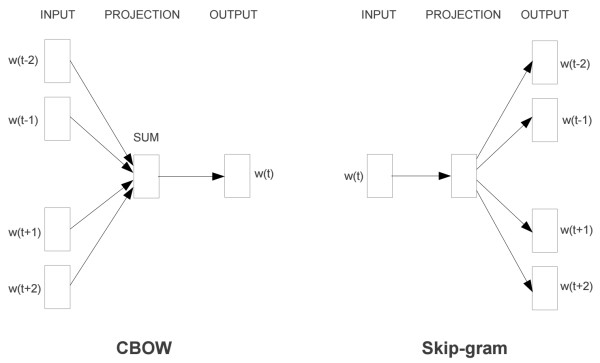
词嵌入------Word2Vector

定义：通过学习文本来用词向量的方式表达词的语义信息，即通过一个嵌入空间使得语义上相近的单词距离很近，反之很远。

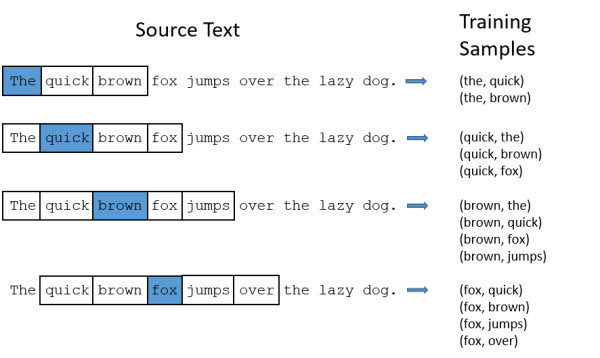
实现方式主要有Skip-gram和 CBOW两种模型，skip-gram是给定input word，预测上下文，CBOW是给定上下文，预测input word。



Word2vec模型分为两个部分，1.建立模型，2.通过模型获取词向量。Word2vec基于训练数据构建一个神经网路模型，模型训练的目的是为了获取模型的参数，而不是用于处理新的任务，该训练过程称为fake task。神经网络的权重参数实际上是我们将要得到的word vector。

例如句子：the dog barked at the mailman.

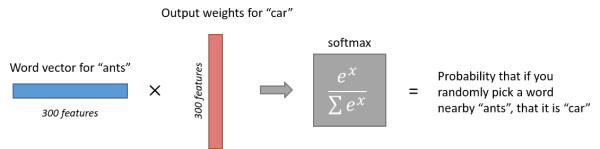
Skip-gram模型下，以dog为input word ，设置skip\_window = 2,即会选取input word左侧2个词和右侧2个词，那最终获得窗口中的词有[‘the’, ‘dog’, ‘barked’, ‘at’], 当num\_skips=2，随机得到两组训练数据（‘dog’, ‘barked’） 和（‘dog’, ‘the’）



输入单词和输出单词均以one-hot编码，feed至神经网络进行训练。最终模型的输出的一个关于词典的概率分布。

如果我们想要使用300个特征来表示一个单词，即每个单词可以被表示为一个300维的向量，那么隐藏层应为300列,N行（N为词典的长度），输出层是一个softmax回归分类器，

它的每个节点都将会输出一个【0-1】的概率值，这些输出概率值之和为1. 下面是训练样本呢为（input word: ants, output word: car）的计算示意图：



提高模型训练效率：

1. 将常见的单词组合或者词组作为单个词处理：

例如United States， United Kingdom，New York， 单个的词表达不出含义，它们应该组合在一起表达词的含义，而不是拆开

1. 对高频单词进行抽样来减少训练样本的个数：

例如 the 这样的常用词出现概率很大，会出现很多（the,..）这样的训练样本，而这些样本远远的超过我们需要学习the的样本总数。

1. 对优化目标采取负采样，每个样本的训练将会只更新神经网络的一部分权重，从而减少计算量：

训练一个神经网络意味着要输入样本并且不断调整权重值，从而不断提高对目标的准确预测。每当神经网络经过一个训练样本，它的权重就会更新一次。假设我们想要将一个词汇量为10000的样本转换为维度为300的向量，那将需要300\*10000的权重矩阵，再算上bias，假设样本以亿计算，每次更新计算量是非常大的