CNNs are successful in computer vision applications
Transformers have revolutionised NLP

- Transformer use in computer vision is limited

Transformers require fewer parameters

Paper says that a pure transformer can work well on sequences of image patches

Transformers

- Transformers are models that operate on sequences (**sets**)
    - You have a set of tokens (usually words)
    - The transformer takes in tokens and computes attention on them
    - **Attention -** quadratic operation
        - You have to calculate the pairwise inner product between each pair of the tokens
            - **Pairwise inner product -** if you have a set of 5 tokens, you compute the product between every pair of tokens, so you have 25 connections
        - Hence, transformers work very well in NLP but are **limited by the memory requirements of computing the attention**
            - Images are therefore much harder for transformers
                - Images are grids / volumes of pixels which have a ton of data
            - *Every single pixel has to attend to every other pixel in the image*
        - This is somewhat similar to convolutional neural networks, except that each pixel has a relatively small receptive field for that layer of only the pixels in its near proximity, as you go deeper into the network, each pixel has a higher effective receptive field as they combine
    - **Problem: Transformers are able to attend to every pixel from every pixel everywhere**

Steps

1. You divide the input image into **patches**
2. You **unroll the patches**
3. Consider it as a **sequence of patches**, much like a sentence
4. You prepend the sequence with a **classifying token CLS**
    - Also passed through the transformer
        - Associated with **no location in the image**
5. You treat the patches as **word embeddings**
6. You put the patches through **one fully connected layer to get the token embeddings**
7. You then put those embeddings through a **transformer**

- The transformer keeps the **length of the sequence the same**
    - *Not necessary but it's just how we do things*
    - Possibly makes it easier to chain transformers