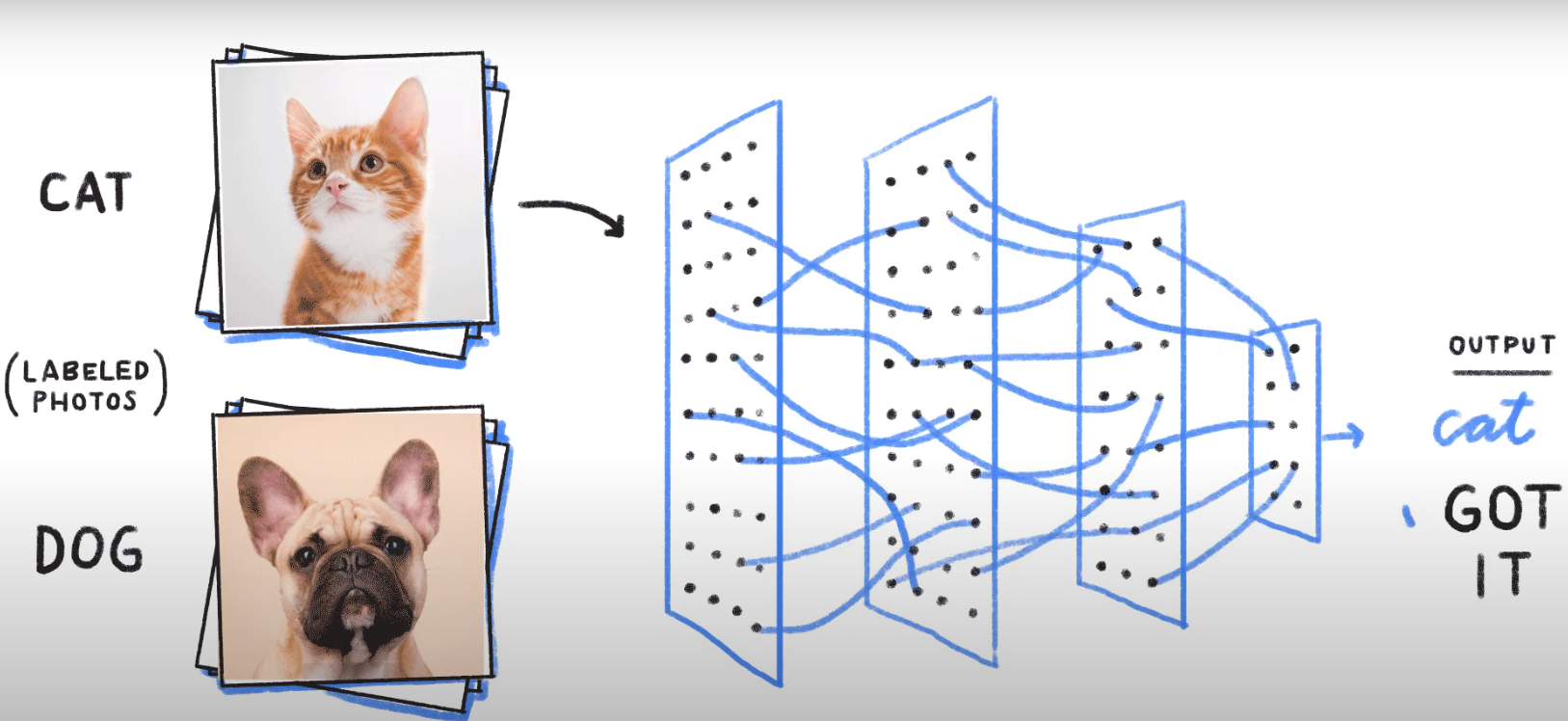
**Vision Transformers- A summary by Utkarsh Babbar**

**How do CNNs work?**

Small filters to compress information towards a general answer.

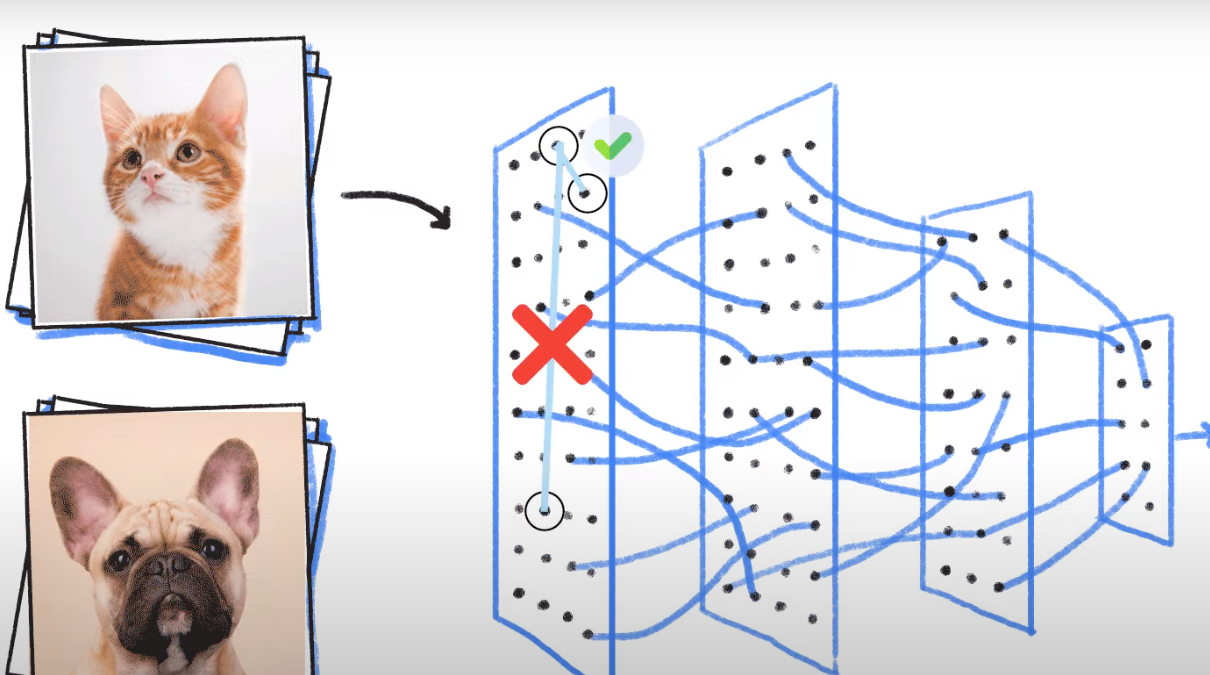


Strengths and weaknesses of CNN

**Strengths:**

1. CNN is capable of recognizing an object even after the image’s orientation is changed or it is zoomed in or zoomed out.

**Weaknesses:**



1. Only designed for images
2. It is domain dependent and each pixel represents varying importance.
3. Main issue is it lacks global understanding of the images, cant capture features over long distance. The features present in the bigger picture might not be captured well.
4. Computationally its very expensive

In the figure above we see that the local pixels

**How does Vision transformers help?**

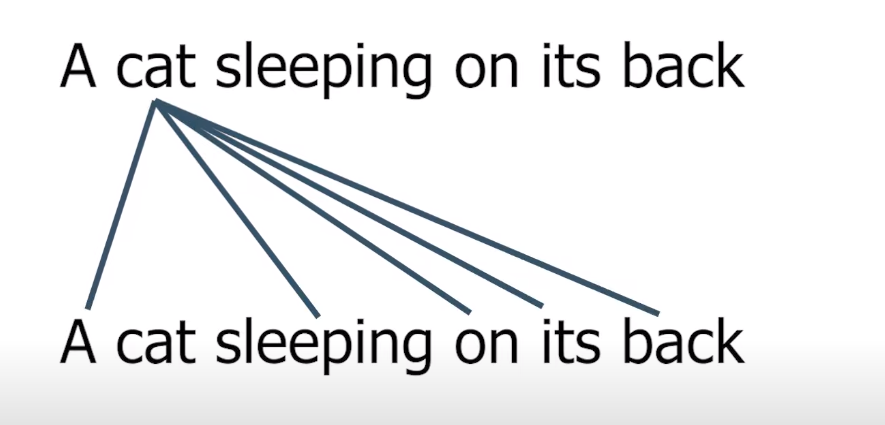
Unlike CNN which uses a pixel array, visual transformers divides the image into visual tokens, or patches ,linearly embed each of them

***Advantages of transformers:***

1. They have a lot more memory.
2. They can extract information simultaneously from the input and the interrelation.

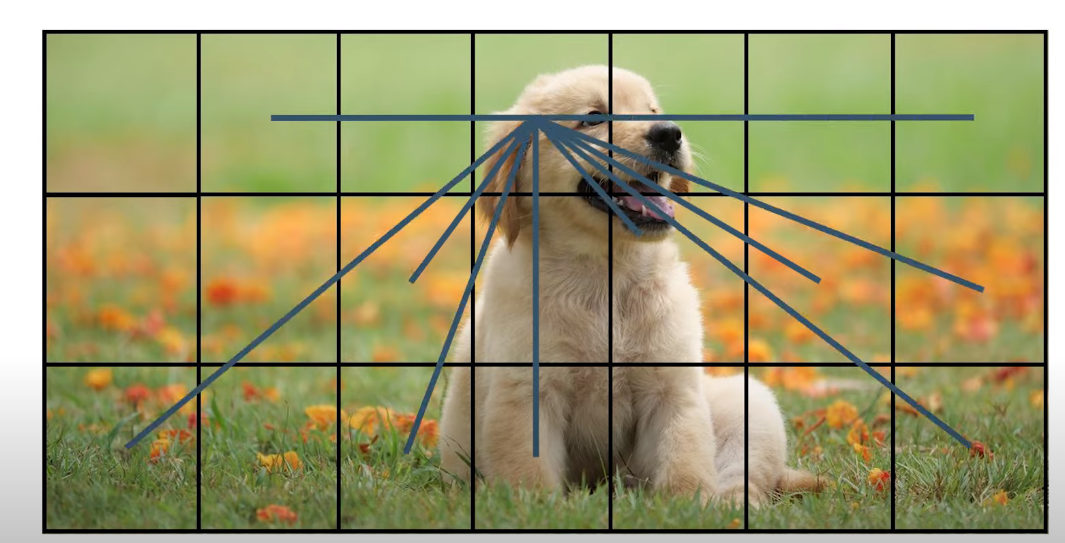
**Transformers in NLP**

The concept of self attention



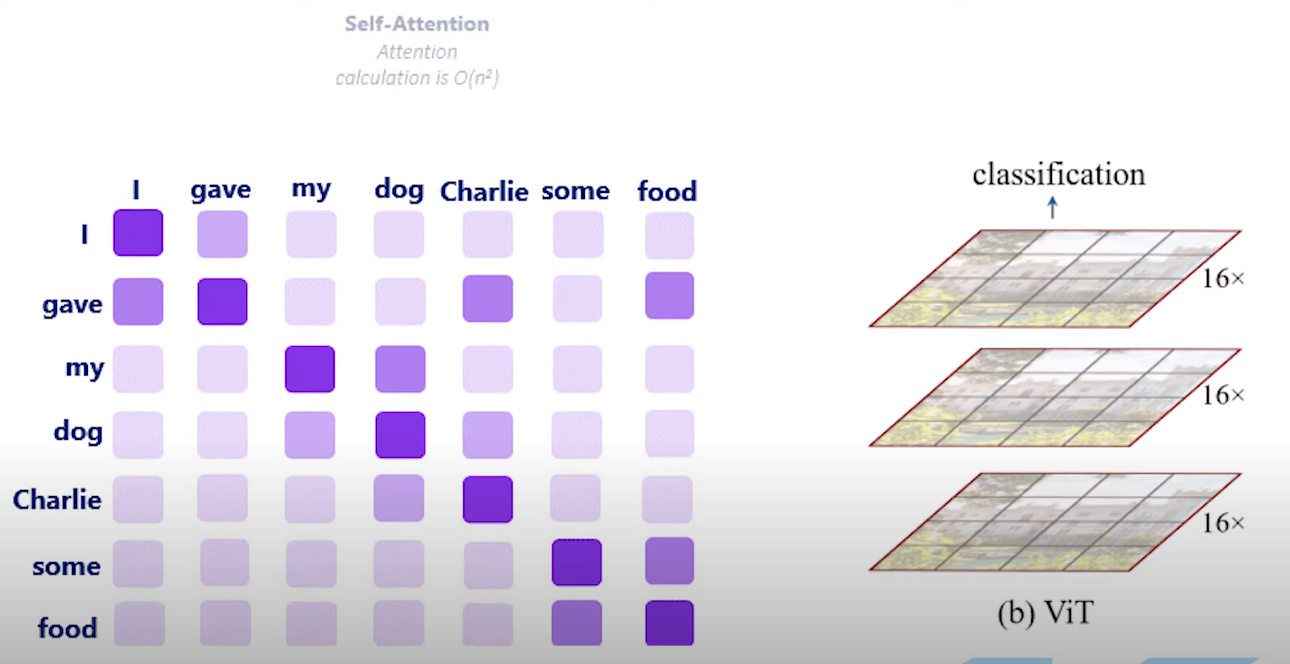
This technique is not just limited to NLP. Researchers have found it to be a great fit for images too.

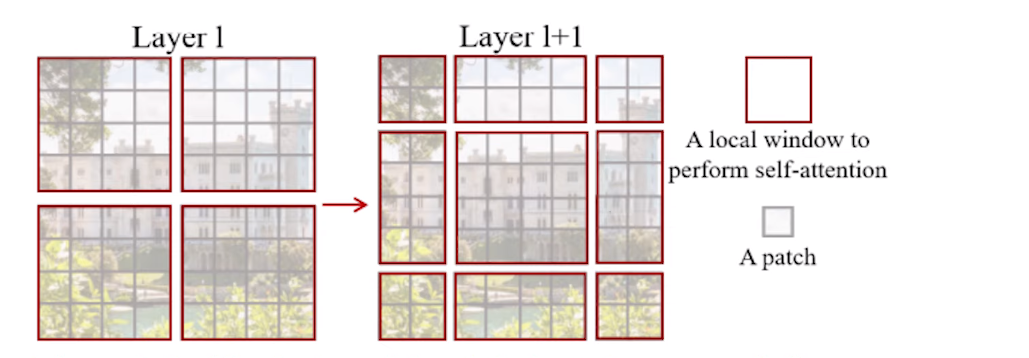
**Use of Transformers in images:**



“A picture is worth a thousand words”, yeah pictures contains much more information than sentences.

The cost of self-attention is quadratic. If we pass each pixel of the image as input, then self-attention would require each pixel to attend to every other pixel. The quadratic cost of the self-attention will be very costly and not scale to realistic input size; hence, the image is divided into patches.



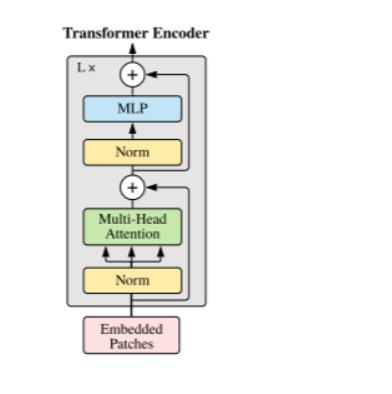


* **Image is first split into fixed-size patches.**

The 2D image of size H \*W is split into N patches where N=H\*W/P²

* **Flatten the 2D patches to 1D patch embedding and linearly embed them**

Each patch is flattened into a 1D patch embedding by concatenating all pixel channels in a patch and then linearly projecting it to the desired input dimension.



* **Multi-Head Self Attention Layer(MSP)** to concatenate the multiple attention outputs linearly to expected dimensions.Themultiple attention heads help learn local and global dependencies in the image.
* **Multi-Layer Perceptrons(MLP) contains two-layer with Gaussian Error Linear Unit(GELU)**
* **Layer Norm(LN) is applied before every block** as it does not introduce any new dependencies between the training images. **Help improve the training time and generalization performance.**
* **Residual connections are applied after every block as they allow the gradients to flow through the network directly without passing through non-linear activations.**

MLP is implemented using one hidden layer and one layer for fine tuning

The higher layers of ViT learn the global features, whereas the lower layers learn both global and local features. This allows ViT to learn more generic patterns.

## **A combination of CNN and Visual Transformer**

