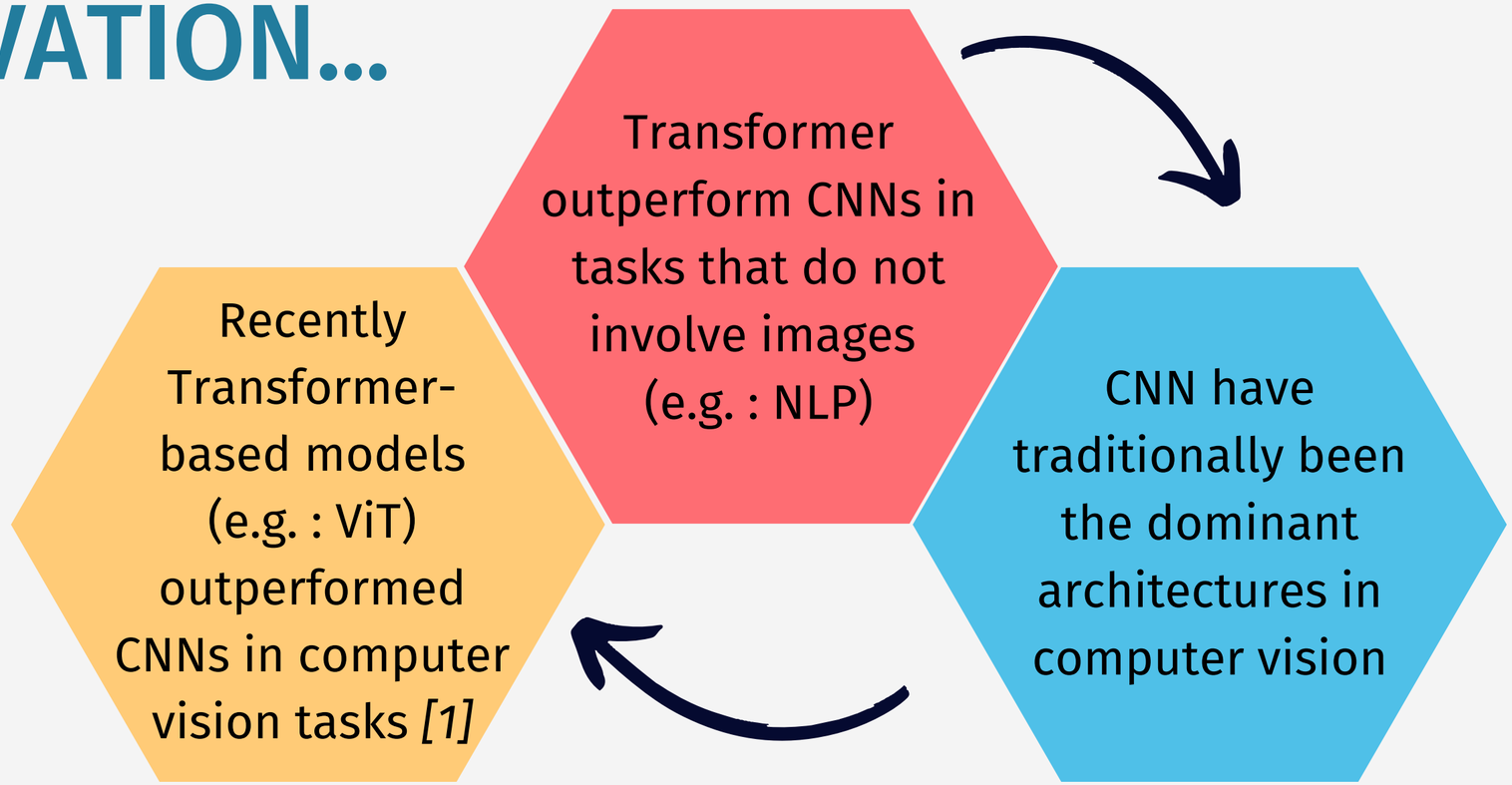
Scene Classification

CESARIO LAURA THOFT (1852596)
CICIANI DIEGO (2140394)
ODDI LIVIA (1846085)
ZELLER DAMIAN (2118831)

MOTIVATION...



...AND TASK

Replicate the Vision Transformer (ViT) model from the research paper published as a conference paper at ICLR 2021 "*An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*" (Dosovitskiy et al.), and apply the same pipeline to a new large dataset Places365.

MODELS

3 Vision Transformers (ViTs) pre-trained on different image datasets (ImageNet-21k, JFT-300M)

CNN models pre-trained as well

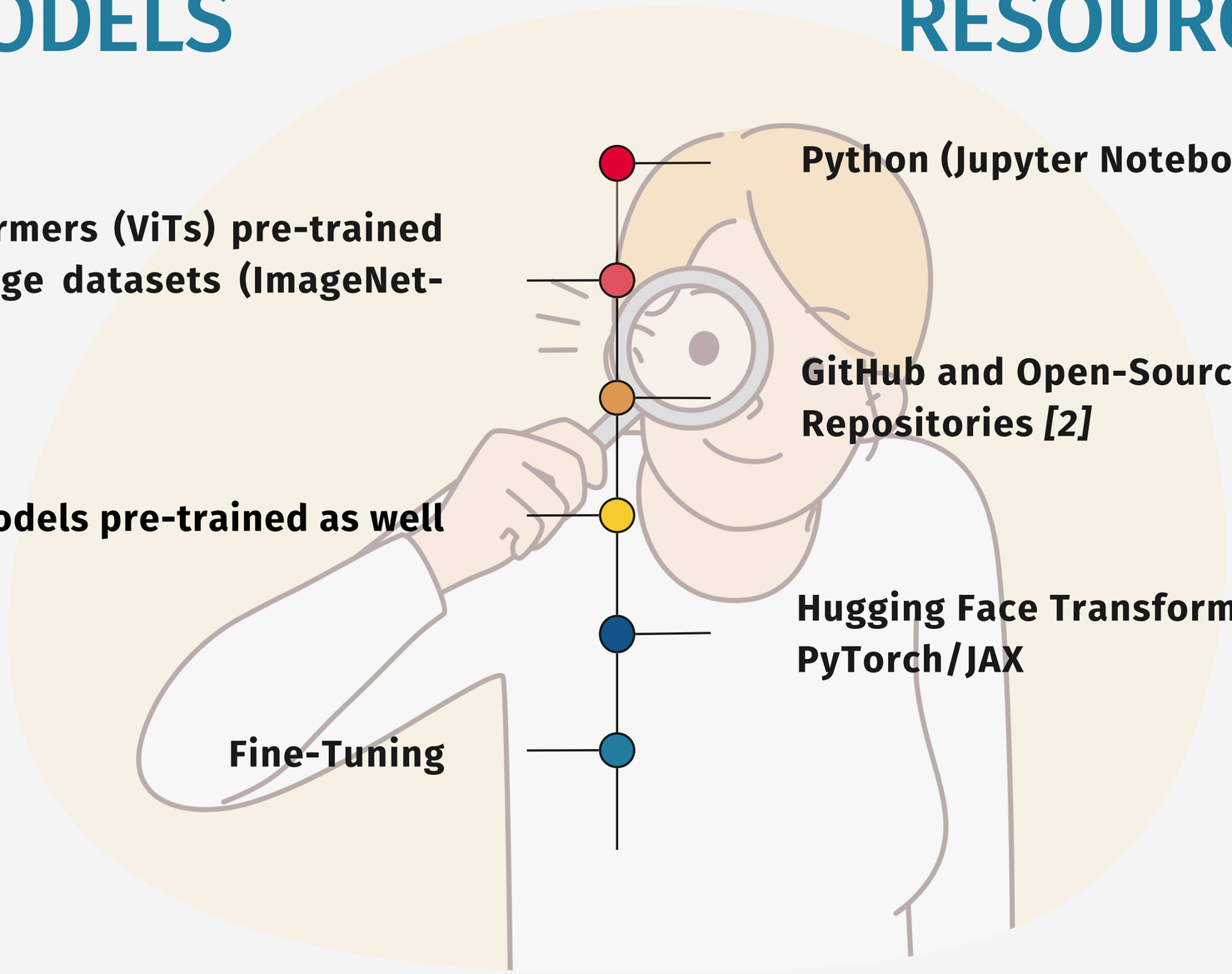
Fine-Tuning

RESOURCES

Python (Jupyter Notebooks)

GitHub and Open-Source Code Repositories *[2]*

Hugging Face Transformers and PyTorch/JAX



ANALYSIS

Attention Visualizations

Use ViT’s attention maps to identify which parts of an image are most influential in decision-making.

Class-Wise Analysis and Confusion matrix

Measure F1-score per category to understand performance disparities. Identify which categories are frequently misclassified.

Data

Places365-Standard dataset *[3] [4]*, a large-scale benchmark with 365 scene categories, used for evaluating scene recognition models

THANK YOU

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

[2] Data Source: Google Research Vision Transformer (GitHub). Available at : https://github.com/google-research/vision_transformer

[3] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6), 1452–1464.
Avaiable at: https://doi.org/10.1109/TPAMI.2017.2723009

[4] Places365 dataset, retrieved from http://places2.csail.mit.edu/download-private.html