Real, real fakes, really fake - image authenticity exploration

(Re)-implement a model for image classification on real vs. synthetic image data, cf. Bird & Lotfi (2024, https://ieeexplore.ieee.org/abstract/document/10409290), to:

Main task
- Test the capabilities of NNs to distinguish between 'real' and 'fake' image data
  - A task which is becoming harder and harder for the human eye to solve, but also
  - Gaining more relevance: e.g. dangers of synthetic image/video 'evidence', deepfakes, …

Research angles for the project:
A) As pointed out by the authors: implement an attention-based approach and generate attention heat maps
  - Discern which areas were relevant for classification
  - Compare results with the reported gradient class activation maps from the paper

B) Test the trained model on synthetic vs real image data from a different dataset to examine…
  - Training dataset 'bias' (if the new test data has the same content classes as the CIFAKEt raining data)
  - Modle limitations (if content classes differ and the model cannot generalise)
  - Possible ethical implications should images be misclassified more frequently in the new dataset, which would have real-world consequences in downstream applications if content class effect are not sufficiently researched (in this regard, confidence scores for classification might be of interest)

Evaluation with quantitative methods could be done on CIFAKE and an additional dataset (e.g. https://www.kaggle.com/datasets/cashbowman/ai-generated-images-vs-real-images/code), both of which would be labeled for supervised methods. To explore difficulty of the task, visualising learning progress per epoch is of interest,

If accessible/feasible, for A) the overlap between gradient class activation maps and attention maps could be calculated (though the code for the original paper does not seem available, so this likely exceeds the scope of the project). Nevertheless, attention maps would aid explainability of results and allow for some speculation about which features seemed to motivate classification.

For B), additional qualitative analysis of false positives/negatives is likely a valid approach to discover concerns arising from misclassification.

Resources
Dataset CIFAKE (https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images/data)

Bird, J. J., & Lotfi, A. (2024). CIFAKE: Image classification and explainable identification of AI-generated synthetic images. *IEEE Access*, *12*, 15642–15650. https://doi.org/10.1109/access.2024.3356122

(…plus some projects on e.g. Kaggle to jumpstart the code for classification)