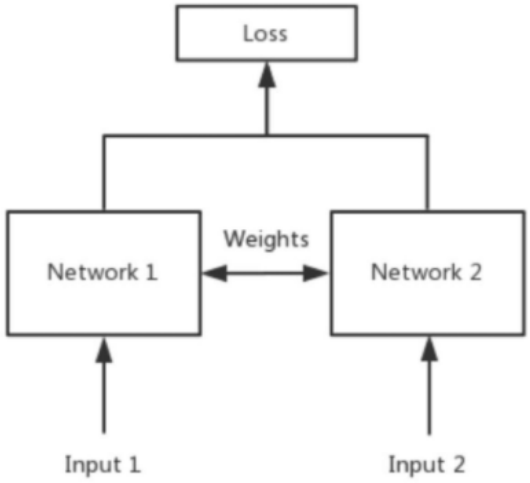
Text Summarization

Siamese Network & Optimization Tricks

1.1. Siamese Network

1.1. 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：

$$(1-Y)\frac{1}{2}(D_W)^2+(Y)\frac{1}{2}\{\max(0,m-D_W)\}^2$$

其中Y是标签，0 表示相似，1 表示不相似，m是一个正的常量，$D_W=\sqrt{\{G_w(X_1)-G_w(X_2)\}^2}$，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，保存两个向量的长度信息，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角。

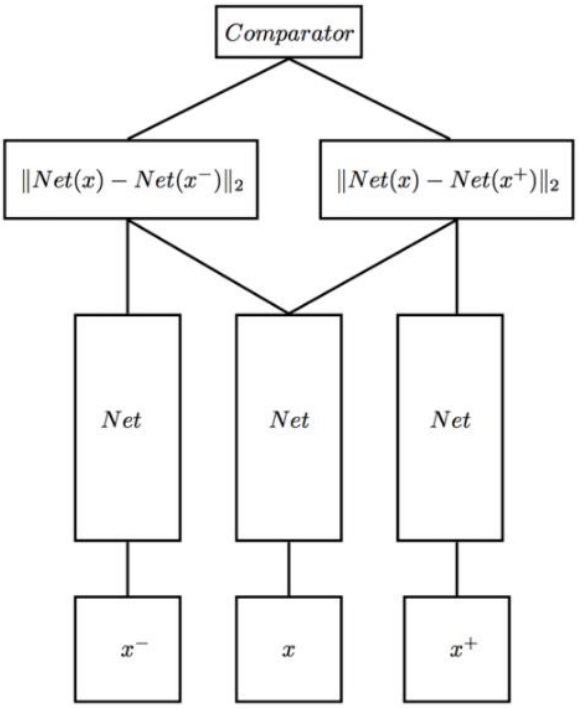
思想是当两个输入相似时，令D_W变小，当不相似时，令D_W大于阈值m。

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。

1.2. 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：

$$L(A,P,N) = \max\left(\left||f(A) - f(P)\right|\right|^2 - \left||f(A) - f(N)\right|\right|^2 + a, 0\Big)$$

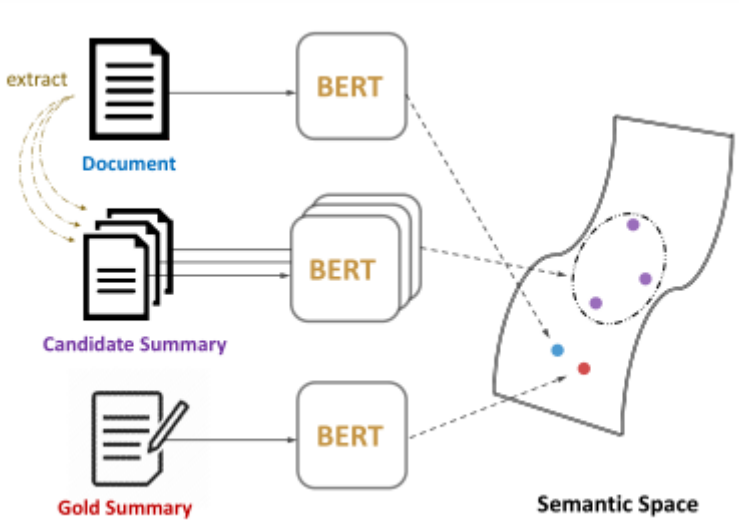
其中A, P, N分别表示三种输入，$f(*)$表示经过网络得出的向量。Loss 的思想是让A与P的距离更近，让A与N的距离更远，公式上来看就是让AN的距离比AP大至少a。

2. MatchSum

2.1. 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。

2.2. 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，其中候选摘要是原文通过 bertsum 选出来得分较高的 m 个句子。通过从候选摘要里选出 n 个句子组成不同的摘要组合，与原文进行相似度计算，选出最佳摘要组合。MatchSum 的 loss 由两部分组成：

第一部分是基于候选摘要与原文档的相似度，其目标函数为：

$$L_1 = \max\left(0, f(D, C) - f(D, C^*) + \gamma_1\right)$$

同样是想让候选组合C与原文D的距离与标准摘要C^*和原文D的距离差小于边界值γ_1。

第二部分考虑候选摘要之间的差异性，即基于 margin loss 的思想，认为得分靠前的与得分靠后的有较大的差异，其损失函数可表示为：

$$L_2 = \max\left(0, f\left(D, C_j\right) - f(D, C_i) + (j - i) * \gamma_2\right), (i < j)$$

这里是对C_k进行了排序，当$f\left(D, C_j\right) > f(D, C_i)$，$i, j \in k$时，令$i < j$。这里的思想是，假设$f(D, C_i) = d_1$，那么排在后面的$f\left(D, C_j\right)$则至少大于$d_1 + (j - i) * \gamma_2$.，以此来拉开$C_i$和$C_j$的距离。

2.3. 总结

在抽取式摘要中，考虑到了多个侯选摘要的组合情况，同时也验证了基于侯选摘要句独立假设来逐句选择摘要句的不合理性。首次选用了语义匹配来选择 Best-Summary，在想法上具有较大的创新性。

但是这种策略似乎在 inference 上极慢，候选摘要句的组合得不少，然后也会基于 bert 来做相似度，时间再次消耗，最终得到单篇摘要猜想至少是秒级。

3. Don't Stop Pretraining

为了使用利用专业领域内的大量未标注数据，可以将官方预训练好的模型继续进行预训练。

3.1. Domain Adaptive Pretraining(DAPT)

DAPT 是利用领域内的语料对模型进行继续的预训练，在当前领域与官方预训练语料重合较少的情况下，DAPT 的提升较大，但是对于领域外的下游任务效果大部分相较于原始预

训练模型还下降了，因此，在大多数情况下，不考虑领域相关性而直接直接暴露于更多数据的持续预训练对最终任务可能是有害的。

3.2. Task Adaptive Pretraining(TAPT)

而 TAPT 选用了任务相关的无标注数据集继续进行预训练（一小部分，大部分还是用于 task 的训练），任务数据集可以看作相关领域数据的一个子集。相比于 DAPT，TAPT 使用的是预训练语料要少得多，但是与特定任务相关的语料要多得多。

TAPT 效果要比 DAPT 差一些。

参考：

Siamese network 孪生神经网络--一个简单神奇的结构
MATCHSUM 论文笔记