CIDEr

See CIDEr.pdf to read the original paper. I'll extract relevant equations and details here.

CIDEr accounts for both the local and global appropriateness of a description by defining a measure that operates on the the term-frequency of an n-gram weighted with its inverse document frequency. The term-frequency of an n-gram ω_k is defined as

$$\tau_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})}$$

where $h_k(s_{ij})$ is the count of ω_k in s_{ij}, Ω is the vocabulary of all n-grams and s_{ij} is reference description j for image I_i. This is to say, the term-frequency is the count of an n-gram in a description divided by the total count of n-grams in the same description. Whereas the term-frequency is large if ω_k is often present in its reference descriptions, the inverse document frequency is low if ω_k is often present in the corpus,

$$\iota_k(s_{ij}) = \log\left(\frac{\|I\|_1}{\min(1, \sum_{I_p \in I} \sum_q h_k(s_{pq}))}\right).$$

The term-frequency inverse-document-frequency is then

$$g_k(s_{ij}) = \tau_k(s_{ij})\iota_k(s_{ij}).$$

When considering n-grams of size n CIDEr$_n$ is defined as

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m}\sum_{j=1}^{m} \frac{\mathbf{g}^n(c_i) \cdot \mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\|_2 \, \|\mathbf{g}^n(s_{ij})\|_2},$$

where c_i is a candidate sentence for image I_i and S_i is the ground-truth set of descriptions for I_i, and m is the size of the set S_i. $\mathbf{g}^n$ is a vector with entries that consist of all n length n-grams. Notice that the dot product will be large when a candidate description shares n-grams with its according ground-truth description, particularly if these n-grams are rare across the corpus. For robustness n is often taken in the range 1 to 4 and the resulting CIDEr$_n$ values are averaged,

$$\text{CIDEr}(c_i, S_i) = \frac{1}{4}\sum_{n=1}^{4} \text{CIDEr}_n(c_i, S_i).$$

For further robustness it is common to use CIDEr-D instead of CIDEr,

$$\text{CIDEr-D}_n(c_i, S_i) = \frac{10}{m}\sum_{j=1}^{m} e^{\frac{\delta^2}{2\sigma^2}} \frac{\min(\mathbf{g}^n(c_i), \mathbf{g}^n(s_{ij})) \cdot \mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\|_2 \, \|\mathbf{g}^n(s_{ij})\|_2}$$

where δ is the length difference between c_i and s_{ij}, and σ is 6. Again averaging for n over the range 1 to 4 gives

$$\text{CIDEr-D}(c_i, S_i) = \frac{1}{4}\sum_{n=1}^{4} \text{CIDEr-D}_n(c_i, S_i).$$

CIDEr-D is intended to prevent gameability where a gaussian length penalty and a minimum function are added to prevent candidate sentences from learning to become unreasonably long. The multiplication by 10 centers scores for closer comparison against BLEU.