Unit Dot Product Similarity

- 你是否曾经质疑余弦相似度只考虑两个向量之间夹角的合理性？
- 点积可以被用作计算相似度吗？如果用点积作为两个向量之间相似度的判断标准，会存在什么问题？

在这篇论文（https://link.springer.com/article/10.1007/s10791-025-09516-2）中，我们讲述了余弦相似度和点积相似度的缺陷，并提出了一种新颖的相似度——单元点积相似度（Unit Dot Product Similarity，UDPS）。这种新的相似度可以完美解决余弦相似度和点积相似度的问题，同时拥有良好的数学性质。

由于毕业要求等原因，这篇论文以语音识别为主，没有对单元点积相似度进行单独的讲解，此外，这篇论文的第一作作者和通讯作者被导师们占有。因此，在这里我对单元点积相似度进行一个重新的讲述，如有错误欢迎联系（2447424163@qq.com）讨论。

余弦相似度和点积相似度

首先，我们回顾一下余弦相似度和点积相似度：

$$S_{\text{cosine}}(\mathbf{a}, \mathbf{b}) = \cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$$

$$S_{\text{dot}}(\mathbf{a}, \mathbf{b}) = \mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\|\cos(\theta)$$

其中，$\mathbf{a}$、$\mathbf{b}$为两个不全为零向量的N维向量，$\mathbf{a} \in \mathbb{R}^N$，$\mathbf{b} \in \mathbb{R}^N$，$\theta$ 为两个N维向量 $\mathbf{a}$、$\mathbf{b}$ 之间的夹角，$\cos$ 为角的余弦计算函数，$\|\cdot\|$ 为向量的范数，也即向量的模长，通常使用向量的2范数。S_{cosine} 为余弦相似度，S_{dot}为点积相似度。

缺陷

余弦相似度只考虑两个向量之间的夹角，这就意味着，无论两个向量是否等长，只要他们之间的方向相同，他们之间的相似度就为最大值1。例如，向量$(1, 2)$与$(1, 2)$之间的相似度为1，向量$(1, 2)$与$(2, 4)$之间的相似度也为1。尽管$(1, 2)$与$(2, 4)$不同，但它们之间的相似度仍为最大值1。

点积相似度的取值范围是无界的，并且在夹角不变的情况下，向量模长越大，点积值就越大。向量$(1, 2)$与$(1, 2)$之间的相似度为5，向量$(2, 4)$与$(2, 4)$之间的相似度为20。这意味着，如果用点积作为相似度，$(2, 4)$与$(2, 4)$之间的相似度大于$(1, 2)$与$(1, 2)$之间的相似度，然而，$(1, 2)$与$(1, 2)$完全相同，$(2, 4)$与$(2, 4)$也完全相同，从常理的角度来说，$(2, 4)$与$(2, 4)$之间的相似度并不大于$(1, 2)$与$(1, 2)$之间的相似度。

造成点积相似度上述缺陷的原因是什么？

点积相似度没有归一化，因此当两个向量被同比放大或缩小时，得到的结果也会被放大或缩小。因此，是否可以尝试通过归一化的手段来解决点积相似度的上述缺陷呢？

解决方法

事实上，余弦相似度就是一种归一化的点积相似度，它相当于先对每个向量进行单位化，然后再做点积，即：

$$S_{\text{cosine}}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}}{\|\mathbf{a}\|} \cdot \frac{\mathbf{b}}{\|\mathbf{b}\|}$$

但是这样做的问题在于，它忽略了两个向量模长比例的关系，只要两个向量方向一样，两个不同的向量之间的相似度，和两个完全一样的向量之间的相似度相同，都为最大值1。从朴素的直观感受上来说，两个不同的向量之间的相似度应该是小于两个完全相同的向量之间的相似度的。例如，$(1, 2)$与$(2, 4)$之间的相似度，在直观感受上小于$(1, 2)$与$(1, 2)$之间的相似度。

如果在归一化的同时，保持两个向量之间模长比例的关系，是否就能解决点积相似度的问题，同时避免余弦相似度的缺陷呢?

一个简单的归一化的同时保持向量模长比例的方法是，将两个向量的模长的和归一化到1，也就是：

$$S(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}}{\|\mathbf{a}\|+\|\mathbf{b}\|} \cdot \frac{\mathbf{b}}{\|\mathbf{a}\|+\|\mathbf{b}\|}$$

性质

向量 $\mathbf{a}$ 被归一化为 $\frac{\mathbf{a}}{\|\mathbf{a}\|+\|\mathbf{b}\|}$，向量 $\mathbf{b}$ 被归一化为 $\frac{\mathbf{b}}{\|\mathbf{a}\|+\|\mathbf{b}\|}$ 。被归一化后两个向量的模长的和为1，即：

$$\left\| \frac{\mathbf{a}}{\|\mathbf{a}\|+\|\mathbf{b}\|} \right\| + \left\| \frac{\mathbf{b}}{\|\mathbf{a}\|+\|\mathbf{b}\|} \right\| = 1$$

当 $\mathbf{a}$ 和 $\mathbf{b}$ 被同比放大或缩小时，被归一化后的两个向量的模长保持不变，即：

$$\frac{k\mathbf{a}}{\|k\mathbf{a}\|+\|k\mathbf{b}\|} = \frac{\mathbf{a}}{\|\mathbf{a}\|+\|\mathbf{b}\|}$$

被归一化后的两个向量的模长比例与归一化前 $\mathbf{a}$ 和 $\mathbf{b}$ 的模长比例相同，因为归一化时 $\mathbf{a}$ 与 $\mathbf{b}$ 被以同样的比例（$\frac{1}{\|\mathbf{a}\|+\|\mathbf{b}\|}$）缩放。

$S(\mathbf{a},\mathbf{b})$是否可以解决余弦相似度和点积相似度的缺陷，是否可以作为一种更好的相似度计算方法?

记被归一化后的 $\mathbf{a}$ 和 $\mathbf{b}$ 分别为 $\mathbf{a}^{'}$ 和 $\mathbf{b}^{'}$，即：

$$\mathbf{a}^{'} = \frac{\mathbf{a}}{\|\mathbf{a}\|+\|\mathbf{b}\|}$$

$$\mathbf{b}^{'} = \frac{\mathbf{b}}{\|\mathbf{a}\|+\|\mathbf{b}\|}$$

由于$\|\mathbf{a}^{'}\| + \|\mathbf{b}^{'}\| = 1$，因此：

$$S(\mathbf{a},\mathbf{b}) = \mathbf{a}^{'} \cdot \mathbf{b}^{'} = \|\mathbf{a}^{'}\|\|\mathbf{b}^{'}\|\cos(\theta) = \|\mathbf{a}^{'}\|(1-\|\mathbf{a}^{'}\|)\cos(\theta)$$

其中 $\mathbf{a}$、$\mathbf{b}$ 不全为零向量，θ 为 $\mathbf{a}$ 和 $\mathbf{b}$ 之间的夹角，也是 $\mathbf{a}^{'}$ 和 $\mathbf{b}^{'}$ 之间的夹角。由于 $0 \le \|\mathbf{a}^{'}\| \le 1$，$0 \le \theta \le \pi$，可以证明 $-\frac{1}{4} \le S(\mathbf{a},\mathbf{b}) \le \frac{1}{4}$，当且仅当 $\mathbf{a}$ 与 $\mathbf{b}$ 完全相同时 $S(\mathbf{a},\mathbf{b}) = \frac{1}{4}$，当且仅当 $\mathbf{a}$ 与 $\mathbf{b}$ 模长相同但方向相反时 $S(\mathbf{a},\mathbf{b}) = -\frac{1}{4}$。具体的证明参考论文。

由于

$$\|\mathbf{a}^{'}\| = \begin{cases} \frac{1}{1+\frac{\|\mathbf{b}\|}{\|\mathbf{a}\|}}, & \|\mathbf{a}\| > 0 \\ 0, & \|\mathbf{a}\| = 0 \end{cases}$$

因此 $\mathbf{a}^{'}$ 的模长只取决于 $\mathbf{a}$ 和 $\mathbf{b}$ 的模长的比值。当 $\mathbf{a}$ 与 $\mathbf{b}$ 的模长相同时，$\|\mathbf{a}^{'}\| = \frac{1}{2}$。当 $\mathbf{a}$ 的模长小于 $\mathbf{b}$ 的模长时，$0 \le \|\mathbf{a}^{'}\| < \frac{1}{2}$ 且 $\frac{\|\mathbf{a}\|}{\|\mathbf{b}\|} < 1$。当 $\mathbf{a}$ 的模长大于 $\mathbf{b}$ 的模长时，$\frac{1}{2} < \|\mathbf{a}^{'}\| \le 1$ 且 $\frac{\|\mathbf{a}\|}{\|\mathbf{b}\|} > 1$。

在 $\mathbf{a}$ 和 $\mathbf{b}$ 的夹角不变的情况下，同比放大或缩小 $\mathbf{a}$ 和 $\mathbf{b}$ 不会影响它们的模长的比值，因此$\|\mathbf{a}^{'}\|$不受影响，由于 $\cos(\theta)$ 不变，因此 $S(\mathbf{a},\mathbf{b})$ 也不会改变。但是如果调整 $\mathbf{a}$ 和 $\mathbf{b}$ 的模长的比值，$\|\mathbf{a}^{'}\|$ 就会跟着改变，尽管夹角没有改变，$S(\mathbf{a},\mathbf{b})$ 通常还是会跟着改变。因此，在夹角不变的情况下，$S(\mathbf{a},\mathbf{b})$ 既不会像点积相似度那样，因为向量被同比放大或缩小而导致结果也被放大或缩小；也不会像余弦相似度那样，只要夹角不变，相似度无论何种情况下都一样。

此外，在 $\mathbf{a}$ 和 $\mathbf{b}$ 的夹角不变的情况下，由于 $\|\mathbf{a}^{'}\|(1-\|\mathbf{a}^{'}\|)$ 先单调增后单调减当且仅当 $\|\mathbf{a}^{'}\|=\|\mathbf{b}^{'}\|=\frac{1}{2}$ 时取最大值，因此 $\frac{\|\mathbf{a}\|}{\|\mathbf{b}\|}$ 越接近1，$S(\mathbf{a},\mathbf{b})$ 越大，这也符合直观的向量相似度的感受——两个向量的模长越接近，两个向量的相似度就越高。

当两个向量的模长不变时，夹角越大，$\cos(\theta)$ 越小，由于 $\|\mathbf{a}^{'}\|(1-\|\mathbf{a}^{'}\|)$ 不变，因此 $S(\mathbf{a},\mathbf{b})$ 也越小。这一特点与点积相似度和余弦相似度保持一致。这也同样符合直观的向量相似度的感受——两个向量的夹角越大，相似度越低。

所以，$S(\mathbf{a},\mathbf{b})$ 既保留了点积和余弦相似度夹角越大相似度越小的良好特点，又避免了点积和余弦相似度的缺陷。

为了让相似度的取值范围与经典的相似度一样为$[-1,1]$，单元点积相似度被定义为：

最终定义

$$S_{\text{udot}}(\mathbf{a},\mathbf{b})=\begin{cases}4\cdot S(\mathbf{a},\mathbf{b})=\frac{4\cdot\mathbf{a}\cdot\mathbf{b}}{(\|\mathbf{a}\|+\|\mathbf{b}\|)^2}, & \|\mathbf{a}\|+\|\mathbf{b}\|>0\\ 0, & \|\mathbf{a}\|=\|\mathbf{b}\|=0\end{cases}$$

$S_{\text{udot}}(\mathbf{a},\mathbf{b})$的取值范围为$[-1,1]$，性质与$S(\mathbf{a},\mathbf{b})$类似，这里不再赘述。

关于单元点积相似度与self-attention

self-attention中的点积可以替换为单元点积嘛？效果怎么样？

对于每一个注意力头，经典的self-attention是这样计算的：

$$\mathbf{A}=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

其中 $\mathbf{A}\in\mathbb{R}^{N\times N}$，$Q\in\mathbb{R}^{N\times k}$，$K\in\mathbb{R}^{N\times k}$。$\mathbf{A}_{i,j}=\frac{q_ik_j^T}{\sqrt{d_k}}$，$q_i$、$k_j$ 分别为 Q、K 的第 i 列和第 j 列。

将q_i与k_j的缩放点积替换为单元点积后，$\frac{q_ik_j^T}{\sqrt{d_k}}$变为 $4\cdot\frac{q_ik_j^T}{(\|q_i\|+\|k_j\|)^2}$ ，$\mathbf{A}$相应地变为：

$$\mathbf{A}=\text{softmax}\left(4\cdot\frac{QK^T}{D}\right)$$

其中 $D\in\mathbb{R}^{N\times N}$，

$$D_{i,j}=(\|q_i\|+\|k_j\|)^2$$

然而，$[-1,1]$的取值范围对于 softmax 来说太小，为了增加区分度，可以在原来的基础上乘上一个可学习的参数，即：

$$\mathbf{A}_{udps}=\text{softmax}\left(4\cdot\alpha\cdot\frac{QK^T}{D}\right)$$

其中 $\alpha\in\mathbb{R}^1$，α 是一个学习的参数，初值可以设为10左右。这便是使用单元点积相似度的self-attention的最原始的形式。在实际应用中，α 也可以改为 α^2 以保证缩放系数大于0，多个注意力头可以共享一个 α ，也可以每个注意力头有一个独立的 α 。

此外，由于 D 的计算可以用并行的方式进行，因此实际的计算时间只增加10%左右。

代码

单元点积相似度和使用单元点积相似度的self-attention的代码实现在`main.py`中。相关的语音识别模型代码（改进的Transformer和Conformer）在`wenet-1.0.0-att_udps`中，尤其在`wenet-1.0.0-att_udps/wenet/transformer/attention.py`和`wenet-1.0.0-att_udps/wenet/transformer/encoder.py`中。