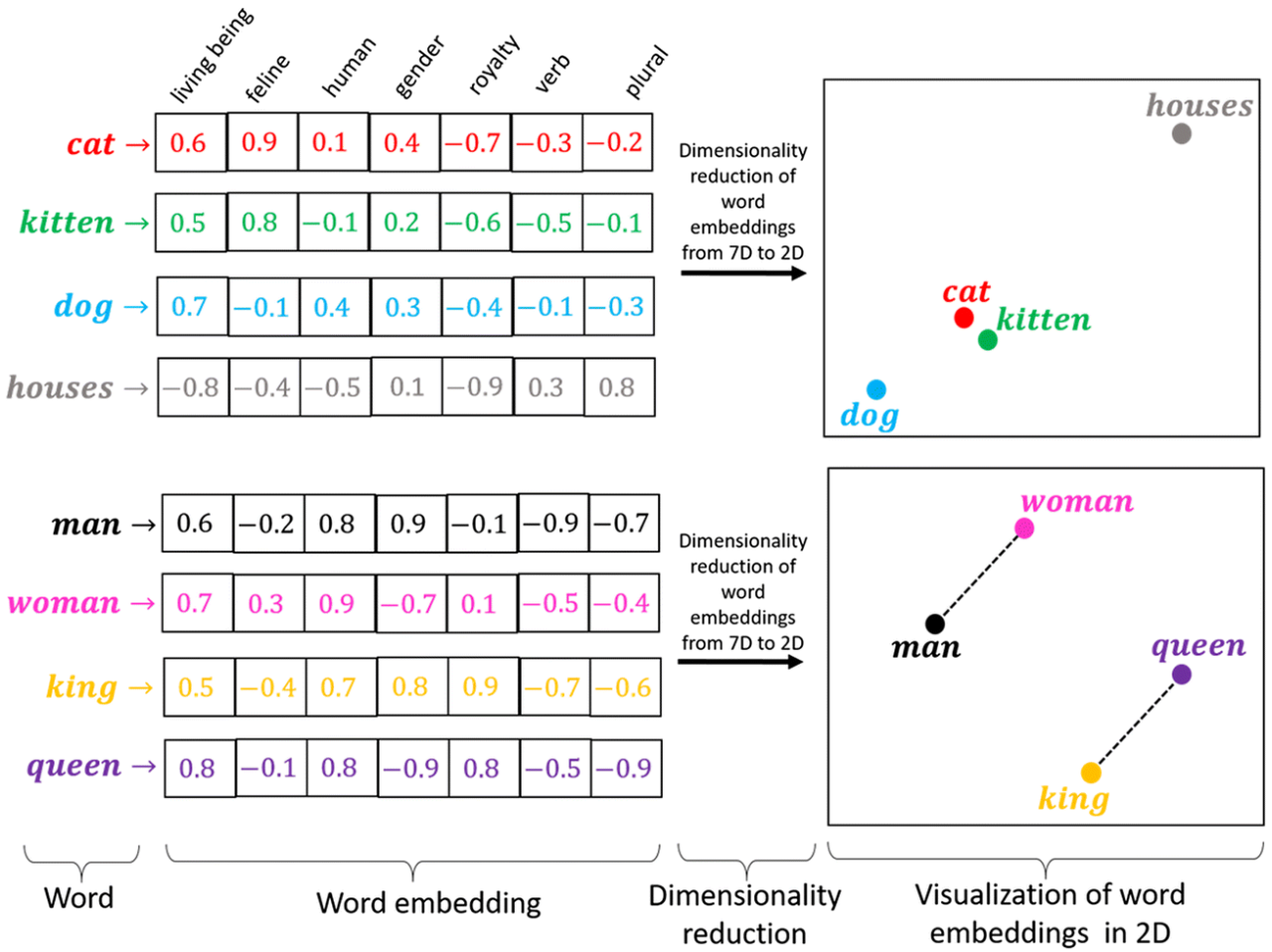
独热编码主要缺点：

1. 维度太多
2. 太稀疏（n^2-n个0）

通过word embedding转换为词向量后明显改善了这两点，且方便了下游任务

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*(6), 1137–1155. https://doi.org/10.1162/153244303322533223

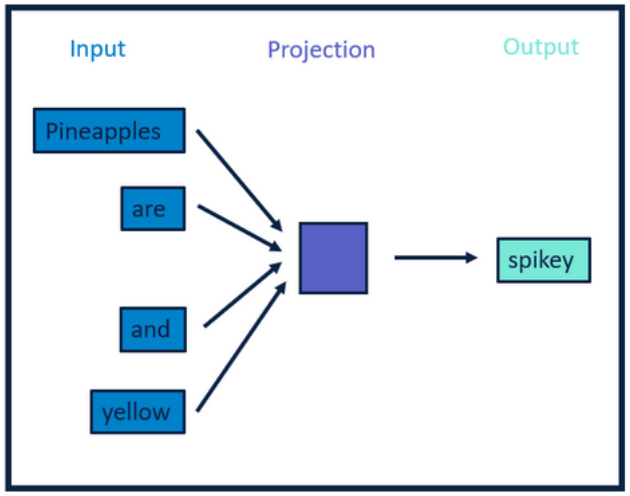
Word Embedding Application



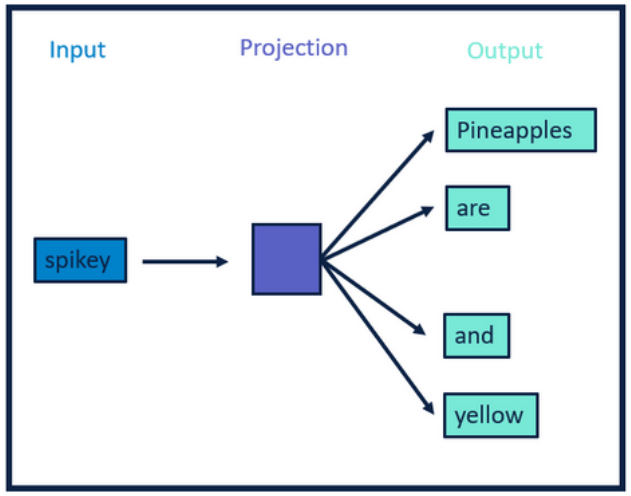
and TM、Supervised ML……

为了强化对于矩阵C的训练

Word2Vec



CBOW



Skip-gram

GloVe

简单来说，更强调全局的共现频率，而非仅捕捉相邻的几个词

图片来源：https://community.alteryx.com/t5/Data-Science/Word2vec-for-the-Alteryx-Community/ba-p/305285

Tutorial: text2vec in R