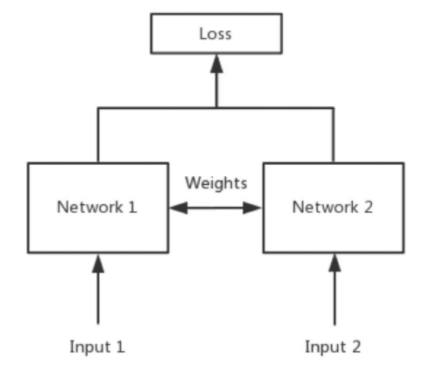
# Text Summarization

## Siamese Network & Optimization Tricks

* 1. **Siamese Network**
  2. Siamese Network

孪生神经网络模型有两个相同的网络结构并且共享参数，对这两个网络输入不同的数据，两个神经网络分别将输入映射到新的空间，形成输入在新的空间中的表示。通过Loss的计算，评价两个输入的相似度。



模型的目标是让两个相似的输⼊距离尽可能的⼩，两个不同类别的输⼊距离尽可能的⼤。Loss公式为：

其中是标签，0表示相似，1表示不相似，是一个正的常量，，表示两个输入的距离，并且使用了欧式距离，是因为更适合句⼦级别、段落级别的⽂本相似性度，保存两个向量的⻓度信息，而余弦距离更适合词汇级别的相似度，仅计算两个向量的夹⻆。

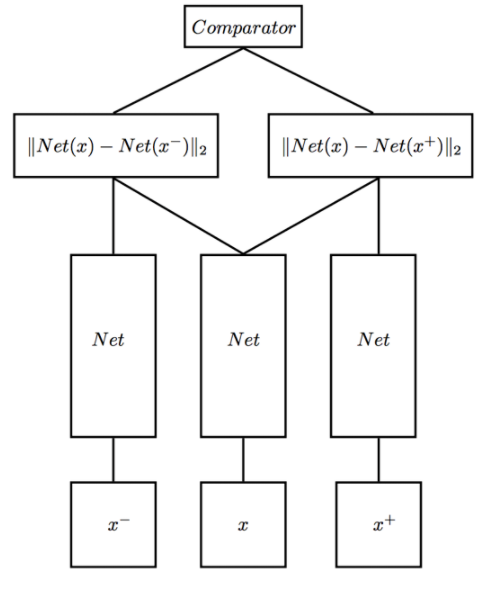
思想是当两个输入相似时，令变小，当不相似时，令大于阈值。

在代码实现的时候可以是同一个网络，不用实现另外一个，因为权值都一样。

**pseudo-siamese network**和siamese network相比区别在于两个网络结构不共享参数。因此siamese network用于处理两个输入"比较类似"的情况。pseudo-siamese network适用于处理两个输入"有一定差别"的情况。比如，我们要计算两个句子或者词汇的语义相似度，使用siamese network比较适合；如果验证标题与正文的描述是否一致（标题和正文长度差别很大），或者文字是否描述了一幅图片（一个是图片，一个是文字），就应该使用pseudo-siamese network。

* 1. Triple Siamese

Triplet network有三个输入，一个是anchor（标准数据），一个positive（anchor的正样本），一个negtive（anchor的负样本）；或者两个正样本一个负样本、一个正样本两个负样本。



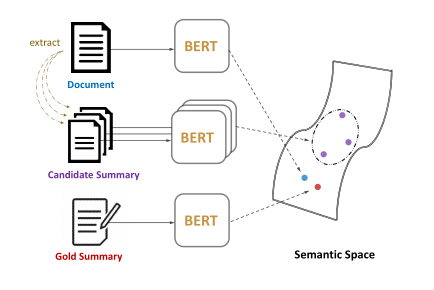
可以看到它是先计算了anchor和正负样本之间的距离，再进行比较。其loss公式如下：

其中分别表示三种输入，表示经过网络得出的向量。Loss的思想是让与的距离更近，让与的距离更远，公式上来看就是让的距离比大至少。

1. **MatchSum** 
   1. 贡献

由于之前的抽取式摘要模型都是基于句子级（Sentence-level）提取的，对所有句子逐个打分，最后取topn的句子为摘要，而MATCHSUM考虑了句子间的关系，是利用BertSum先抽取m个句子组成候选集，再从中选出n个句子组成摘要级组合（Summary-level），利用摘要级组合整体与标准摘要进行计算得出Summary-level Score。即基于候选句间的组合句与原文档的相似度来判断文档摘要的模型。

* 1. 模型结构



其模型结构与Triple Siamese相似，有原文、候选摘要、标准摘要三个输入，

1. **Don't Stop Pretraining**

参考：

[Siamese network 孪生神经网络--一个简单神奇的结构](https://zhuanlan.zhihu.com/p/35040994)

[MATCHSUM论文笔记](https://zhuanlan.zhihu.com/p/269132744)