QUESTIONS

* Now that I don’t use pretrained model what image dimension should I use ?
* I need 2 papers. There is the first basic one, what about the second ? Can I use the Keras tuto, the French website, another paper bringing what ?
* Let’s check mon plan prévisionel.
* L’analyse de la feature importance globale et locale du nouveau modèle → Je fais comment avec ma classification d-images ?
* Je peux/dois mettre des images dans la note méthodologique ?
* la méthodologie de modélisation, la métrique d'évaluation retenue et sa démarche d'optimisation ?

<https://www.kernix.com/article/la-vision-transformers-vit-cette-nouvelle-technologie-qui-revolutionne-le-traitement-des-images/> → Good text explaining in French the details.

1. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (2020)

This is the seminal paper that introduced Vision Transformers. While it's from late 2020, it's the foundational work for ViT models and is crucial for understanding the basic concepts.

Authors: Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby

ArXiv link: <https://arxiv.org/abs/2010.11929>

Key points:

* Introduces the concept of treating image patches as "words" for transformer models
* Demonstrates that transformers can be directly applied to image classification tasks
* Shows that with sufficient training data, ViTs can outperform convolutional neural networks

1. "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet" (2021)

This paper proposes an improved ViT architecture that addresses some limitations of the original ViT model.

Authors: Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, Shuicheng Yan

ArXiv link: <https://arxiv.org/abs/2101.11986>

Key points:

* Introduces a new tokens-to-token (T2T) process to overcome limitations of the original ViT
* Proposes an efficient backbone for vision transformers
* Demonstrates improved performance on ImageNet without requiring large-scale pre-training

These papers provide a solid foundation for understanding Vision Transformers and their evolution. The first paper introduces the core concept, while the second one shows how the model has been refined and improved upon.

When reading these papers, pay attention to:

1. How images are divided into patches and processed
2. The role of positional embeddings in maintaining spatial information
3. How self-attention mechanisms are applied to image data
4. The differences in architecture and performance compared to traditional convolutional neural networks

Understanding these concepts will give you deeper insight into how your simplified ViT model works and how it relates to the broader field of computer vision using transformers.