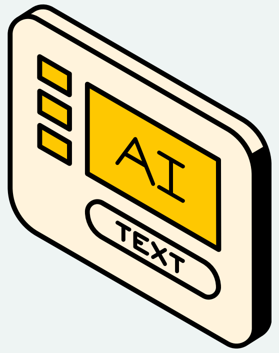
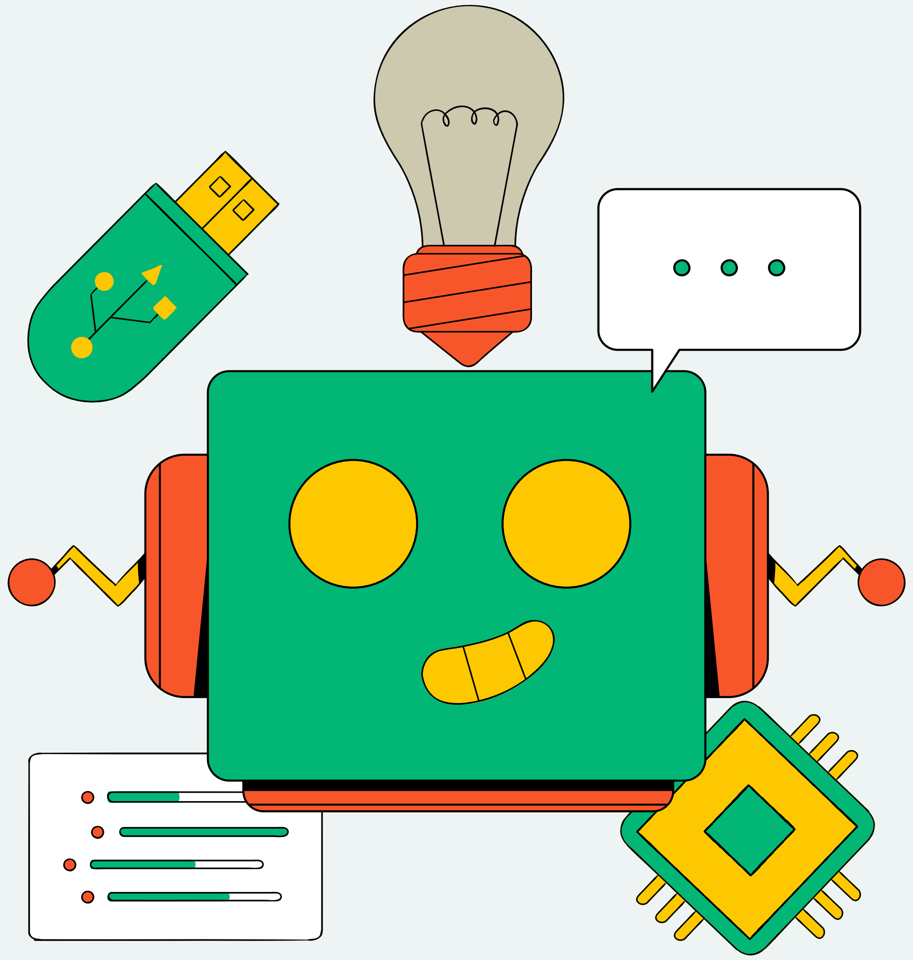


SCENE CLASSIFICATION

FINAL PROJECT PRESENTATION



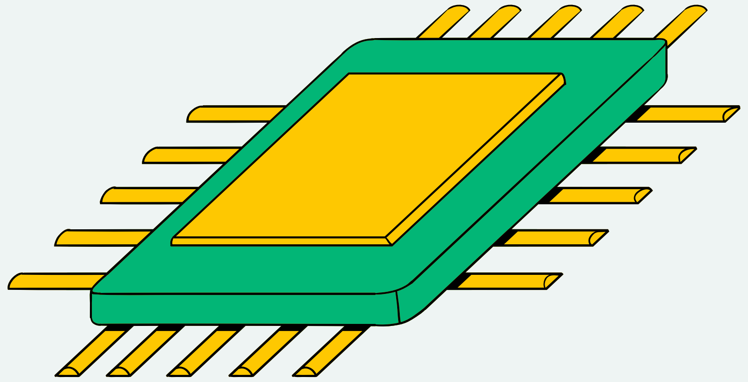
PRESENTED BY:

CESARIO LAURA THOFT (1852596)

CICANI DIEGO (2140394)

ODDI LIVIA (1846084)

ZELLER DAMIAN (2118831)



PRESENTATION OUTLINE

- Problem Statement
- Model selection
- Dataset
- Fine-Tuning
- Experiments and results
  - Accuracy metrics
  - Computational cost vs. Accuracy
  - Attention maps
- Future Work

PROBLEM STATEMENT

Replication and fine-tuning of 5 pre-trained models from ICLR 2021 paper:
"An Image is Worth 16x16 Words" (Dosovitskiy et al.)

Evaluate and compare fine-tuning performance of Vision Transformers (ViTs), Hybrid CNN-Transformer models, and advanced CNN-based models on the Places365 dataset.

A scene classification dataset differing from object-centric datasets like ImageNet-21K

Focus shifted from general image recognition tasks to scene classification on Places365.

Extend existing research by:

- Analyzing model performance on scene-centric tasks
- Exploring computational trade-offs in fine-tuning
- Using attention maps to visualize the decision-making of the model

MODEL SELECTION

VIT'S

- **ViT-B/32:** Computationally efficient baseline, captures global spatial relationships
- **ViT-B/16:** Finer-grained feature extraction, balances cost and performance

HYBRID MODELS

- **R50+ViT-B/16:** Combines ResNet-50's local feature extraction with ViT's global reasoning.
- Evaluates whether merging CNN inductive biases with Transformers improves scene classification.

CNN'S

- **EfficientNet-L2 :** Provides a strong computational efficiency baseline
- **BiT-L (ResNet152x4):** Represents traditional CNN approaches for comparison in spatial contexts

DATASET

-

PLACES 365

365 scene classes and 1.8M images

Computational complexity

Train/Val/Test Split (60/20/20)

based on diversity and real-world relevance

balanced representation

20 classes (4,000 images , 200 per class) created with stratified subsampling

FINE-TUNING

Places 365

Trial Number	Learning Rate (lr)	Momentum	Validation Accuracy
0	2.35E-05	0.898099309	36
1	0.00085604	0.773994161	40
2	6.87E-05	0.791928096	38
3	2.61E-05	0.915534275	38
4	0.002361648	0.972915928	36
5	0.000594294	0.704122817	34
6	0.000127275	0.736889905	34
7	0.000868238	0.870753416	36
8	0.001042364	0.899503888	34
9	0.002427862	0.825105154	34

Adopted fine-tuning protocol from paper, if reasonable

For all models:

- **Hyperparameter Tuning:** Automated tuning with Optuna for learning rate and momentum
- **Optimizers**: SGD with momentum from tuning, weight decay = 0 and gradient clipping (max_norm=1.0)
- **Learning Rate**: Cosine decay with warm-up; learning rate set based on tuning results
- **Batch Size**: 32
- **Epochs**: 10 epochs
- **Loss Function**: Cross-entropy loss

ACCURACY METRICS 1/3

MODEL	OVERALL ACCURACY
R50 + ViT-B-16	87.75
ResNet152x4	86.0
ViT-B-16	85.25
ViT-B-32	82.75
EfficientNet-L2	8.25

ACCURACY METRICS 2/3

EXPECTED RESULTS

- ViT models excel in global context understanding, achieving strong performance.
- ResNet152x4 performs well, due to its deep architecture enabling capturing complex spatial relationships
- EffNet's low performance reflects its simpler architecture and limitations in adapting ImageNet-21K pre-training.

UNEXPECTED OBSERVATIONS

- Classes like Canyon and Castle challenge all models (possibly) due to limited similar examples in ImageNet-21K, leading to suboptimal transfer learning for these specific categories
- R50 + ViT-B-16 perform exceptionally well, (possibly) benefiting from hybrid designs and dataset compatibility.

ACCURACY METRICS 3/3

FINDING CONSISTENT WITH THE RESEARCH PAPER?

- Results largely align with research findings:
- ViT-B16 Outperforms ViT-B32

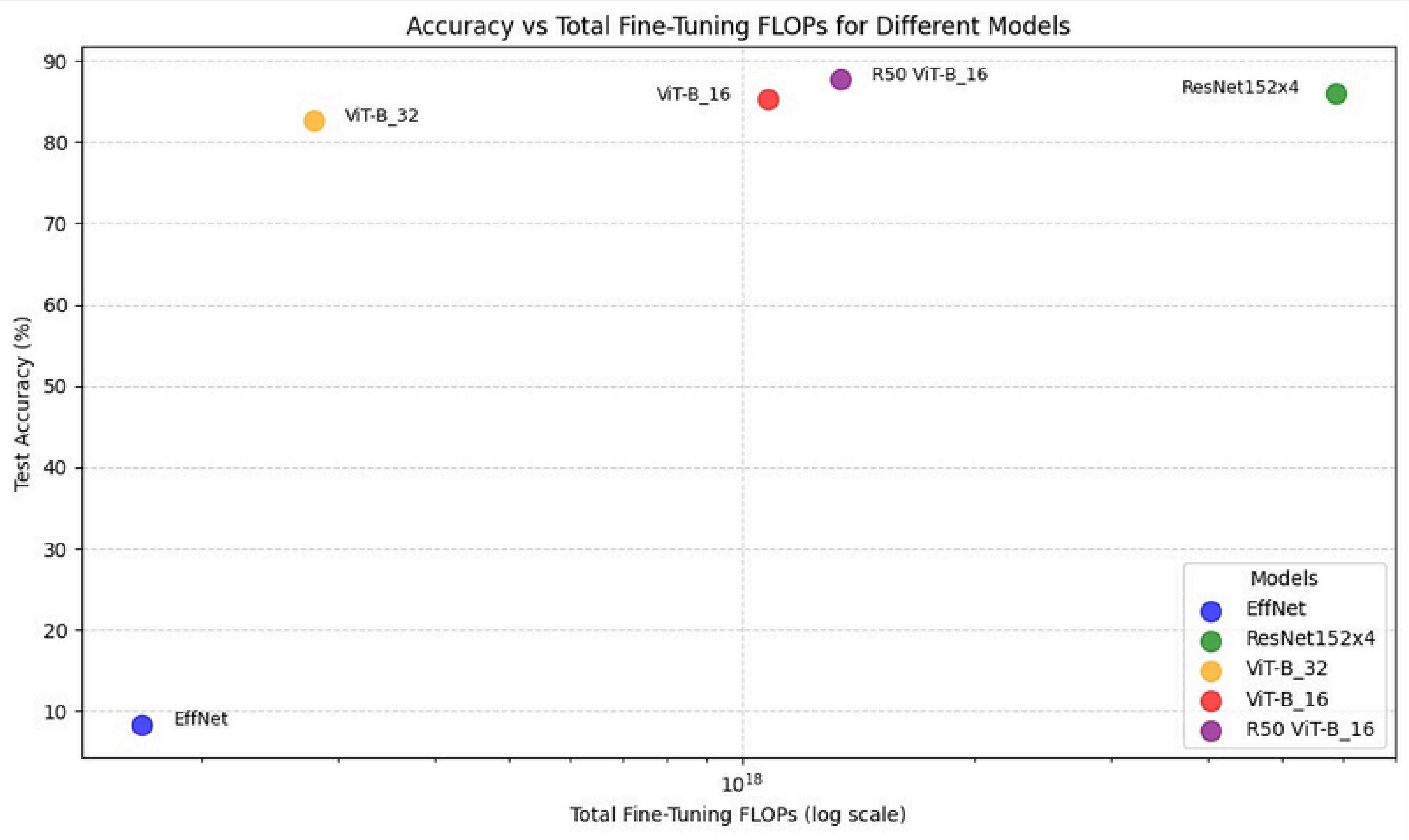
DISCREPANCIES

- EffNet underperforms compared to research due to smaller model size or training scale.
- R50+ViT-B/16 outperforming all other models

DATASET EFFECT

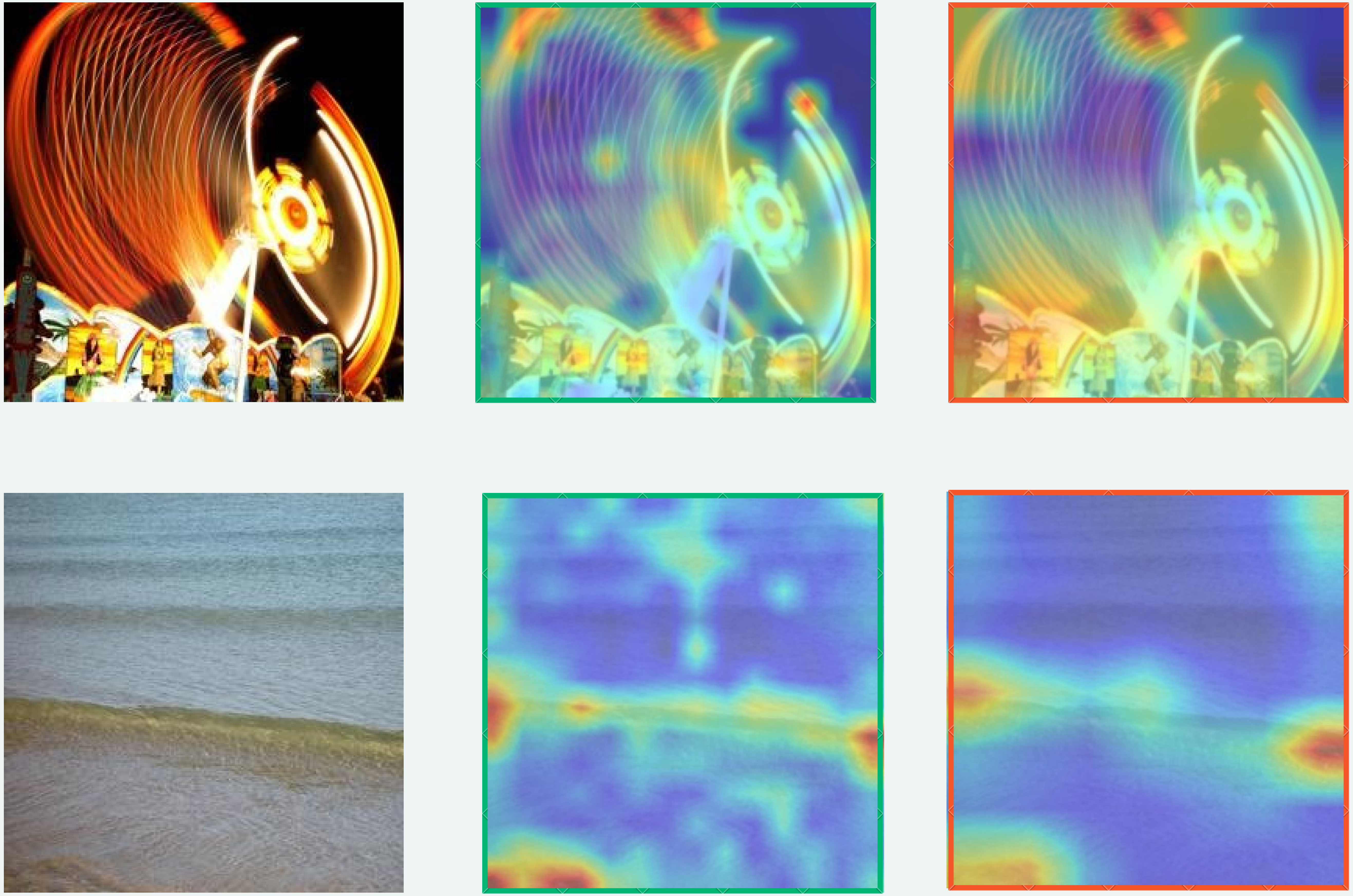
- Small fine-tuning datasets favor hybrid models, due to ResNet backbone
- ImageNet-21k Pretraining: JFT-300M pretraining in the research paper allows for further performance gains, particularly for ViT models
- Scene-Specific Challenges: Limited representation of scene-centric classes in ImageNet-21k

ACCURACY VS FLOP



ORIGINAL IMAGE | ViT-B_16 | ViT-B_32

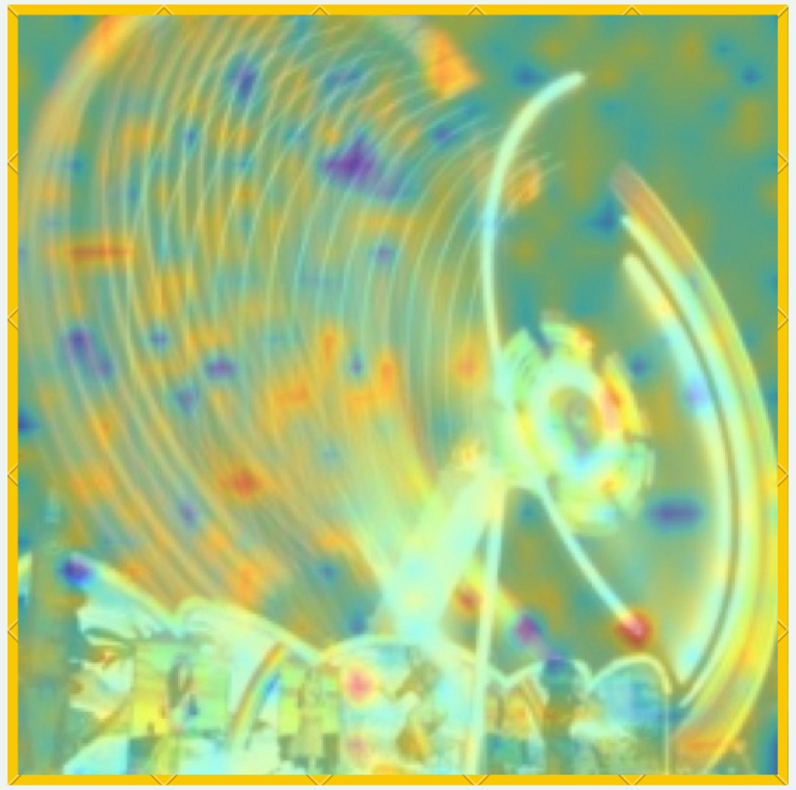
ATTENTION MAPS

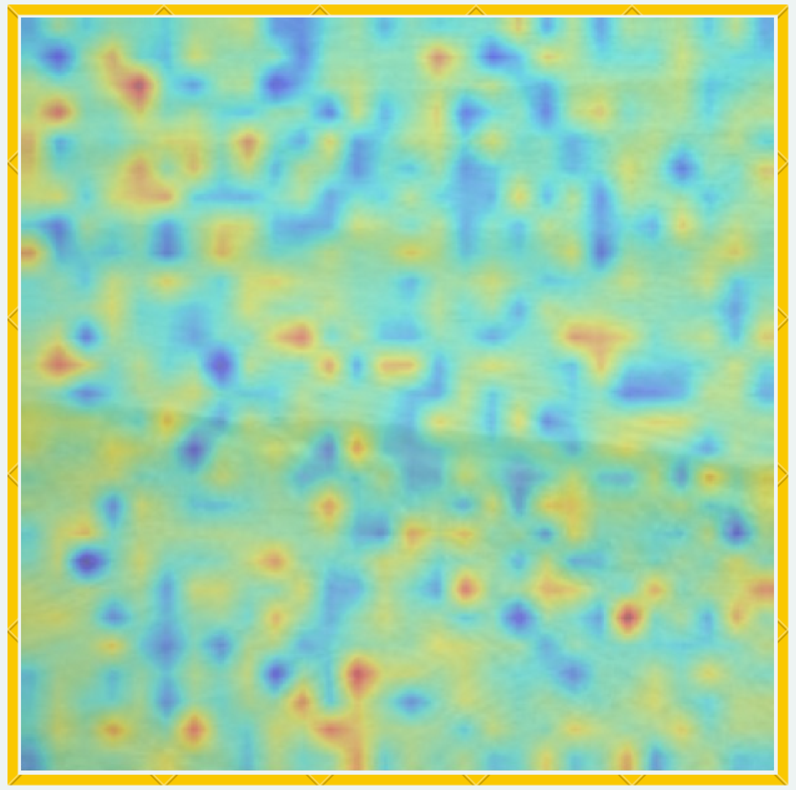


ORIGINAL IMAGE

RESNET50_ViT-B_16

ATTENTION MAPS





TAKEAWAYS

Quality of attention maps

ViT-B_16 + ViT-B_32 highly interpretable attention maps

ResNet50-ViT-B_16 struggles to generate clear attention maps

Impact of patch sizes

ViT-B_16 smaller patches

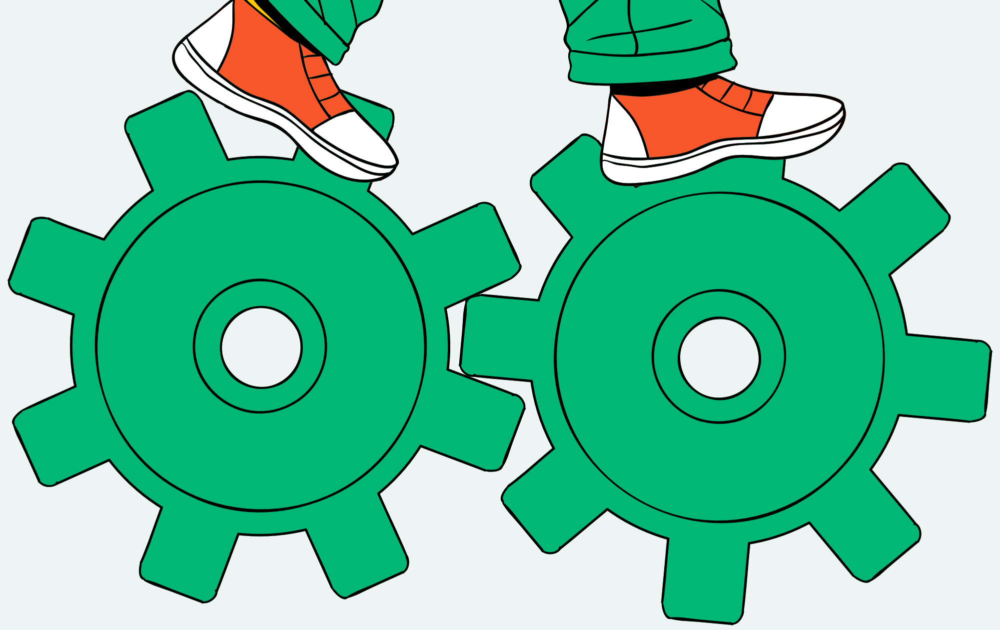
ViT-B_32 larger patches

Design trade-off

ResNet50-ViT-B_16
- high accuracy
- compromise on interpretability

ViT models seem to be the more reliable choice.

FUTURE WORK



Explore self-supervised pre-training for ViTs on Places365

Test on other datasets for broader generalization.

Investigate efficient ViTs for resource-constrained scenarios.

Explore domain-specific pre-training on Places365 for improved performance.

THANK YOU!