SU-330-GPD-T-544

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

Attribution Energy as a Companion Interpretability Metric for Grad-CAM in Deep Learning Image-Quality Assessment

J. Luce[1,2], H. Nguyen[1], S. Gros[1], N. Darwish[1], M. Bhandari[1], I. Lasko, A. Block[1], H. Kang[1], G. K. Thiruvathukal[2], J.C. Roeske[3]
(1) Department of Radiation Oncology, Stritch School of Medicine, Cardinal Bernardin Cancer Center, Loyola University of Chicago, Maywood, IL
(2) Department of Computer Science, Loyola University of Chicago, Chicago, IL
(3) Department of Radiation Oncology, Mayo Clinic Arizona, Phoenix, AZ

ABSTRACT

Purpose:
Deep learning models can match human judgments in perceptual rating tasks, yet the cues driving their predictions are often difficult to interpret. Grad-CAM and related saliency methods are used to visualize attribution, but heatmaps may be diffuse despite accurate predictions or visually compelling despite errors. We evaluate whether Attribution Energy (AE), a quantitative summary of Grad-CAM attribution, can serve as a companion interpretability signal that tracks expert-aggregated (consensus) image quality ratings.

Methods:
A CNN-based model observer was trained to predict consensus Likert-scale image quality scores (1–5) from axial head-and-neck CBCT slices across five categories: parotid, submandibular glands, larynx, mandible, and overall. Consensus scores were computed as a weighted average across expert raters. Grad-CAM maps were computed per slice. AE was defined as the volumetric sum of raw (pre-normalization) Grad-CAM values across each CBCT volume. AE was computed for 16 CBCT volumes per category and compared with consensus scores using Spearman rank correlation. Correlations were summarized across leave-one-out folds using median Spearman (ρ) and p-values.

Results:
AE was significantly correlated with consensus score for Parotid (median ρ = 0.71), Submandibular (median ρ = 0.89), Mandible (median ρ = 0.70), and Overall (median ρ = 0.67) (all p ≤ 0.01), except for Larynx (median ρ = 0.25, p = 0.37). The weaker Larynx association suggests structure-dependent behavior, with greater divergence between model attribution and expert scoring cues.

Conclusion:
AE complements Grad-CAM inspection by quantifying overall attribution strength, helping separate human-facing visual interpretability from the broader information the model may use to generate predictions.

Poster Number:
SU-330-GPD-T-544

CONTACT

Jason Luce
Stritch School of Medicine, Cardinal Bernardin Cancer Center, Loyola University of Chicago
2160 S 1st Ave, Maywood, IL 60153
jaluce@luc.edu

INTRODUCTION

- AI-enhanced CBCT may support reduced-dose image guidance, but structure visibility must be assessed before clinical use.
- Grad-CAM heatmaps (Fig. 1) are useful to qualitatively examine AI behavior yet are difficult to use in cross-case comparisons.
- Attribution Energy (AE) summarizes the whole-volume Grad-CAM attribution strength as a compact scalar companion to visual saliency maps.
- Objective: To test whether AE correlates with expert consensus image-quality ratings across head-and-neck structures in clinical-dose, low-dose, and AI-enhanced CBCT images.

METHODS AND MATERIALS

- This dataset included clinical-dose, simulated low-dose, and AI-enhanced H&N CBCT volumes.
- Reduced-projection CBCT simulated low-dose acquisition; AI-enhanced CBCT was generated from low-dose inputs using U-Net and cGAN image-to-image models.
- Five expert raters scored axial slices on a 1–5 Likert scale for parotids, submandibular glands, larynx, mandible, and overall quality.
- Weighted expert consensus scores served as the CNN model-observer target.
- AE was computed as the whole-volume sum of raw Grad-CAM values and compared with consensus score using Spearman correlation.
- Test set: 16 CBCT volumes per category.

RESULTS

- Positive association in 4 of 5 categories (Fig. 2)
  - Parotid: ρ = 0.71, p ≤ 0.01
  - Submandibular: ρ = 0.89, p ≤ 0.01
  - Mandible: ρ = 0.70, p ≤ 0.01
  - Overall: ρ = 0.67, p ≤ 0.01
  - Larynx: ρ = 0.25, p = 0.37

DISCUSSION

- AE may serve as a lightweight companion signal for cross-case interpretability comparisons.
- Positive associations suggest overall attribution strength often tracks perceived image quality.
- The weaker larynx association suggests anatomy-specific dependence.
- AE is a global summary and is not intended to validate spatial localization.
- Future work could add concentration, dispersion, or attribution-stability metrics

CONCLUSIONS

- AE provides a compact companion to Grad-CAM visualizations.
- In this proof-of-concept study, AE aligned with consensus quality ratings in most categories.
- AE may help summarize and compare attribution behavior across CBCT volumes

FIGURES

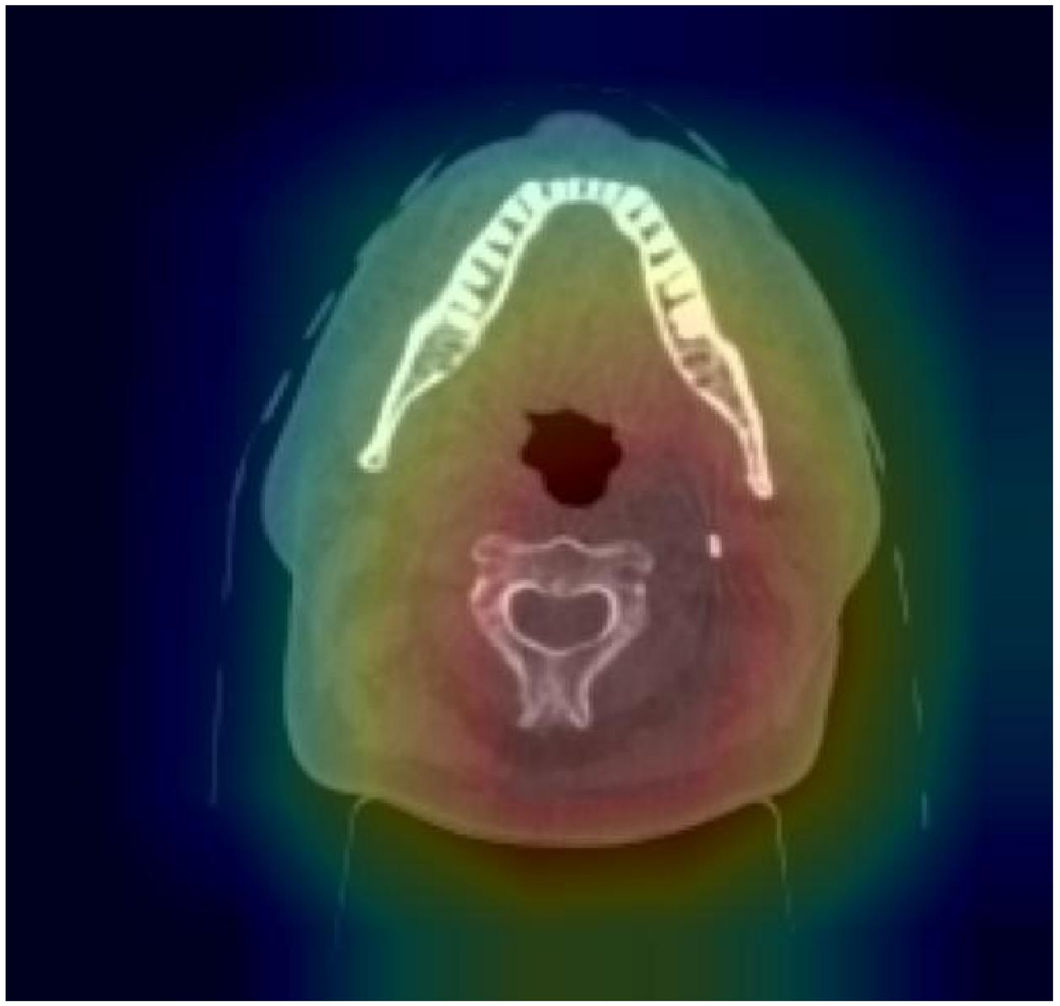

Figure 1. Example Grad-CAM attribution map rendered as an RGB heatmap, where warmer colors indicate higher attribution.

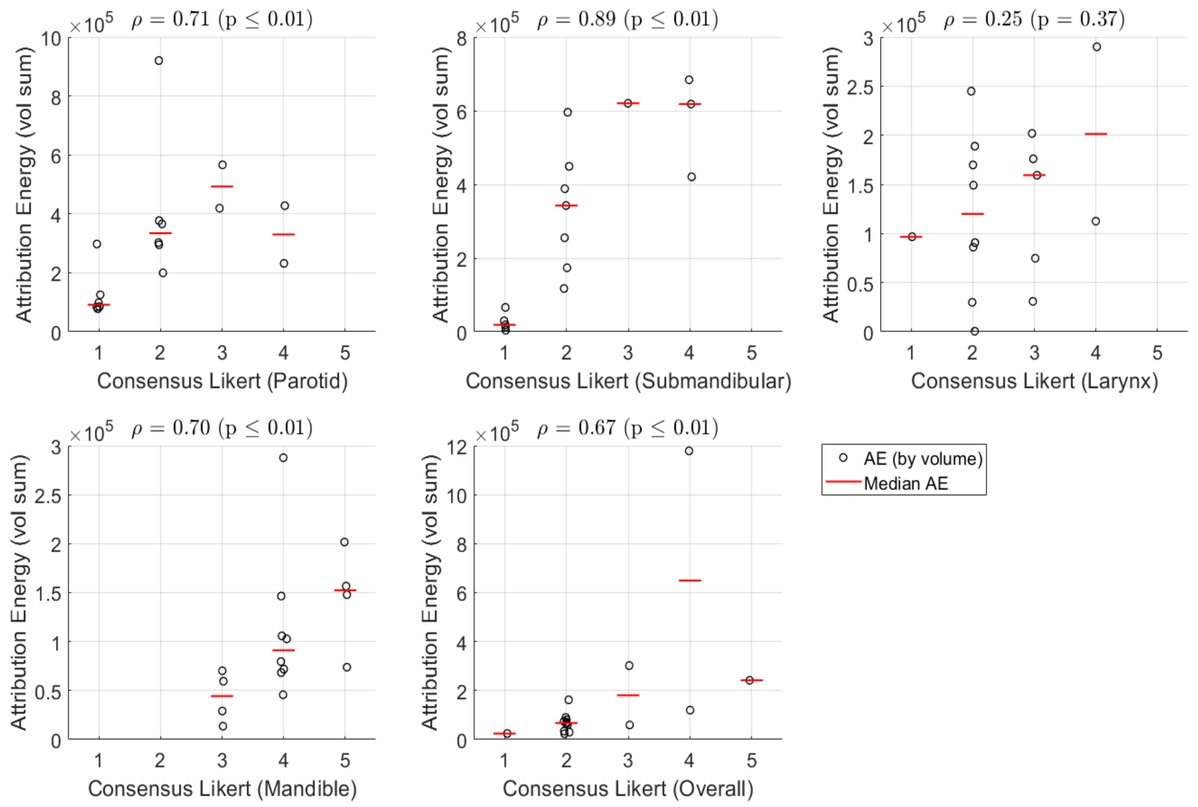

Figure 2. Volume-level AE versus consensus Likert score by head-and-neck category. Markers indicate volumes; red bars show the median AE within each Likert level. Y-axis ranges are scaled independently by category to improve visualization of within-structure trends. Annotations report median leave-one-out Spearman ρ (p-value). Trends are presented as monotonic relationships, not classification thresholds.