Video Quality Assessment

Learning Progress Report

James Hsu
May 16th 2025

Learning Goals and Topics

- Understand metrics of non-reference VQA and their mathematical basis
  - SRCC, KRCC, PLCC, RMSE, nonlinear four-parametric logistic regression

1. Spearman Rank-Order Correlation Coefficient (SRCC)

- Measures **monotonic relationships** between predicted scores and subjective Mean Opinion Scores (MOS).
- Uses **rank differences**:
  - Given two sets of samples $X = \{x_1, x_2, \ldots, x_n\}, Y = \{y_1, y_2, \ldots, y_n\}$ assign ranks to each set as $R(x_i)$ and $R(y_i)$ respectively, then compute:

    $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$
  - $d_i = R(x_i) - R(y_i)$ is the rank difference for each pair
  - n is the number of samples
- Interpretation
  - $\rho = +1$: Perfect positive monotonic relationship (ranks are identical)
  - $\rho = -1$: Perfect negative monotonic relationship (ranks are exactly reversed)
  - $\rho = 0$: No monotonic relationship

1. Spearman Rank-Order Correlation Coefficient (SRCC)

- Example
  - A set of video $V = \{v_1, v_2, \ldots, v_n\}$
    - Each video has Mean Opinion Score (MOS) $Y = \{y_1, y_2, \ldots, y_n\}$
    - Each video would generate a predict score $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$

Video	MOS (Y)	Predicted (Ŷ)
V1	4.5	4.8
V2	3.2	3.9
V3	2.8	2.5
V4	1.7	1.9
V5	4.0	3.7

1. Spearman Rank-Order Correlation Coefficient (SRCC)

- Step 1: Calculate d_i^2 the rank difference for each pair

Video	MOS	Rank(Y)	Predicted	Rank(Ŷ)	$d_i = R_Y - R_{\hat{Y}}$	d_i^2
V1	4.5	1	4.8	1	0	0
V5	4.0	2	3.7	3	-1	1
V2	3.2	3	3.9	2	1	1
V3	2.8	4	2.5	4	0	0
V4	1.7	5	1.9	5	0	0

- Step 2: Calculate SRCC

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(0+1+1+0+0)}{5(25-1)} = 1 - \frac{12}{120} = 0.9$$

2. Kendall Rank Correlation Coefficient (KRCC)

- Measures **pairwise ranking agreement**:

  $\tau = \frac{C-D}{\binom{n}{2}}$

  - C: Concordant pairs
  - D: Discordant pairs
  - n: Number of sample
  - $\binom{n}{2}$: All possible pairs

- Concordant v.s. Discordant: sample (i, j)

  - Concordant: $x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$
  - Discordant: $x_i > x_j$ but $y_i < y_j$, vice versa

- Interpretation

  - $\tau = +1$: ranks are identical, $\tau = -1$: ranks are exactly reversed, $\tau = 0$: No significant relationship

2. Kendall Rank Correlation Coefficient (KRCC)

- Example
  - A set of video $V = \{v_1, v_2, \ldots, v_n\}$
    - Each video has Mean Opinion Score (MOS) $Y = \{y_1, y_2, \ldots, y_n\}$
    - Each video would generate a predict score $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$

Video	MOS (Y)	Predicted (Ŷ)
V1	4.5	4.8
V2	3.2	3.9
V3	2.8	2.5
V4	1.7	1.9
V5	4.0	3.7

2. Kendall Rank Correlation Coefficient (KRCC)

- Step 1: Compare all pairs
  - We'll compare each pair (i, j) where $i < j$ and check whether the ordering of Y and Ŷ is consistent (concordant) or opposite (discordant).

Pair (i, j)	Y_i vs Y_j	$\hat{Y}_i$ vs $\hat{Y}_j$	Type
(V1, V2)	4.5 > 3.2	4.8 > 3.9	Concordant
(V1, V3)	4.5 > 2.8	4.8 > 2.5	Concordant
(V1, V4)	4.5 > 1.7	4.8 > 1.9	Concordant
(V1, V5)	4.5 > 4.0	4.8 > 3.7	Concordant
(V2, V3)	3.2 > 2.8	3.9 > 2.5	Concordant

Pair (i, j)	Y_i vs Y_j	$\hat{Y}_i$ vs $\hat{Y}_j$	Type
(V2, V4)	3.2 > 1.7	3.9 > 1.9	Concordant
(V2, V5)	3.2 < 4.0	3.9 > 3.7	Discordant
(V3, V4)	2.8 > 1.7	2.5 > 1.9	Concordant
(V3, V5)	2.8 < 4.0	2.5 < 3.7	Concordant
(V4, V5)	1.7 < 4.0	1.9 < 3.7	Concordant

- Compute KRCC (Kendall's tau)

$$\tau = \frac{C-D}{\binom{n}{2}} = \frac{9-1}{10} = \frac{8}{10} = 0.8$$

SRCC v.s. KRCC

Aspect	SRCC	KRCC
Rank Transformation	Convert all data to ranks, then compute squared differences	Compare pairs to see which is higher
Sensitivity to Outliers	More sensitive (large d^2 has a big impact)	More stable
Value Range	$[-1, 1]$	$[-1, 1]$
Applicability	Suitable for measuring overall ranking trends	Suitable for measuring local consistency

3. Pearson Linear Correlation Coefficient (PLCC)

- Measures **linear correlation** between predicted scores and MOS:
  - Given two sets of samples $X = \{x_1, x_2, \ldots, x_n\}, Y = \{y_1, y_2, \ldots, y_n\}$

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

  - $\bar{x}$ and $\bar{y}$: the means of X and Y
  - The numerator is the covariance
  - The denominator is the product of the standard deviations of X and Y

- Interpretation
  - r = 1: Perfect positive linear correlation
  - r = -1: Perfect negative linear correlation
  - r = 0: No linear correlation (nonlinear correlation may still exist)

4. Root Mean Square Error (RMSE)

- Measures **the average error** between predicted values and subjective MOS scores:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  - y_i: Ground truth value of the i^{th} sample (MOS)
  - $\hat{y}_i$: Model-predicted value of the i^{th} sample
  - n: Number of samples

- Interpretation:
  - Smaller value → Prediction is closer to human judgment
  - Larger value → Higher prediction error, lower model accuracy

- Because of the squaring, RMSE is particularly sensitive to outliers.

Nonlinear Four-Parameter Logistic Regression

- Used to **align predicted scores to MOS scale** before computing PLCC/RMSE:

  $$f(x) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp[-\beta_3(x - \beta_4)]}$$

  - x: The predicted quality score from the model
  - $f(x)$: The mapped prediction score (used for PLCC comparison with MOS)
  - β_1: Upper asymptote (maximum limit)
  - β_2: Lower asymptote (minimum limit)
  - β_3: Slope, controls the rate of change of the curve
  - β_4: Shift, the center point of the sigmoid curve

- Why use this?

  - Model predictions ($\hat{Y}$) may differ in scale and aren't always linearly related to MOS.
  - A logistic function reshapes predictions to better match how humans perceive quality changes.
  - This improves PLCC/RMSE accuracy by reducing the impact of scale differences.

Learning & Question

- Learning: In VQA tasks, the model outputs a predicted quality score for each video. These are then compared against human-annotated MOS (Mean Opinion Scores).
  - PLCC measures correlation
  - SRCC/KRCC measure rank consistency
  - RMSE measures the actual size of prediction errors
- Question:
  - Since MOS is inherently subjective and may vary depending on the content type or domain, would it make sense to consider training separate models on domain-specific datasets (e.g., gaming, AI-generated content, animation, sports) to improve relevance and performance?
  - How much do the choice of evaluation metrics impact the final outcome or perceived quality?

Thank You