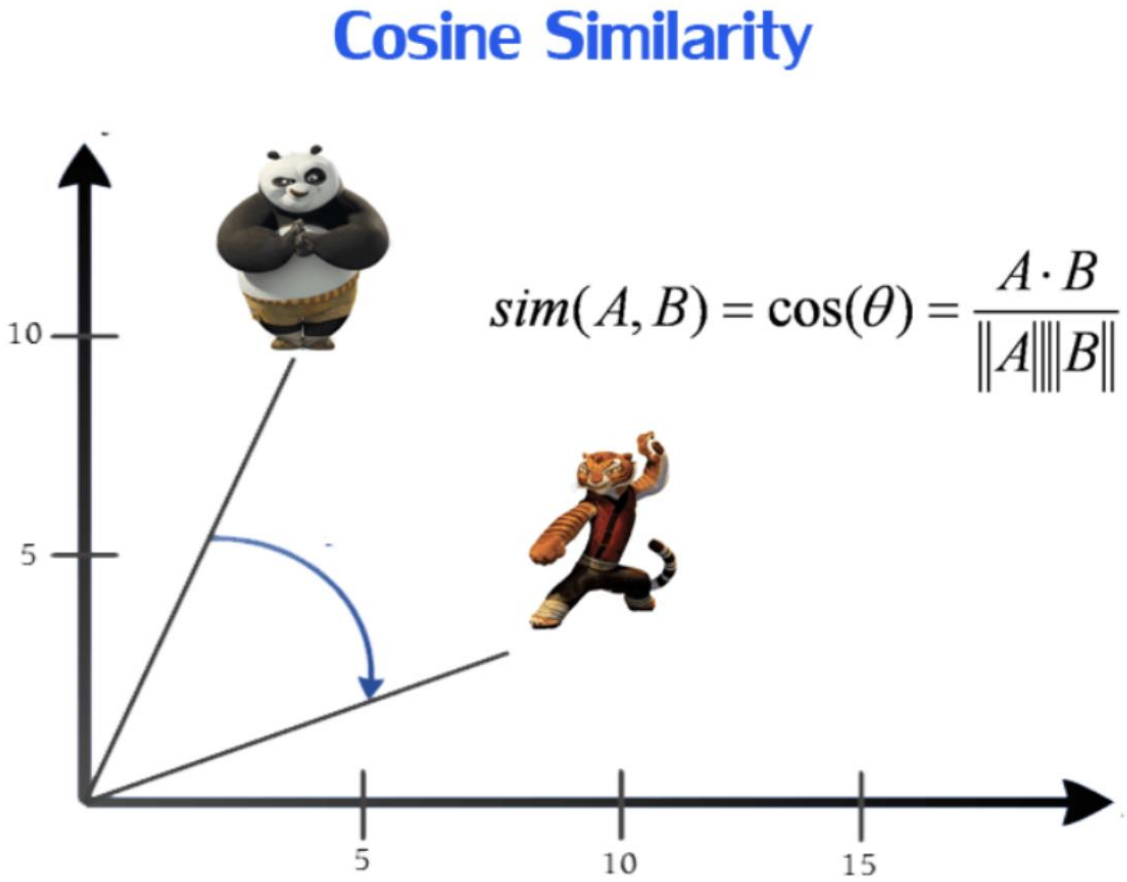
Is Cosine-Similarity of Embeddings Really About Similarity?

N e t f i l x

Is Cosine-Similarity of Embeddings Really About Similarity?

- 余弦相似度可以等价地视作归一化后向量的点积，通常被用于量化高维对象的语义相似度
- 底层逻辑是嵌入向量的范数（长度）不如方向（归一化后方向的一致性）重要
- 某些研究指出某些场景下直接使用未归一化的嵌入向量点积反而表现更好



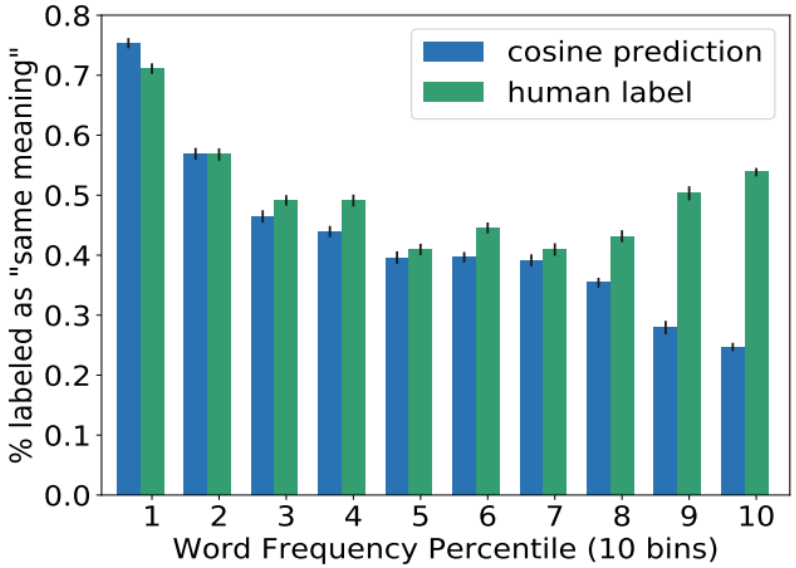


Figure 2: Percentage of examples labeled as having the "same meaning". In high frequency words, cosine similarity-based predictions (blue/left) on average **under**-estimate the similarity of words as compared to human judgements (green/right).

Is Cosine-Similarity of Embeddings Really About Similarity?

北京大学 PEKING UNIVERSITY

- 采取具有闭式解的矩阵分解模型（matrix-factorization model）进行研究和解析

$$X = PQ^T \in R^{\{p \times p\}} \qquad P, Q \in R^{\{p \times k\}} \qquad k \leq p$$

$$X \approx XAB^T$$

$$\left(XAB^{\mathrm{T}}\right)_{u,i} = \langle \overrightarrow{x_u} \cdot A, \overrightarrow{b_i} \rangle$$

- 有余弦相似度

$$cosSim\left\langle \overrightarrow{b_i}, \overrightarrow{b_{i'}} \right\rangle$$

$$cosSim\left\langle \overrightarrow{x_u} \cdot A, \overrightarrow{x_{u'}} \cdot A \right\rangle$$

$$cosSim\left\langle \overrightarrow{x_u} \cdot A, \overrightarrow{b_i} \right\rangle$$

Is Cosine-Similarity of Embeddings Really About Similarity?

北京大学 PEKING UNIVERSITY

- 添加两种不同的L2正则项

$$\min_{A,B} ||X - XAB^\top||_F^2 + \lambda||AB^\top||_F^2 \quad (1)$$

$$\min_{A,B} ||X - XAB^\top||_F^2 + \lambda(||XA||_F^2 + ||B||_F^2) \quad (2)$$

- (1)式为整体正则化
- (2)式近似于 $\min_W ||X - PQ^\top||_F^2 + \lambda(||P||_F^2 + ||Q||_F^2)$，分别对$P$和$Q$进行正则化，类似于权重衰减

Is Cosine-Similarity of Embeddings Really About Similarity?

变换研究

- 对于任意旋转矩阵R，AR和BR也依然都是解，且余弦相似度也不变
- 对于(1)，对A和B的列向量放缩后，余弦相似度也不变
- 当AB^T为第一个正则化方法下的解时，$ADD^{-1}B^T$也为解，其中D是任意对角矩阵。

此时有新解 $\hat{A}^{(D)} := \hat{A}D$

$\hat{B}^{(D)} := \hat{B}D^{-1}.$

归一化结果为 $(X\hat{A}^{(D)})_{(\text{normalized})} = \Omega_A X\hat{A}^{(D)} = \Omega_A X\hat{A}D$ and

$\hat{B}^{(D)}_{(\text{normalized})} = \Omega_B\hat{B}^{(D)} = \Omega_B\hat{B}D^{-1},$

其中Ω_a和Ω_b为对角矩阵，依赖于D，可记作$\Omega_a(D)$和$\Omega_b(D)$

- item-item $\text{cosSim}\left(\hat{B}^{(D)}, \hat{B}^{(D)}\right) = \Omega_B(D)\cdot\hat{B}\cdot D^{-2}\cdot\hat{B}^\top\cdot\Omega_B(D)$
- user-user $\text{cosSim}\left(X\hat{A}^{(D)}, X\hat{A}^{(D)}\right) = \Omega_A(D)\cdot X\hat{A}\cdot D^2\cdot(X\hat{A})^\top\cdot\Omega_A(D)$
- user-item $\text{cosSim}\left(X\hat{A}^{(D)}, \hat{B}^{(D)}\right) = \Omega_A(D)\cdot X\hat{A}\cdot\hat{B}^\top\cdot\Omega_B(D)$

Is Cosine-Similarity of Embeddings Really About Similarity?

- 对于(1)，有闭式解 $\hat{A}_{(1)}\hat{B}_{(1)}^{\top} = V_k \cdot \mathrm{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...)_k \cdot V_k^{\top}$, where $X =: U\Sigma V^{\top}$
- 取 $\hat{A}_{(1)} = \hat{B}_{(1)} := V_k \cdot \mathrm{dMat}(..., \dfrac{1}{1+\lambda/\sigma_i^2}, ...)_k^{\frac{1}{2}}$
- 考虑$k = p$的情况下，对D进行取值

1. 选取 $D = \mathrm{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...)^{\frac{1}{2}}$，有 $\hat{A}_{(1)}^{(D)} = \hat{A}_{(1)} \cdot D = V \cdot \mathrm{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...)$ and $\hat{B}_{(1)}^{(D)} = \hat{B}_{(1)} \cdot D^{-1} = V$

   由于V已归一化，因此 $\Omega_b = I$。此时相似度为 $\mathrm{cosSim}\left(\hat{B}_{(1)}^{(D)}, \hat{B}_{(1)}^{(D)}\right) = VV^{\top} = I,$

$$\begin{aligned}\mathrm{cosSim}\left(X\hat{A}_{(1)}^{(D)}, \hat{B}_{(1)}^{(D)}\right) &= \Omega_A \cdot X \cdot V \cdot \mathrm{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...) \\ &= \Omega_A \cdot X \cdot \hat{A}_{(1)}\hat{B}_{(1)}^{\top},\end{aligned}$$

2. 选取 $D = \mathrm{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...)^{-\frac{1}{2}}$，有

   $\hat{B}_{(1)}^{(D)} = V \cdot \mathrm{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...)$, and $\hat{A}_{(1)}^{(D)} = V$ 和余弦相似度

$$\mathrm{cosSim}\left(X\hat{A}_{(1)}^{(D)}, X\hat{A}_{(1)}^{(D)}\right) = \Omega_A \cdot X \cdot X^{\top} \cdot \Omega_A$$

$$\mathrm{cosSim}\left(X\hat{A}_{(1)}^{(D)}, \hat{B}_{(1)}^{(D)}\right) = \Omega_A \cdot X \cdot \hat{A}_{(1)} \cdot \hat{B}_{(1)}^{\top} \cdot \Omega_B$$

$$\mathrm{cosSim}\left(\hat{B}_{(1)}^{(D)}, \hat{B}_{(1)}^{(D)}\right) = \Omega_B \cdot V \cdot \mathrm{dMat}(..., \frac{1}{1+\lambda/\sigma_i^2}, ...)^2 \cdot V^{\top} \cdot \Omega_B$$

Is Cosine-Similarity of Embeddings Really About Similarity?

北京大学 PEKING UNIVERSITY

- 对于(2) $\min_{A,B} ||X - XAB^\top||_F^2 + \lambda(||XA||_F^2 + ||B||_F^2)$ (2)

  有闭式解 $\hat{A}_{(2)} = V_k \cdot \mathrm{dMat}(..., \sqrt{\frac{1}{\sigma_i} \cdot (1 - \frac{\lambda}{\sigma_i})_+}, ...)_k$ and

  $\hat{B}_{(2)} = V_k \cdot \mathrm{dMat}(..., \sqrt{\sigma_i \cdot (1 - \frac{\lambda}{\sigma_i})_+}, ...)_k$

  此处有 $\hat{P} = X\hat{A}_{(2)} = U_k \cdot \mathrm{dMat}(..., \sqrt{\sigma_i \cdot (1 - \frac{\lambda}{\sigma_i})_+}, ...)_k$ 与 $Q = B$

- 正则项中P与Q分开，只可进行旋转，有唯一余弦相似度
- $\mathrm{dMat}(..., \sqrt{\sigma_i \cdot (1 - \frac{\lambda}{\sigma_i})_+}, ...)_k$ 是否代表实际语义最佳相似度

Is Cosine-Similarity of Embeddings Really About Similarity?

北京大学 PEKING UNIVERSITY

实验设置

- 前面用理论推导了满秩情况下的不可靠性，再设计实验证明低秩的不可靠性
- 设计user-item recommendation实验，生成模拟聚类数据

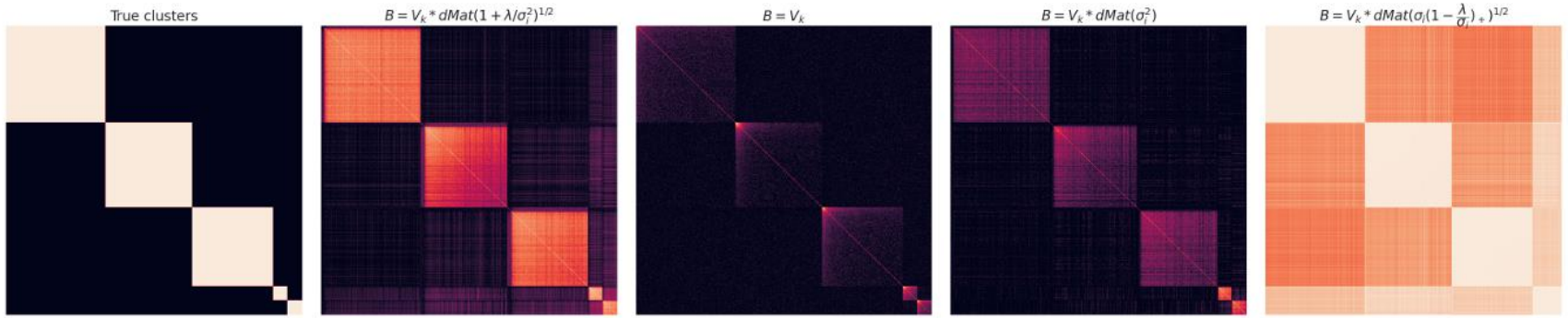


Figure 1: Illustration of the large variability of item-item cosine similarities $cosSim(B, B)$ on the same data due to different modeling choices. Left: ground-truth clusters (items are sorted by cluster assignment, and within each cluster by descending baseline popularity). After training w.r.t. Eq. 1, which allows for arbitrary re-scaling of the singular vectors in V_k, the center three plots show three particular choices of re-scaling, as indicated above each plot. Right: based on (unique) B obtained when training w.r.t. Eq. 2.

Is Cosine-Similarity of Embeddings Really About Similarity?

Discussion

- 深度学习模型通常采用不同正则化技术的组合，这可能会对最终嵌入的余弦相似度产生意想不到的影响。
- 将余弦相似度应用于通过点积优化学习到的嵌入的做法可能会导致不透明且可能毫无意义的结果

Remedies and Alternatives

- 直接以余弦相似度为训练目标，如在模型中引入层归一化或余弦损失，使学到的embedding自带归一化性质
- 避免在嵌入空间中计算余弦相似度，可以将其投影回原始空间中计算。
- 在学习前或学习中做标准化 / 去偏。
  - 标准化
  - 负采样或逆倾向评分（IPS）
  - (余弦相似度与范式相互影响)

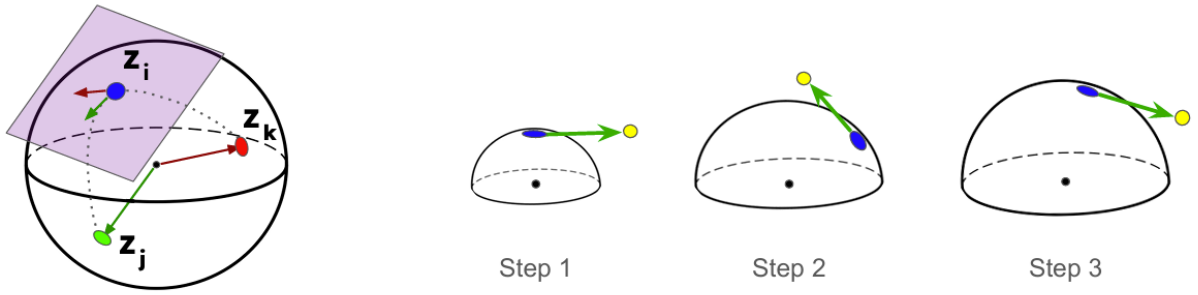
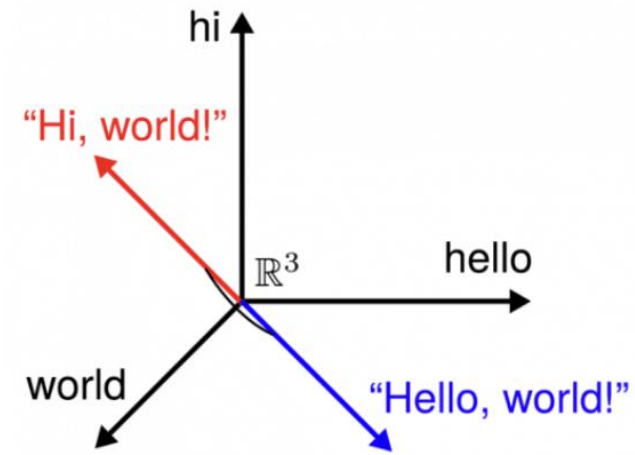


Figure 1: *Left*: The gradients w.r.t. z_i in Proposition 1 and Corollary 4 exclusively exist in the tangent space at z_i. *Right*: The growing embeddings in Corollary 2. Blue points represent z_i at iterations $t = 1, 2, 3$. Yellow points represent z_i', i.e. the result of each step of gradient descent.

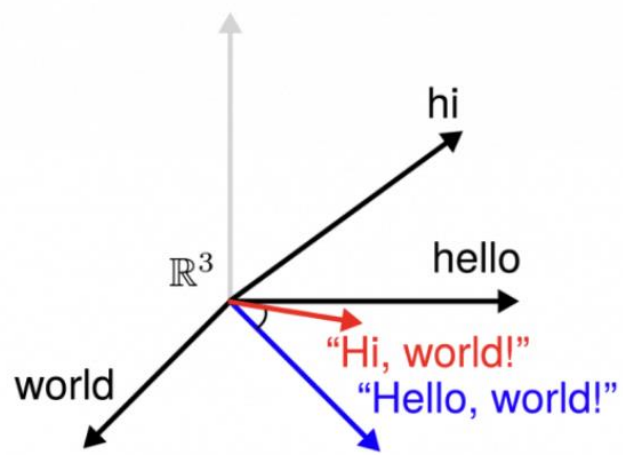
Is Cosine-Similarity of Embeddings Really About Similarity?

Remedies and Alternatives

- 欧氏距离：对向量大小（范数）较为敏感，对嵌入向量进行恰当归一化时，它仍然可以发挥良好效果。
- 点积：在密集段落检索和问答任务中，未经归一化的向量点积有时比余弦相似度更强。
- 软余弦相似度：在计算向量相似度时，除了利用词向量本身的表示，还引入词与词之间的语义相似度，从而实现更细腻的比较。 $\text{soft\_cos}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top S \mathbf{v}}{\sqrt{\mathbf{u}^\top S \mathbf{u}}\sqrt{\mathbf{v}^\top S \mathbf{v}}}$.
- Semantic Textual Similarity：专门针对语义相似度任务微调的模型（如STSScore）
- 归一化嵌入结合余弦相似度（层归一化）



Cosine Similarity



Soft Cosine Measure

Semantics at an Angle: When Cosine Similarity Works Until It Doesn’t

Kisung You

Semantics at an Angle: When Cosine Similarity Works Until It Doesn't

- 余弦相似度的模长不变性让它天然对长度或频率干扰具有鲁棒性
- 词嵌入使语义关系以几何形式体现
- 对比学习和多模态架构中余弦相似度具有训练稳定性、易解释性和评估一致性

- Invariance to Magnitude $\cos\_\text{sim}(\boldsymbol{u}, \boldsymbol{v}) := \frac{\boldsymbol{u}^\top \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|}$
- Semantic Expressivity Through Directionality

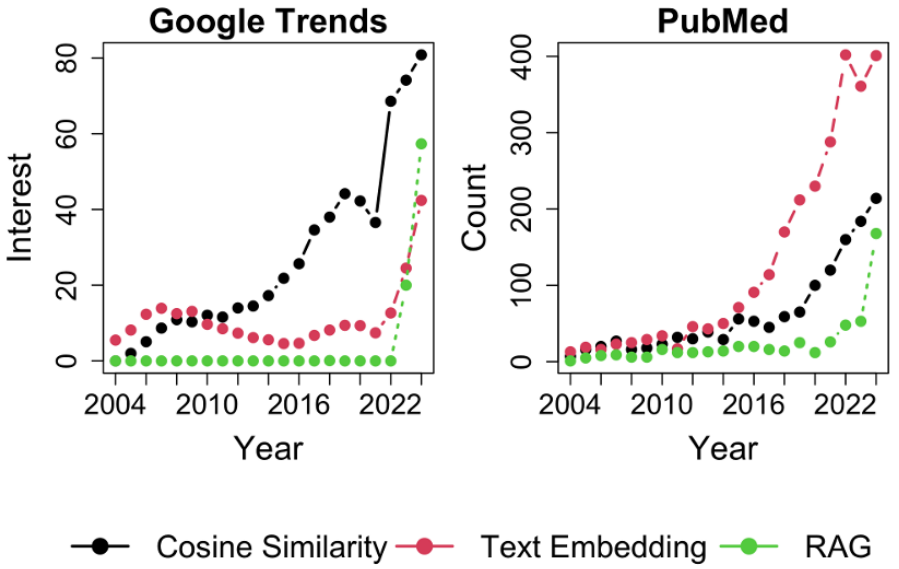
$$\mathcal{L} = -\log \sigma(\boldsymbol{v}_w^\top \boldsymbol{v}_c) - \sum_{i=1}^{k} \log \sigma(-\boldsymbol{v}_{n_i}^\top \boldsymbol{v}_c)$$

- Compatibility with Contrastive Losses

$$\mathcal{L}_{\text{INFONCE}} = -\log \frac{\exp(\text{sim}(\boldsymbol{u}, \boldsymbol{v})/\tau)}{\sum_{i=1}^{N} \exp(\text{sim}(\boldsymbol{u}, \boldsymbol{v}_i)/\tau)}$$

- Mitigating the Curse of Dimensionality

Manifold Interpretation: Cosine Lives on the Sphere



Not All Angles Are Created Equal

Vector norms frequently encode:

- Certainty or alignment strength in multimodal embeddings such as CLIP
- Informativeness in word embeddings
- Prediction confidence or token salience in contextual models

$$\|\mathbf{x}\| \longmapsto \text{certainty, salience, or informativeness,}$$
$$\hat{\mathbf{x}} \longmapsto \text{semantic direction.}$$

- Anisotropy(各向异性) in Embedding Space
- Frequency Bias in Token-Level Embeddings $\text{cos\_sim}(w_{\text{high-freq}}, w') < \text{cos\_sim}(w_{\text{low-freq}}, w')$
- Hubness in High Dimensions $\mathcal{N}(v,k) \gg \frac{1}{|V|}\sum_{u \in V}\mathcal{N}(u,k)$
- Loss of Calibration and Semantic Granularity

Empirical Failure Scenarios:

- Sentence embeddings dominated by high frequency or function words yield misleadingly high similarity scores between unrelated texts.
- Cross-modal embeddings, such as those produced by CLIP, can overvalue syntactically similar prompts, even when their semantic intent diverges.
- Indistinguishable similarity scores for weak and strong hypotheses due to shared directionality.

Semantics at an Angle: When Cosine Similarity Works Until It Doesn't

Norm-Aware Similarity Functions

$$\text{SCALED\_SIM}(\mathbf{x},\mathbf{y}) = \alpha \cdot \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} + (1-\alpha)\cdot(\|\mathbf{x}\|+\|\mathbf{y}\|)$$

$$\text{WRD}(\mathbf{x},\mathbf{y}) = |\|\mathbf{x}\|-\|\mathbf{y}\|| + \lambda \cdot \arccos\left(\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}\right)$$

Post-Hoc Isotropization and Whitening

- Mean-Centering: Subtracting the dataset mean from each embedding to remove the dominant direction.
- Whitening: Applying a linear transformation to normalize the covariance matrix to the identity.
- Principal Component Removal: Dropping the top-k directions with the most variance, which often dominate and distort similarity metrics.

Query-Normalized Adjustments for Retrieval

$$\tilde{s}(\mathbf{q},\mathbf{d}) = \frac{\cos\_\text{sim}(\mathbf{q},\mathbf{d}) - \mu_q}{\sigma_q}$$

Hybrid Measures and Angular–Radial Decoupling

- Radially Weighted Angles: Similarity functionshat modulate cosine scores using norm-based confidence weights.
- Feature Augmentation: Concatenating $\hat{x}$ and $\| x \|$ as distinct features in downstream scoring models.
- Norm-Sensitive Training: Contrastive objectives that incorporate penalties or regularization on vector norms.