1 Word Embedding

1. 传统的one-hot编码没办法表示两个词之间的距离。

1.1 如何利用contenxt信息来获取word embedding

- **Count based**: 如果两个词总是共现，那么两个词的向量应该接近。例如Glove Vector.
- **predition based** 例如给一词，预测下一个词。例如下图中“蔡英文宣誓就职”和“马英九宣誓就职”。对于模型而言，正确答案就是“宣誓就职”，在给定不同的前一个词的情况下，为了让模型得到相同的答案，其实“马英九”和“蔡英文”在word space空间中应该比较靠近。

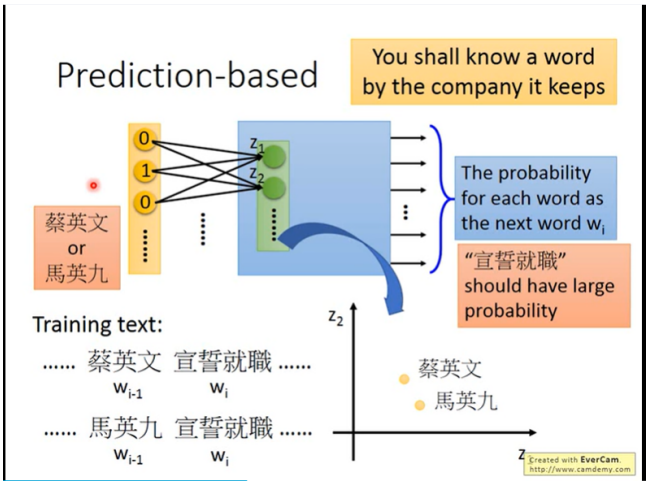
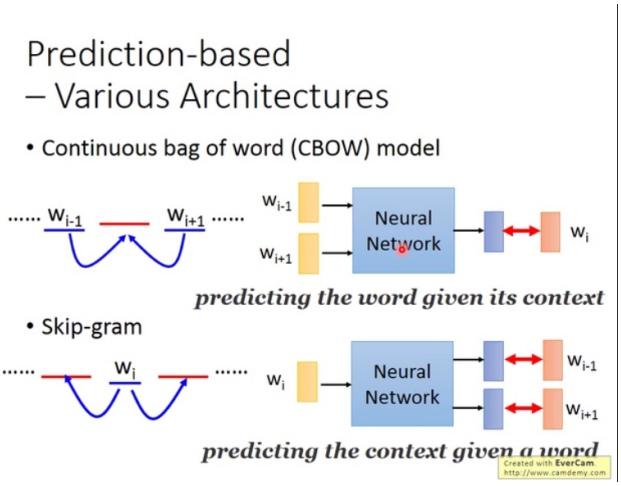


图 1: prediction based的方法示意图

现在用的方法 Continuous bag of word(CBOW) model、Skip-gram。词袋模型是知道前后context单词去预测中间单词。skip gram是知道中间单词去预测前后单词。注意图中的Neural Network中不是DNN，原文中只是一个最简单的linear。



H]

图 2: 词袋模型和skip gram模型示意图

1.2 一些有意思的事情

- 左图是首都和国家之间的关系，可以发现这个关系在不同国家和首都之间存在一些共性，有点TransE的味道。右图是单词的不同时态之间的关系。

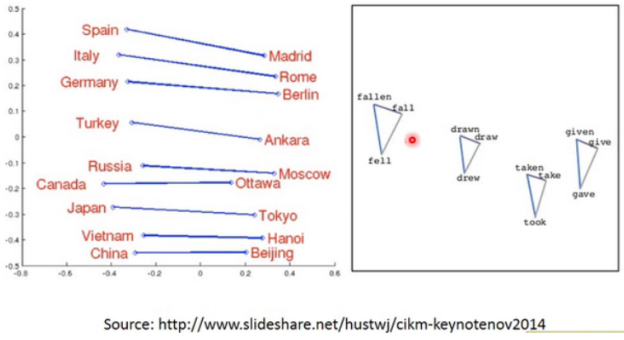


图 3: 词和词之间的关系

-

在做图像分类时，对于机器没有见过的类别，模型往往不能得到正确的分类。例如在训练的时候只给了狗，汽车，马的图片，那么模型将无法将图片分类成猫。但是如果将图片和文字映射到同一个space下面，并且让对应的

图片的embedding分布在对应的词语附近，那么只要知道了cat的位置，模型就有很大可能会把猫的图片放在cat这个词的附近，就能知道新图片的分类。

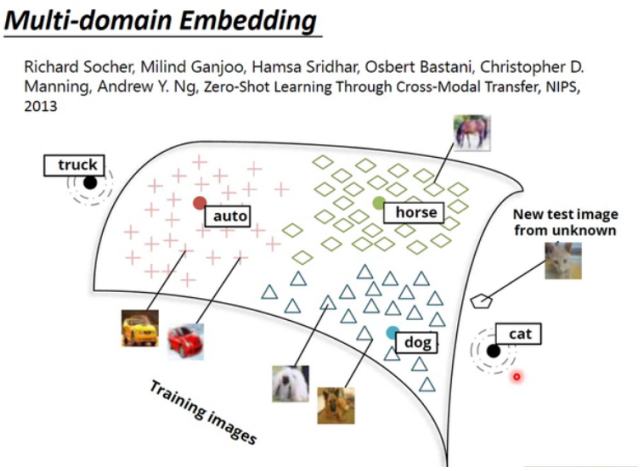


图 4: 多模态，有点zero shot的意思

2 事件抽取

2.1 什么是时间抽取？

事件抽取是从描述事件信息的文本中，识别并抽取出事件信息，并以结构化的形式呈现出来，包括发生的时间、地点、参与角色以及与之相关的动作或者状态的改变。

- **事件描述（Event Mention）**：描述事件的词组/句子/句群，包含一个 trigger 以及任意数量的 arguments.
- **事件触发（Event Trigger）**：事件描述中最能代表事件发生的词汇，决定事件类别的重要特征，一般是动词或者名词
- **事件元素（Event Argument）**：事件的重要信息，或者说是实体描述（entity mention），主要由实体、属性值等表达完整语义的细粒度单位组成

- **元素角色（Argument Role）**：事件元素在事件中扮演的角色，事件元素与事件的语义关系，可以理解为 slot
- 事件类型（Event Type）

事件抽取基础任务是在 mention 中抽取一个 trigger 和多个 arguments，并找到每个 argument 对应的 role，以及 trigger 的 type。因此基础方法可以分成四步：

1. Trigger Identification
2. Trigger Type Classification
3. Argument Identification
4. Argument Role Classification

2.2 常用的方法

2.2.1 基于模式匹配

2.2.2 基于传统机器学习

2.2.3 基于深度学习