Eye Tracking with Speech Recognition Can Automatically Produce Labels for Deep Learning from Standard Clinical Interpretations: A Proof-of-principle Application to Fully Simulated Interpretation of Brain Metastases on MRI

Joseph N. Stember, MD, PhD, Columbia University Medical Center-NYPH; Andrei Holodny, MD; Robert Young, MD; Nathaniel Swinburne, MD; David Gutman, MD; Krishan Juluru, MD; Sarah Eskreis-Winkler, MD, PhD; Haydar Celik, PhD; Ulas Bagci, PhD; Sachin Jambawalikar, PhD; Elizabeth A. Krupinski, PhD, FSIIM; Peter D. Chang, MD

Introduction

Despite advances in deep learning (DL) using convolutional neural networks (CNNs) in radiology, a key hindrance remains generating sufficient amounts of labeled data. Our goal is to address this bottleneck by obtaining a potentially enormous windfall of labeled data automatically from clinical workflow, for essentially all lesions types, here demonstrating proof-of-principle applied to brain metastases on MRI.

Our approach is based on the simple and intuitive premise that radiologists look at the structures they are analyzing and describing. For this reason, eye tracking (ET) is the path toward harvesting labeled data automatically from clinical radiologic interpretations. ET allows a computer to know precisely where and when a radiologist is looking within an image. As such, combining ET with speech recognition (SR) can harness expert-labeled imaging data automatically during routine clinical image interpretation. Again, we illustrate a proof-of-principle here using brain metastasis detection and localization.

Hypothesis

We hypothesized that within a fully simulated clinical interpretation of brain metastases on MRI, by capturing ET data and the times at which certain key words were dictated via SR, we could accurately localize the lesions. We further hypothesized that these could serve as DL labels, and training a CNN using them could accurately detect lesions on new images.

Methods

We analyzed 785 images from the BraTs public brain tumor dataset. The first 700 were viewed during ET and while simulating the typical voice dictation that occurs during clinical radiological interpretations. We used Google Cloud Speech-to-Text to record the times at which any of three keywords (“tumor”, “mass” or “lesion”) was uttered. Using these times, we extracted mean gaze positions for each image. Finally, we trained a keypoint detection CNN using these 700 images + gaze points. We trained the CNN for 50 epochs with a learning rate of 0.001, using an Adam optimizer with mean average error as the loss function. We then applied the trained CNN to predict lesion location on the remaining test set of 85 images.

Results

ET with SR points were within lesion bounding boxes for 658 out of 700 training set images, a localization accuracy of 93%. The resulting trained CNN predicted lesion location for the test set with an accuracy of 85%, improving to 89% for predictions within 5 mm of the bounding box, which we judged adequate to draw one’s attention to the lesion. The receiver-operating-characteristic area-under-the-curve was calculated to be 0.97.

Conclusion

We have demonstrated proof-of-principle that one can automatically extract labeled images for deep learning from clinical radiology interpretations without added radiologist input. We have shown that the resulting trained network can accurately predict lesion location.

Statement of Impact

This work has significantly broader implications than brain metastases alone. We have demonstrated feasibility for dramatically accelerating deep learning to detect essentially any lesion type or structure of interest by automatically extracting tremendous amounts of labeled images from clinical radiological interpretations.

Keywords

deep learning, convolutional neural networks, keypoint detection, eye tracking, brain tumors