Literature review
A ConvNet for the 2020s

Introduction of the Transformer model into the domain of computer vision in the form of ViT has surpassed the CNN model in image classification tasks replacing it as the state-of-the-art model for image classification. The ViT however faced difficulties in certain computer vision tasks such as object detections and image segmentation.

These difficulties are addressed in the SWIN Transformer model which uses some of the techniques used in CNNs.
For example, the ViT model has a global attention layer while the SWIN Transformer model divides the tokens into smaller “windows" and performs a local self attention function on these windows and uses the mechanism of shifting window for the tokens to get information from other tokens not in their local window. This mechanism is very similar to how kernels work in CNNs. They operate on the fact that for images, pixels which are close together, have stronger dependencies than pixels far apart from each other.

This paper introduces a pure CNN model called the ConvNeXt. Just like how SWIN Transformer takes inspiration from CNNs and applies CNN techniques to Transformers, ConvNeXt uses techniques used in Transformers and applies them on a ResNet model.

This paper uses all kinds of approaches used in transformer models and uses them in the CNN model.
The following approaches are picked from Transformer models and used in ConvNeXt:

- Using the AdamW optimizer and image augmentation techniques
- Using the same stage compute ratio and using patchify stem.
  This is an interesting way of selecting layers of the model. Transformers and CNNs don’t have much similarity but the compute ratio used in transformers is used in CNNs and this results in greater accuracy. Same is the case with the downsampling method.
- Grouped convolutions (inspired by ResNeXt) are used in the form of depthwise convolutions which is a special case. This depthwise convolution is similar to the weighted sum operation of the self attention layer. Both of these operate on a per channel basis and the use of depthwise convolutions reduce the model FLOPs. The number of channels is also increased to match the number of channels in SWIN Transformer.

- Inverse bottleneck layers are an important part of transformer designs and so are used.
- The transformer model is known for its non local attention block. Swin transformer uses a local attention block but it's window size (7x7) is still much larger than the gold standard of CNNs (3x3). For this reason a larger kernel size is used in ConvNeXt.
- In Transformers, the self attention layer is placed before the feed forward layer. Since we have an inverse bottleneck introduced in the CNN model, we move up the depthwise convolution layer to match the transformer model. This is done because as mentioned before, depthwise convolution layers are analogous to the self attention layers. This results in a degradation of the model performance which the authors of the paper call "temporary".
- ReLU and Batch Normalisation layers used in CNNs have been replaced by GELU and Layer Normalisation layers. These have also been used less frequently.
  This is a surprise as it is a convention to use a Convolution layer, Batch Normalisation layer and ReLU layer together all the time.

A point to note is that each of these changes applied have all individually been tested before. This paper is the first that combines all the approaches and achieves a model which outperforms the SWIN transformer on all CV benchmarks while having similar FLOPs.

This ConvNeXt model, for classification tasks, is tested on imagenet dataset, with pre-training on imagenet-22k and fine tuning on imagenet-1k dataset. In this dataset it outperforms the Swin ViT model.

For object detection and segmentation, it is tested on the COCO dataset. Here again it outperforms the Swin ViT model.

Semantic segmentation is tested on the ADE20K dataset where it achieves competitive performance.

This paper notes the accuracy of the model after each change made. This shows how each specific technique used in transformers can be used in a regular CNN. The same way how SWIN transformers used techniques of CNNs and achieved a better result than ViTs and CNNs, ConvNeXt does the same thing and achieves better accuracy than the SWIN ViT.

Moving the depthwise convolution layer upwards reduces the model accuracy. This leaves the possibility that not making this specific change can improve the overall accuracy of the model. The goal of this paper is to prove that CNNs are still relevant in the age of Transformers and **NOT** to make it similar to transformers. It is

understood that this change is made to use Transformer techniques in CNNs but in this case it is going against the goal of the paper.
There is also the possibility that not making this particular change results in a lower overall accuracy. If that is the case, the paper does not mention it.

This paper focuses mostly on classification tasks and not as much on other computer vision tasks. Several advances have been made on transformers and models like Visformer are very strong at object detection and segmentation.

CNNs have many advantages over Transformer models. They are built on the basic principles of image processing(convolutions) and are much easier to fine tune than transformers but transformers are evolving very fast.

I believe that to get an overall better performance in general computer vision tasks, more CNN techniques should be used in transformers (Convit) as transformers have lots of potential as it is a very new model. SWIN ViT uses self attention while some other models have been proposed where cross attention is used. Some new transformer models also use Fast Fourier Transforms for token mixing.

Considering that the transformer model works in a very different way as compared to CNNs, they are not direct competitors in the exact same field. The paper recognises this fact and mentions that the architecture should be decided based on the requirements.

This paper also shows that even though transformers are becoming a dominant model in the field of computer vision, CNNs are still relevant.