

The transfer learning process is also commonly called fine-tuning because the model is tuned and repurposed to solve a different problem.

The next slide shows an example of how the source and target datasets are used in

this process.

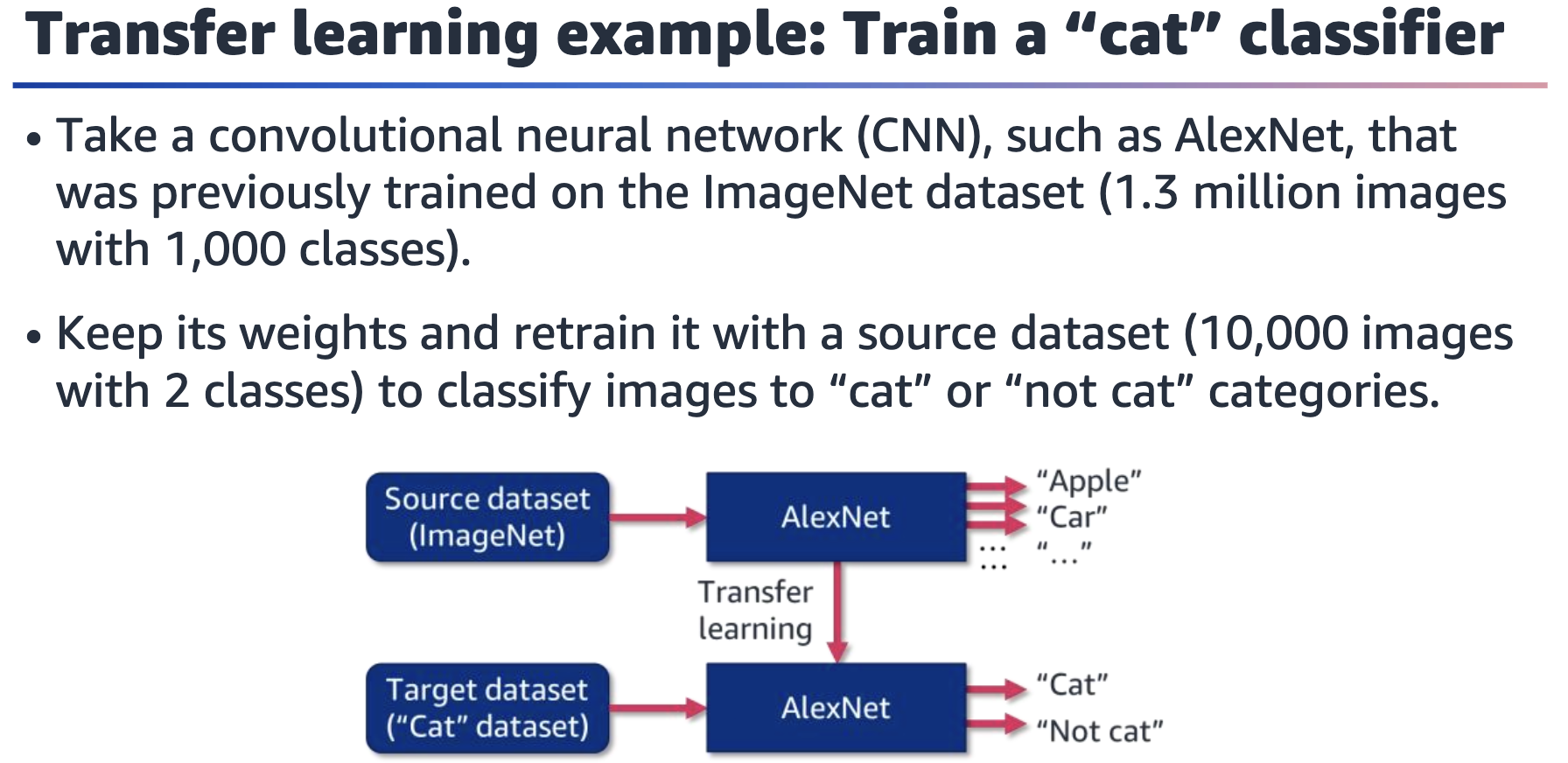


Figure of transfer learning. First, an AlexNet model is trained with the source dataset. This creates a pretrained model. Then, this pretrained model is trained on the target dataset.

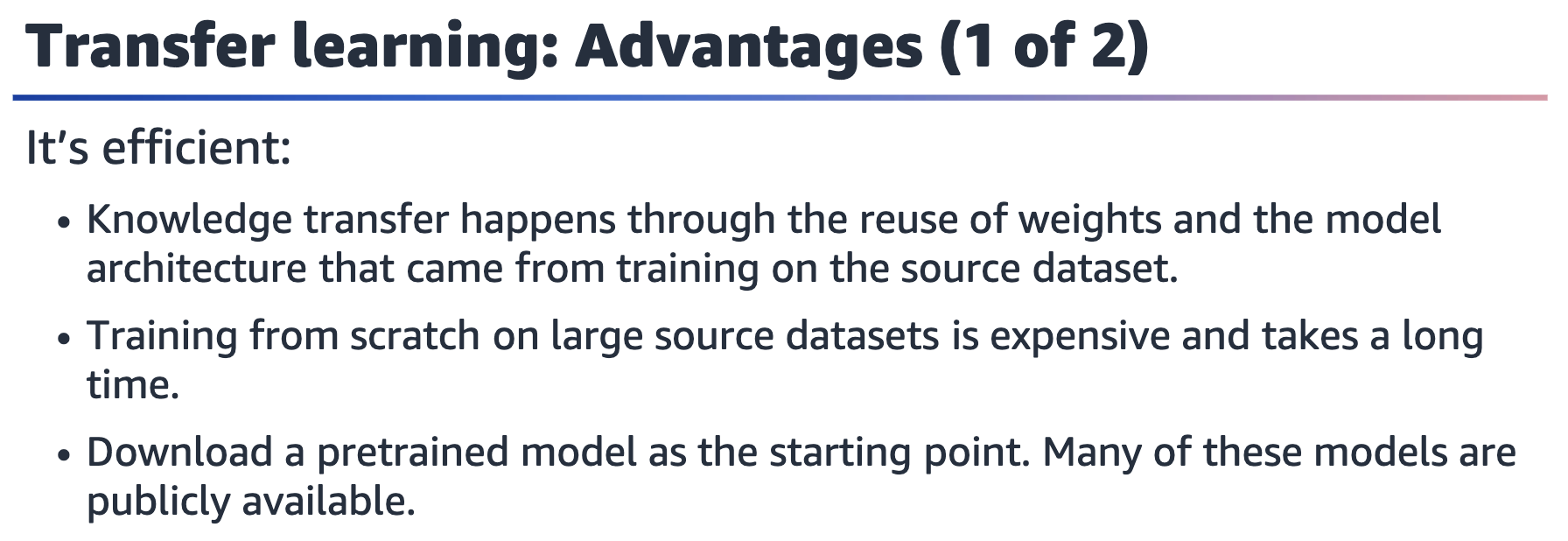
This example of transfer learning uses an AlexNet model that was previously trained

on the ImageNet dataset and retrains it on a target dataset. The source dataset has

1,000 classes, whereas the target dataset has 2. The model’s output layer is changed

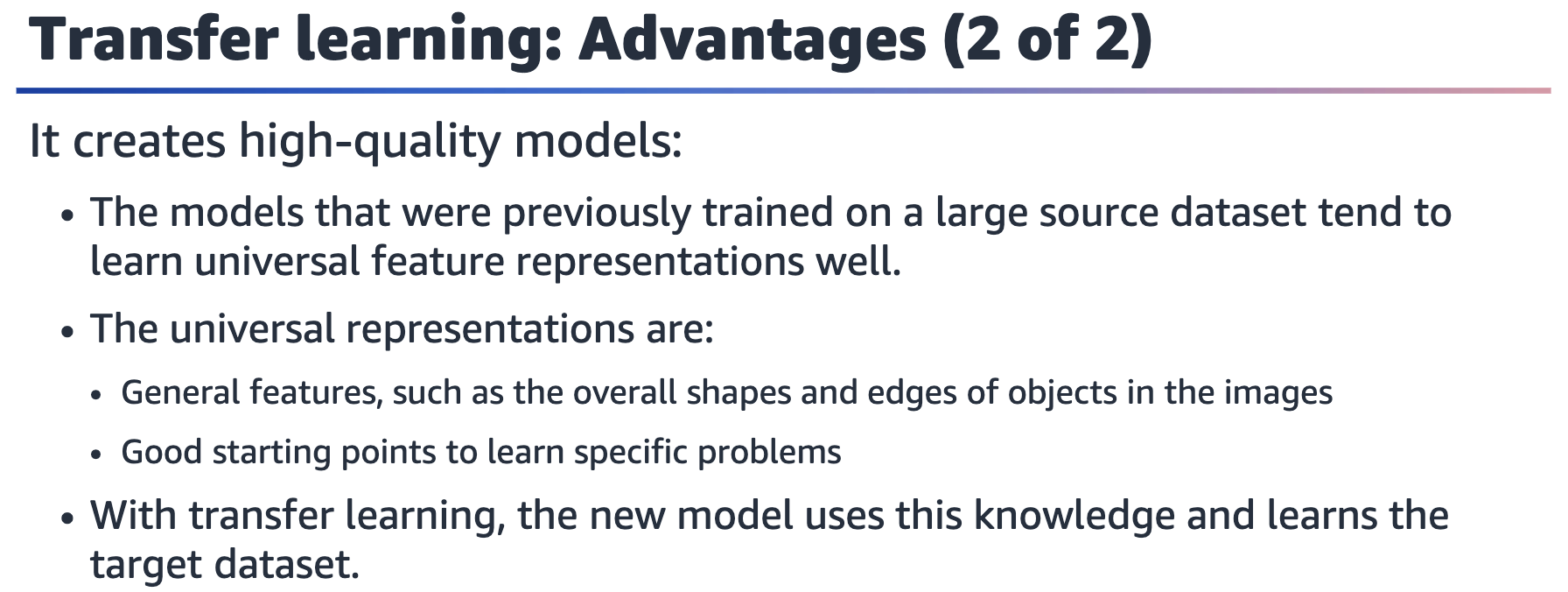
to have a shape of two to make this work.

The steps in transfer learning will be described in more detail later in this lesson.



Transfer learning is efficient. The original weights that are learned from the source

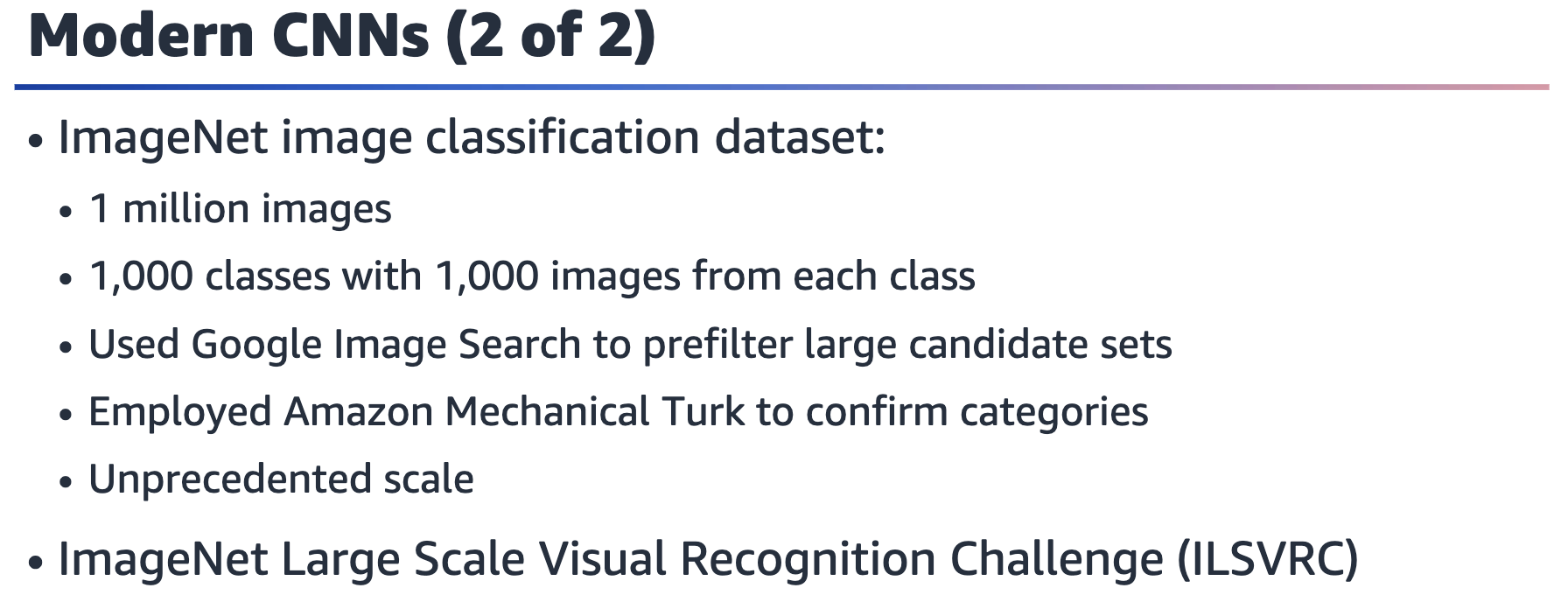
dataset are used as a starting point.



Transfer learning produces high-quality models. The models that are trained on large

source datasets learn universal feature representations well. These models learn how

to handle many images from many categories. These learned universal representations are passed to the model that is trained with the target dataset through the use of the weights.



The ImageNet dataset is a large-scale image classification dataset that has been widely used in computer vision (CV) research. The dataset consists of more than 1 million labelled images, and each image belongs to one of 1,000 object categories.

Researchers at Princeton University and Stanford University created the dataset in

2009. The researchers, led by Fei-Fei Li, used Google Image Search to prefilter large

candidate sets for each category. They also employed the Amazon Mechanical Turk

crowdsourcing pipeline to confirm whether each image belonged to the associated

category. The scale of this process was unprecedented at the time, and it required

significant computational resources to process and store the large dataset.

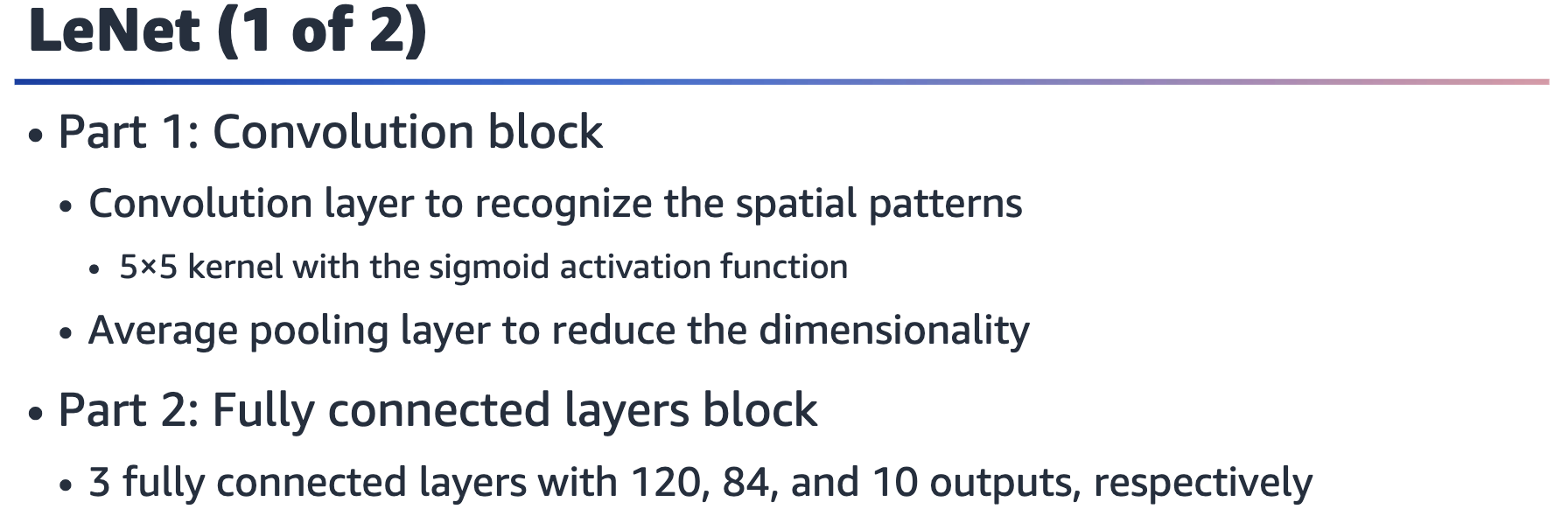
The ImageNet dataset has played a crucial role in advancing the field of CV, particularly in the development of deep learning algorithms, such as CNNs. In 2012, a team of researchers from the University of Toronto used a CNN architecture called AlexNet to achieve a significant breakthrough in image recognition accuracy. The team won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with a top-5 error rate of 15.3 percent, which was significantly better than the next best algorithm at the time.

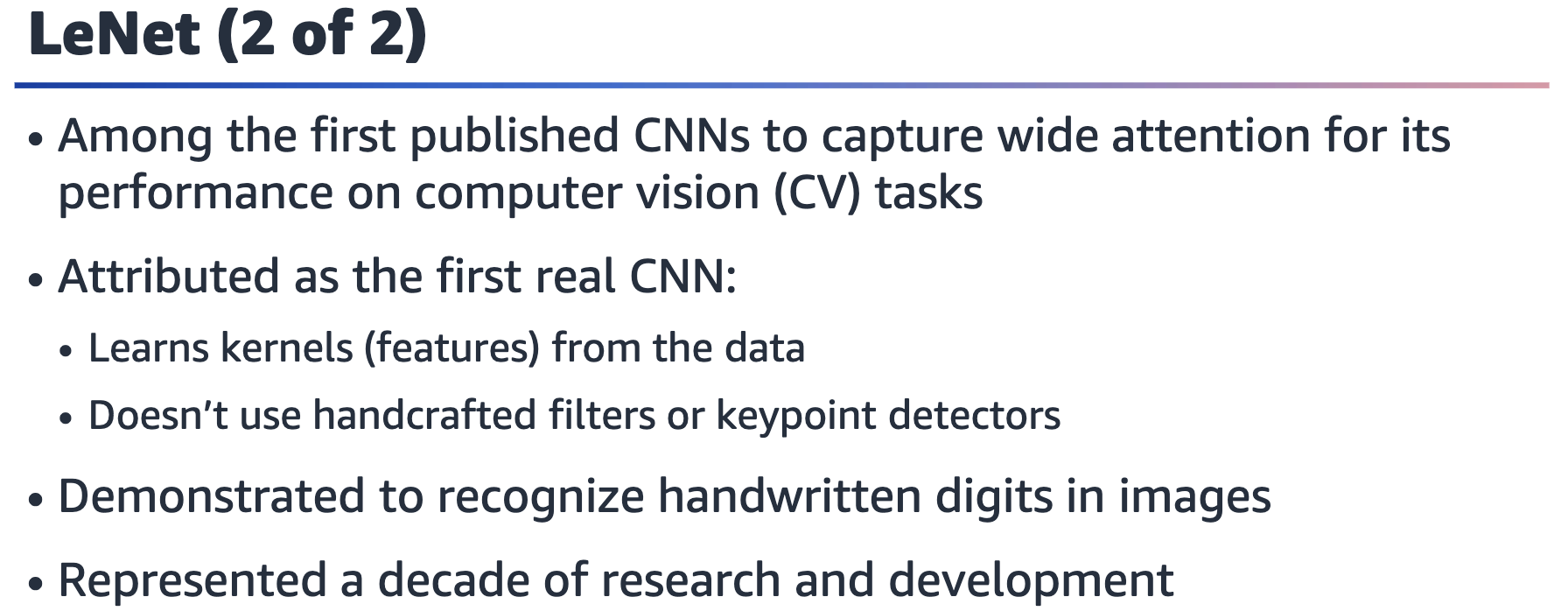
Since then, the ImageNet dataset has been used as a benchmark to evaluate the

performance of image-recognition algorithms. The dataset has been expanded to include other tasks, such as object detection, segmentation, and localization. The dataset and the ILSVRC challenge have also inspired the creation of other large-scale

visual recognition datasets and challenges, such as Common Objects in Context (COCO), PASCAL Visual Object Classes (VOC), and Open Images.







LeNet was among the first published CNNs to capture wide attention for its performance on CV tasks. The model was introduced by (and named for) Yann LeCun, a researcher at AT&T Bell Labs at the time. The model was used to recognize handwritten digits in images. This work represented the culmination of a decade of research to develop the technology.

In 1989, LeCun published the first study to successfully train CNNs through

backpropagation.

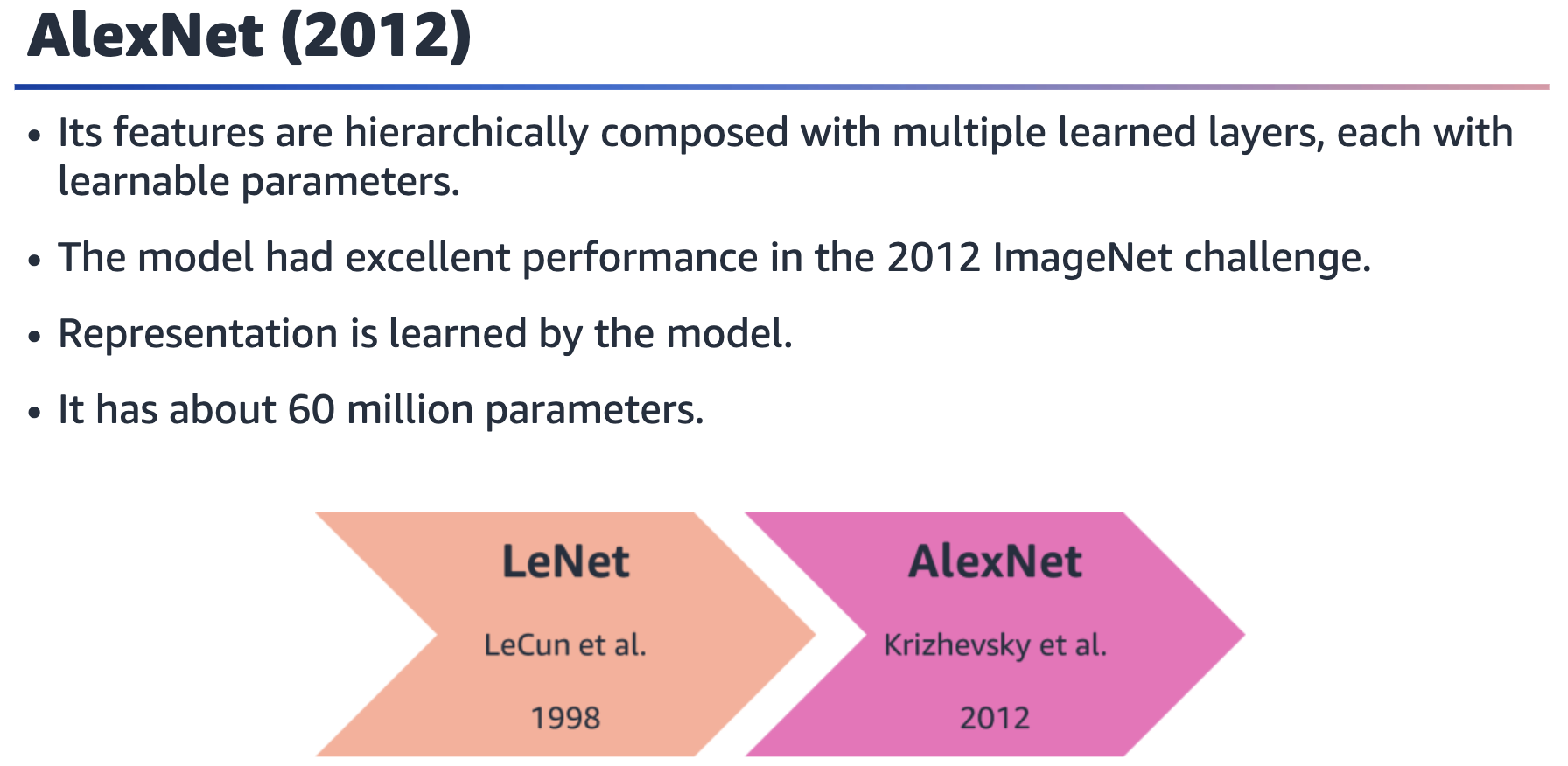
References:

• Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D.

Jackel. “Backpropagation Applied to Handwritten Zip Code Recognition.” Neural Computation 1, issue 4. (December 1989). <https://doi.org/10.1162/neco.1989.1.4.541>.

• Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. “Gradient-Based Learning Applied to Document Recognition.” Proceedings of the IEEE 86, issue 11. (November 1998). <https://doi.org/10.1109/5.726791>.





AlexNet is a deeper CNN model that Alex Krizhevsky, Ilya Sutskever, and Geoffrey

Hinton developed in 2012. It has a much deeper architecture with eight layers, including five convolutional layers and three fully connected layers.

In 2012, AlexNet became the first CNN model to win the ILSVRC, achieving a significant breakthrough in image-recognition accuracy. AlexNet was the only model to achieve an error rate of less than 25 percent. Following the competition, many researchers followed the deep CNN idea and have developed better-performing models year after year.

AlexNet also introduced several novel techniques that weren’t used in LeNet. These

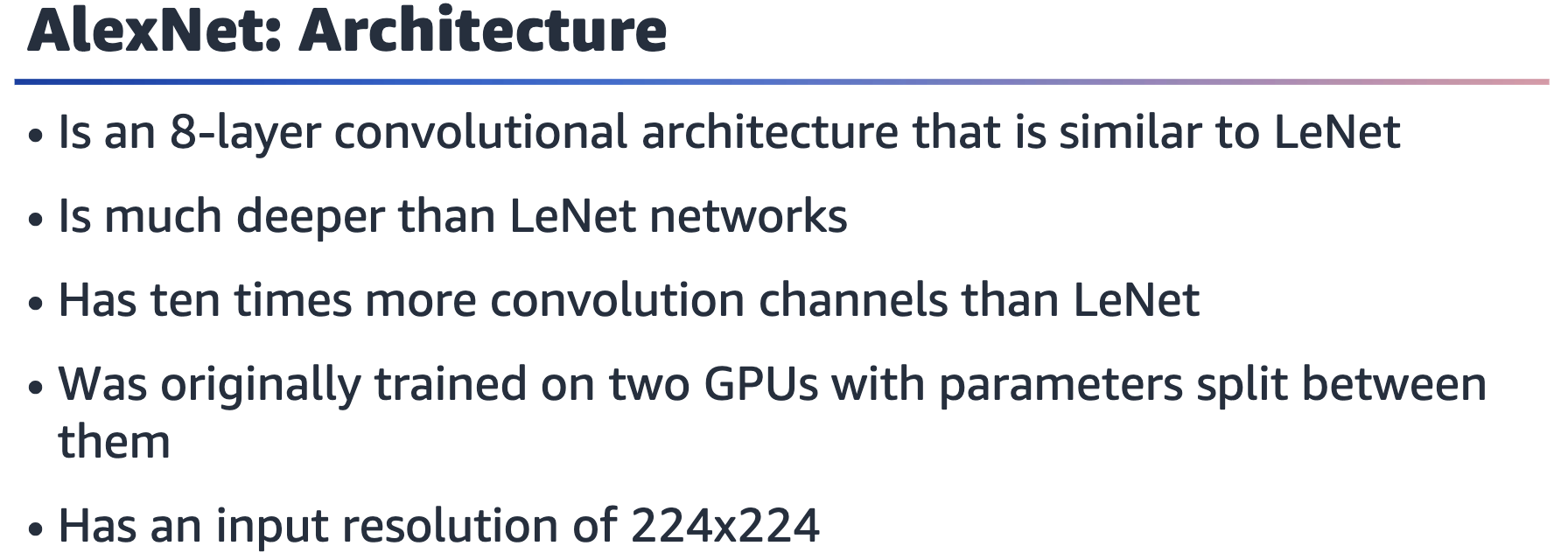
include the use of rectified linear unit (ReLU) activation functions, overlapping pooling, and data augmentation. These techniques, along with the larger architecture and dataset, contributed to the significant improvement in performance that AlexNet achieved.

LeNet and AlexNet were both significant contributions to the development of deep

learning and CV. LeNet was an early pioneering work in the field of CNNs, while

AlexNet represented a significant breakthrough in the field, with its larger

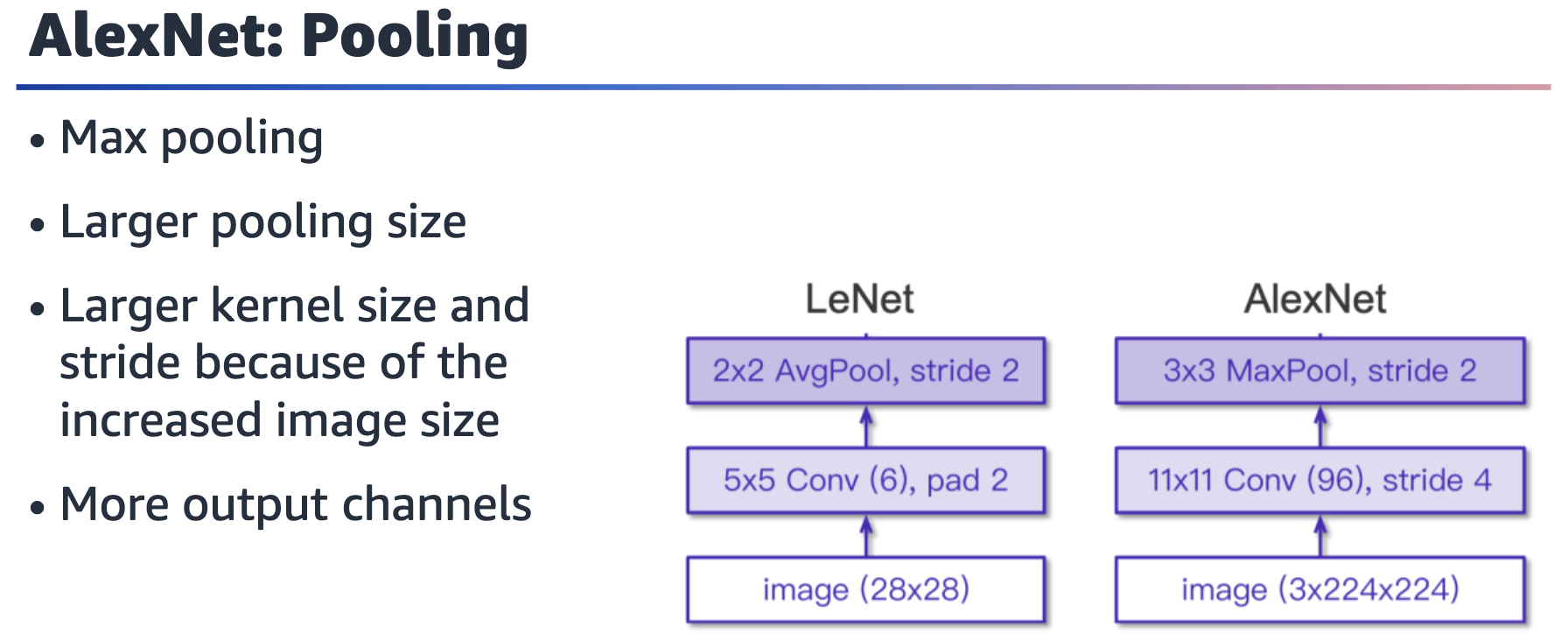
architecture, innovative techniques, and use of a much larger dataset.



For more information about the architecture of the AlexNet model, see “Deep

Convolutional Neural Networks (AlexNet)” in Dive into Deep Learning at

<https://d2l.ai/chapter_convolutional-modern/alexnet.html>.



Compared to the LeNet architecture, which was designed to recognize small, handwritten digits, AlexNet was designed to handle much larger and more complex

images. As a result, several changes were made to the architecture to improve its

performance on this task.

One of the main changes was the use of max pooling instead of average pooling. Max

pooling is a form of nonlinear downsampling that helps to reduce the spatial dimensions of the output feature maps. This helps to reduce the number of parameters in the model and can also help to prevent overfitting. Additionally, larger pooling sizes, such as 3x3, were used in AlexNet compared to LeNet, which used smaller pooling sizes, such as 2x2.

Another change was the use of larger kernel sizes, such as 11x11, in the first convolutional layer. This allowed the model to capture larger-scale features in the

input images. In addition, the use of a stride of 4 in the first convolutional layer

helped to reduce the spatial dimensions of the output feature maps, which reduced

the computational cost of the subsequent layers.

AlexNet also had more output channels than LeNet, which allowed it to learn more

complex representations of the input images.

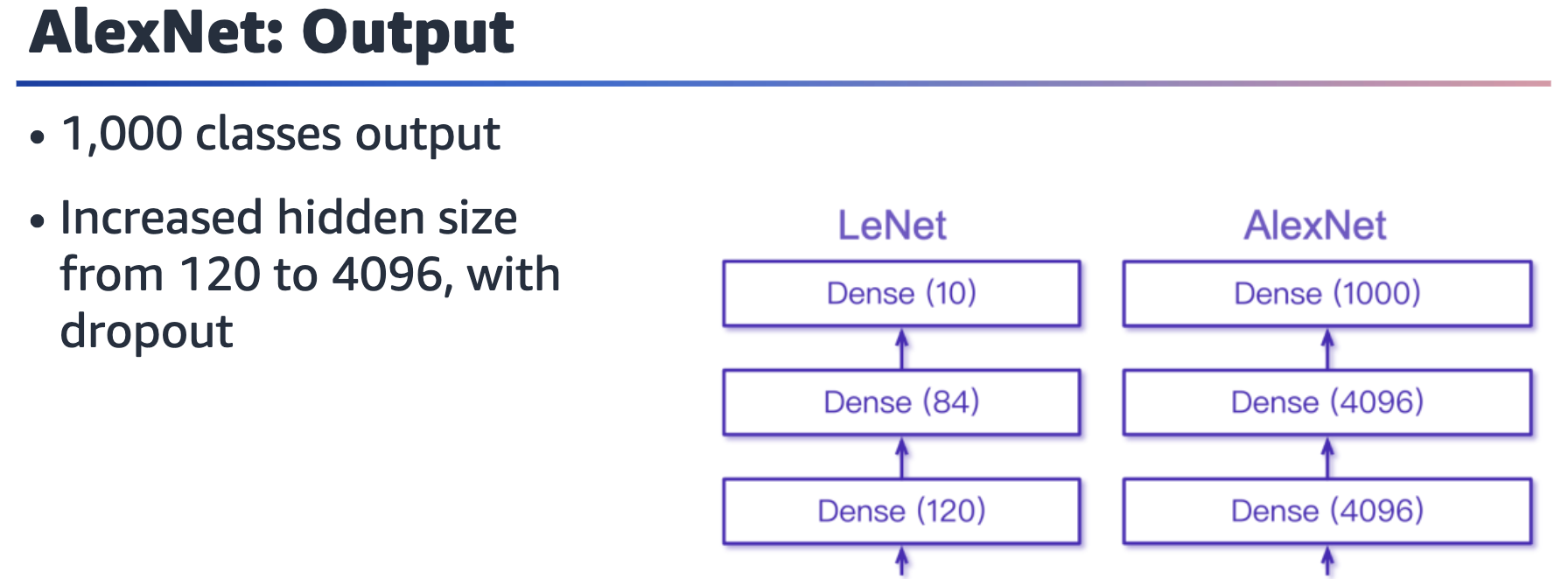
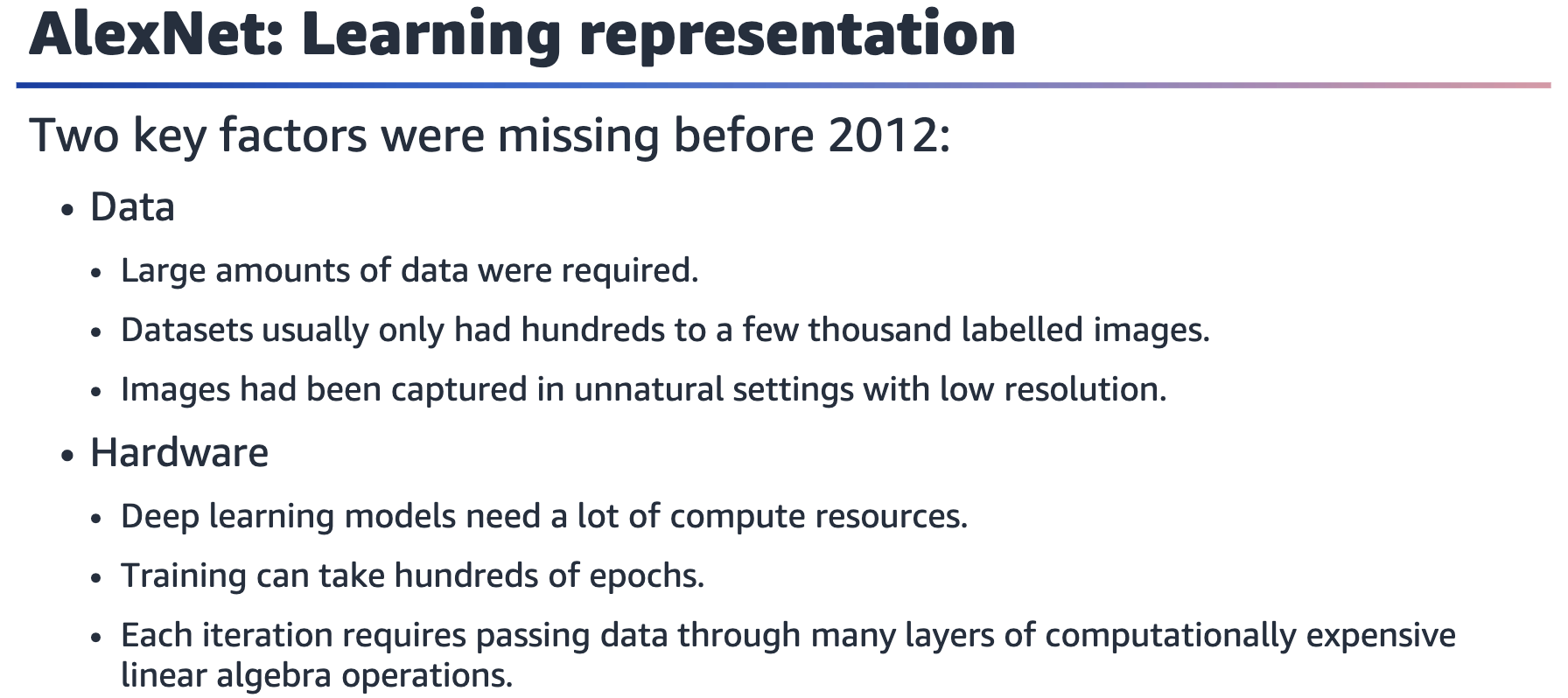


Diagram showing the fully connected layers of LeNet and AlexNet. LeNet has 3 fully connected layers with 120, 84, and 10 neurons for each layer. AlexNet has 3 fully connected layers with 4096, 4096, and 1000 neurons for each layer.

AlexNet was trained for 1,000 classes, while LeNet was trained for 10 classes. AlexNet

also used more units in the dense layers.



The success of AlexNet was because of two key factors that were missing before

2012: data and hardware.

The availability of large amounts of data was a critical factor in the success of AlexNet. Prior to 2012, few large-scale image datasets were available to train deep neural networks. The ImageNet dataset, which was used to train AlexNet, contained over 1 million images with 1,000 object categories. This provided a much larger and more diverse training set than what had been available previously. By using the larger dataset, the model was able to learn more representative and robust features, which

contributed to its improved performance.

Hardware was another critical factor in the success of AlexNet. Deep learning models

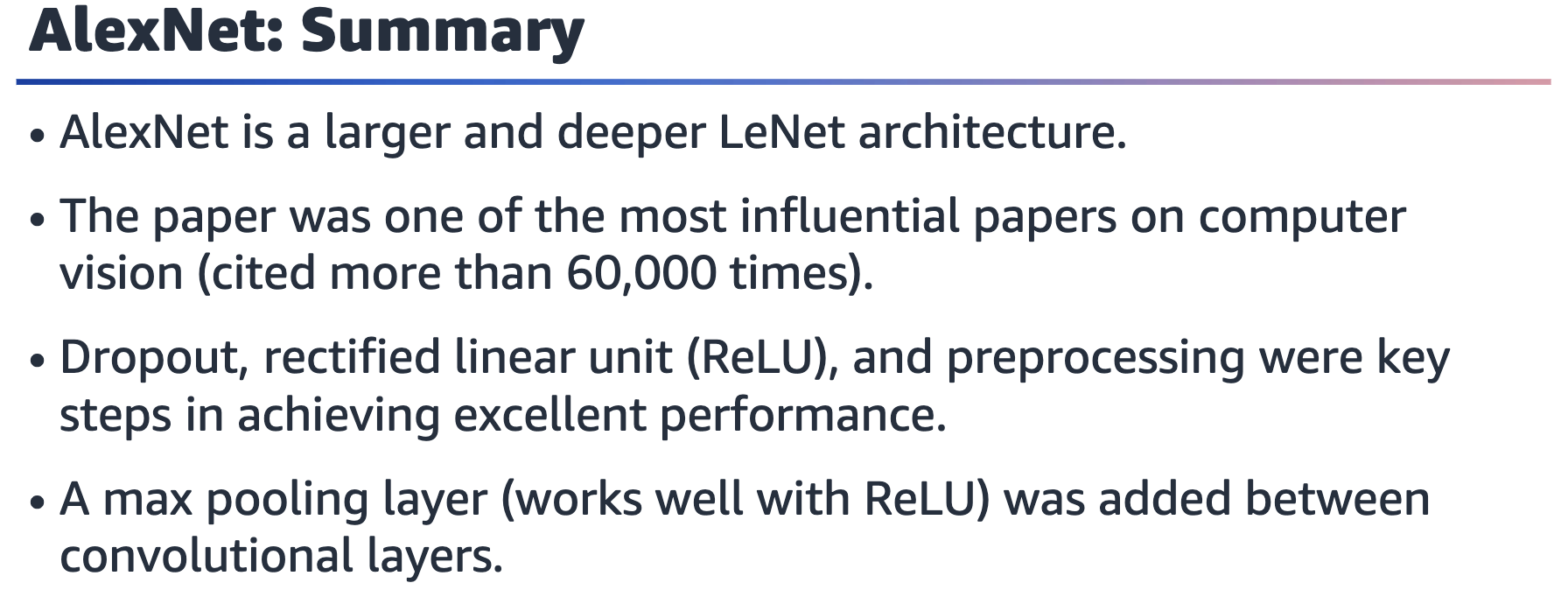
require a significant amount of computational resources. Training can take hundreds

of epochs, with each iteration requiring data to be passed through many layers of

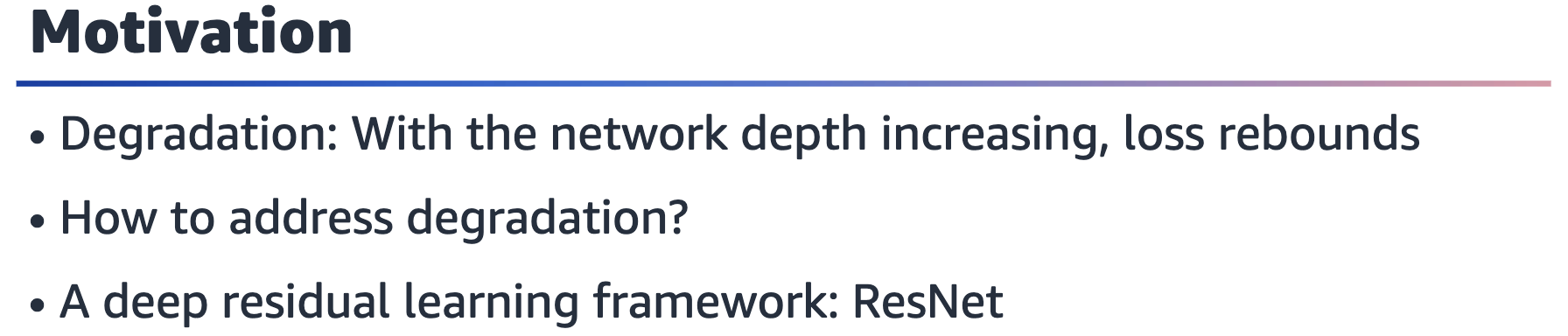
computationally expensive linear algebra operations. The development of graphical

processing units (GPUs) provided a significant increase in compute power, which

enabled faster and more efficient training of deep neural networks. AlexNet was trained on two GPUs, which allowed for the training time to be reduced significantly compared to previous deep learning models.







The degradation problem is believed to be caused by the vanishing gradient problem,

where the gradients become too small to effectively update the parameters of the

lower layers. This can make it difficult for the lower layers to learn useful representations and can ultimately result in a decrease in performance.

ResNet addresses this problem by introducing residual skip connections, which allow

information to be directly passed between layers, bypassing some of the layers

altogether. This means that even if the gradients become small, the information can

still flow through the network and be used to update the parameters.

The motivation behind ResNet is to create deeper neural networks that can learn

more complex representations without being hindered by the degradation problem.

By introducing residual connections, ResNet has been able to achieve state-of-the-art

results on a wide range of CV tasks.

The Xavier initialization (Glorot et al., 2010) and the He initialization (He et al., 2015)

are some techniques to perform robust weight initialization, which mitigates

degradation in deep neural networks.

.

References:

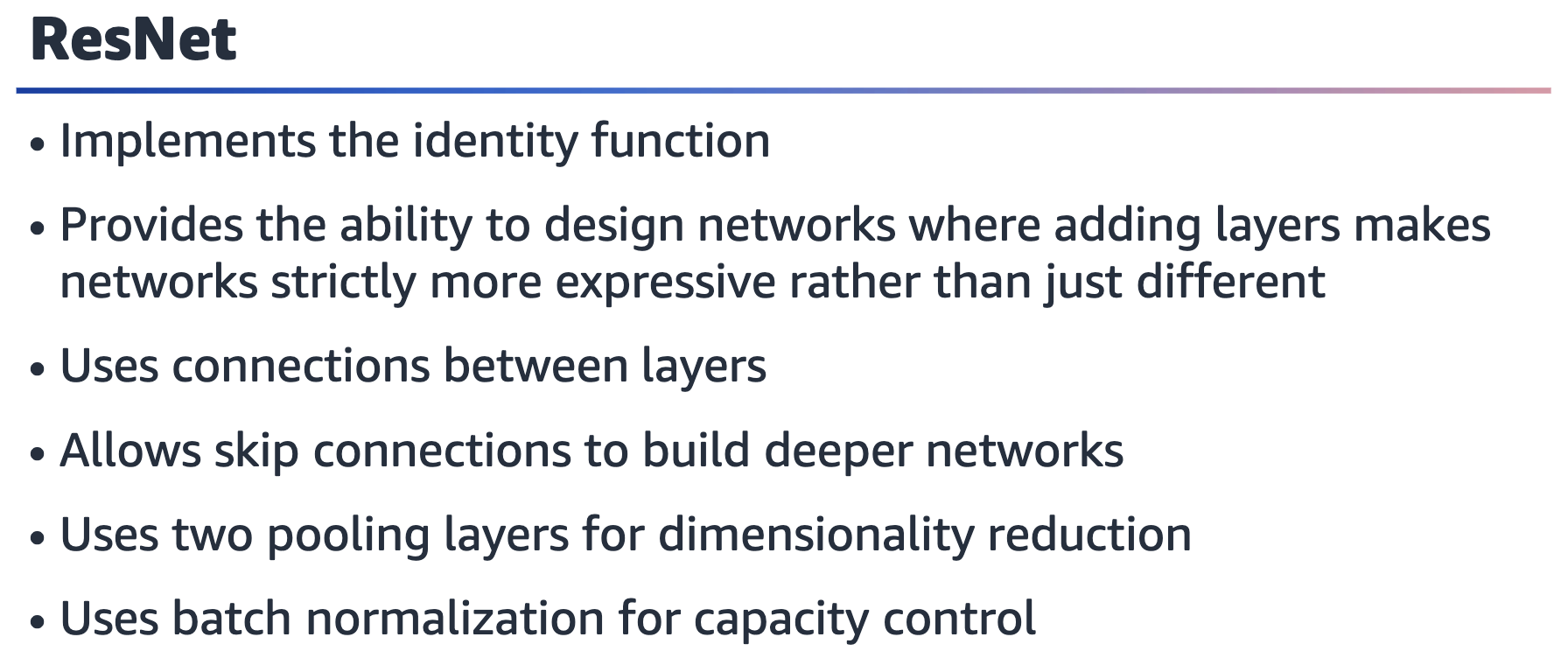
• Xavier Glorot and Yoshua Bengio. “Understanding the Difficulty of Training Deep

Feedforward Neural Networks.” Proceedings of Machine Learning Research 9:

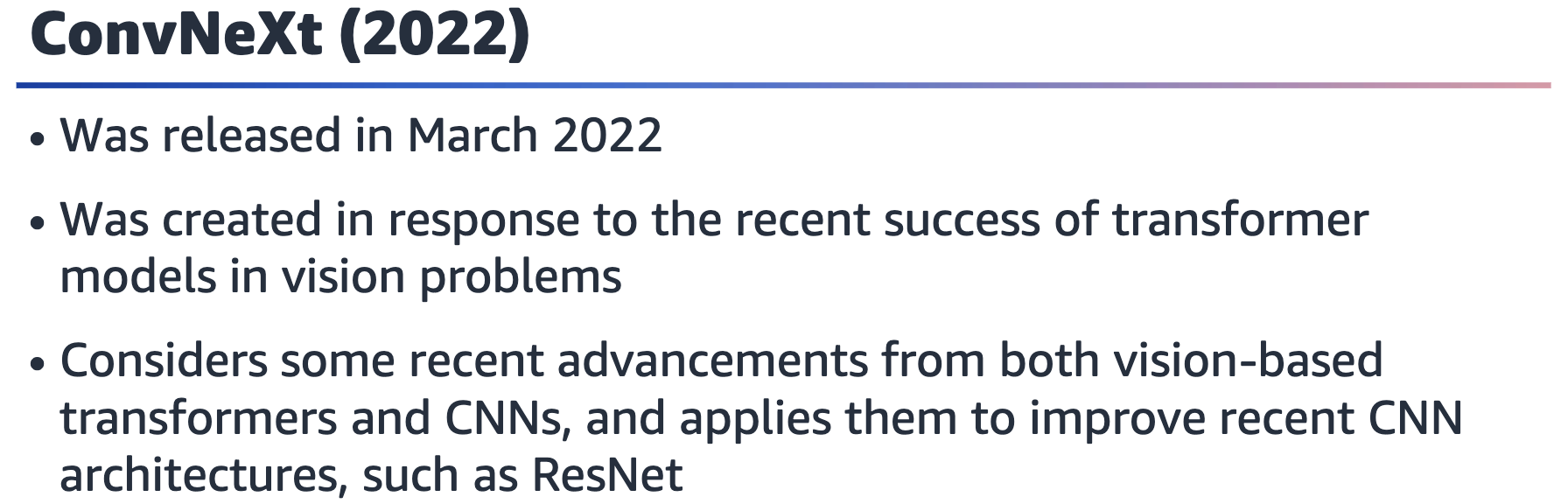
249-256. (2010). <https://proceedings.mlr.press/v9/glorot10a.html>.

• Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Delving Deep into

Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.” Paper presented at the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, December 2015. <https://doi.org/10.1109/ICCV.2015.123>.







Introduced in late 2018, transformers completely revolutionized the natural language

processing (NLP) field. Their success in NLP also made them popular architecture

choices for other areas, including CV. Vision transformers aren’t covered in this class.

ConvNeXt is a popular CNN model that combines ideas from multiple recent developments in vision-based transformers and other recent CNNs. This model isn’t a

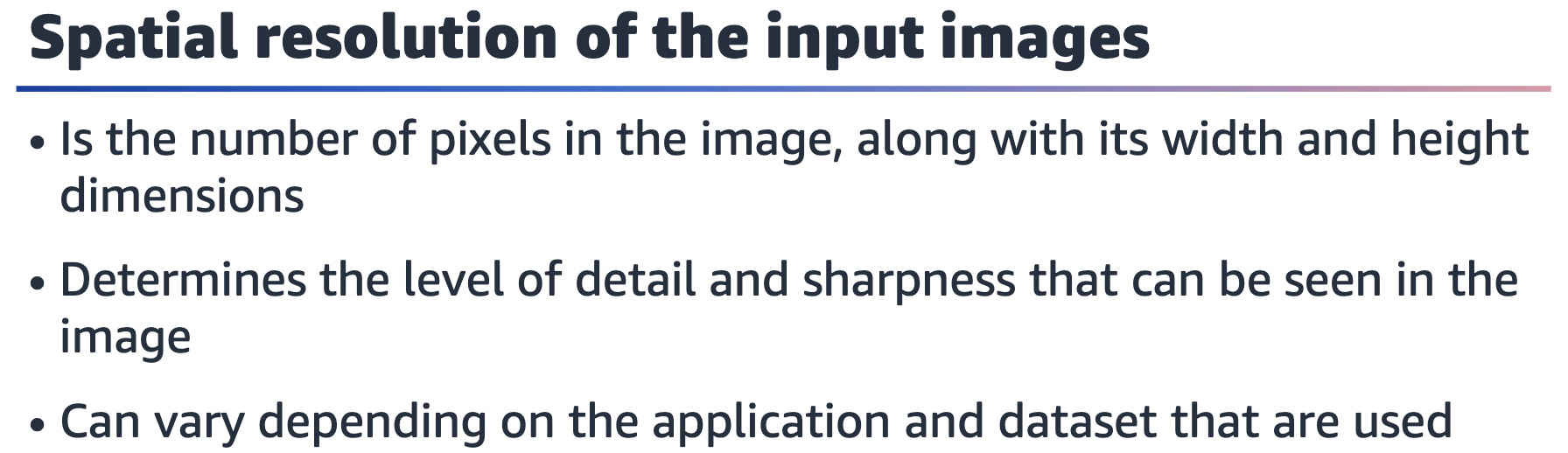
transformer-based model. It uses traditional CNN modules that have been updated based on recent developments in CV.

Reference: Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor

Darrell, and Saining Xie. “A ConvNet for the 2020s.” Paper presented at the 2022

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New

Orleans, LA, June 2022. <https://doi.org/10.1109/CVPR52688.2022.01167>.



Spatial resolution determines the level of detail and sharpness that can be seen in an

image. For example, an image with a spatial resolution of 1280x720, which is 1280

pixels wide and 720 pixels high, has more pixels and higher resolution than an image

with a spatial resolution of 640x480, which is 640 pixels wide and 480 pixels high.

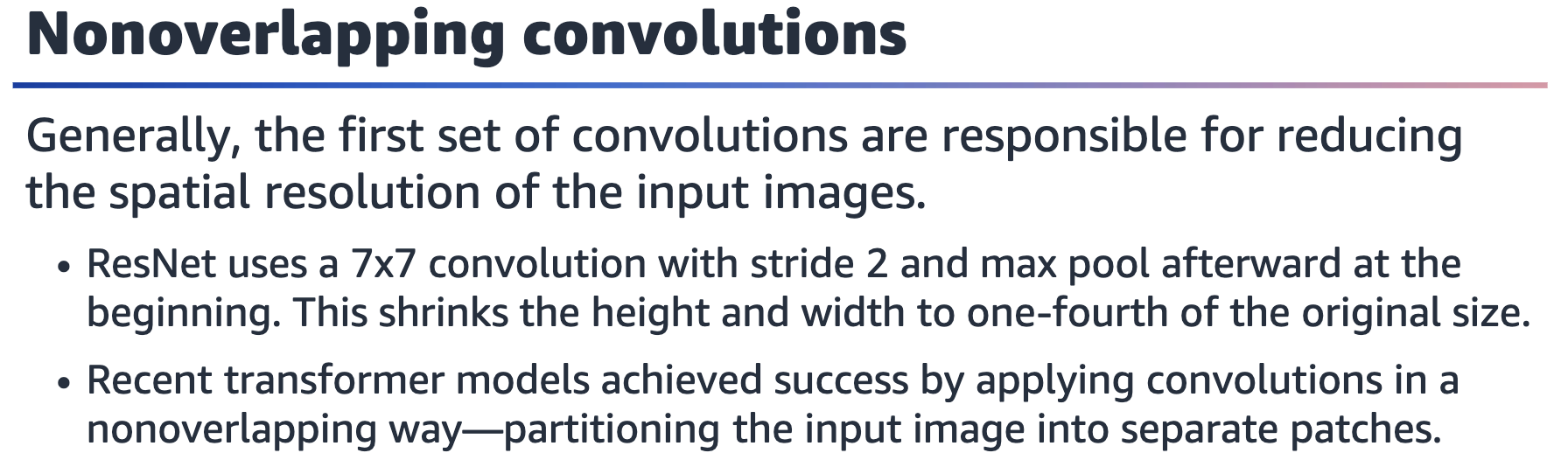
The spatial resolution of input images can vary depending on the application and

dataset that are used. For example, in the case of image classification tasks that use

the CIFAR-10 dataset, the input images have a spatial resolution of 32x32 pixels.

However, in the case of image classification tasks that use the ImageNet dataset, the

input images have a spatial resolution of 224x224 pixels.



ConvNeXt uses nonoverlapping convolutions to reduce the spatial resolution of

feature maps.

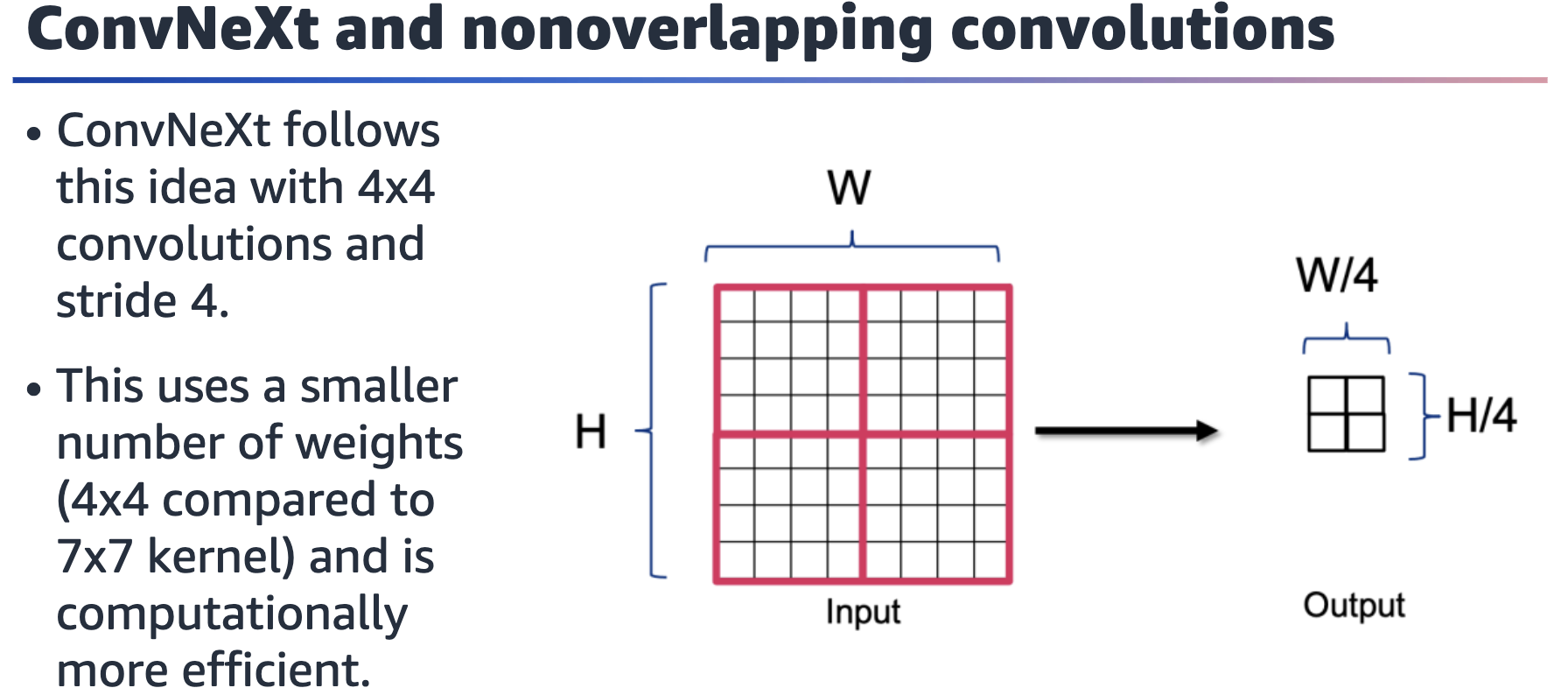


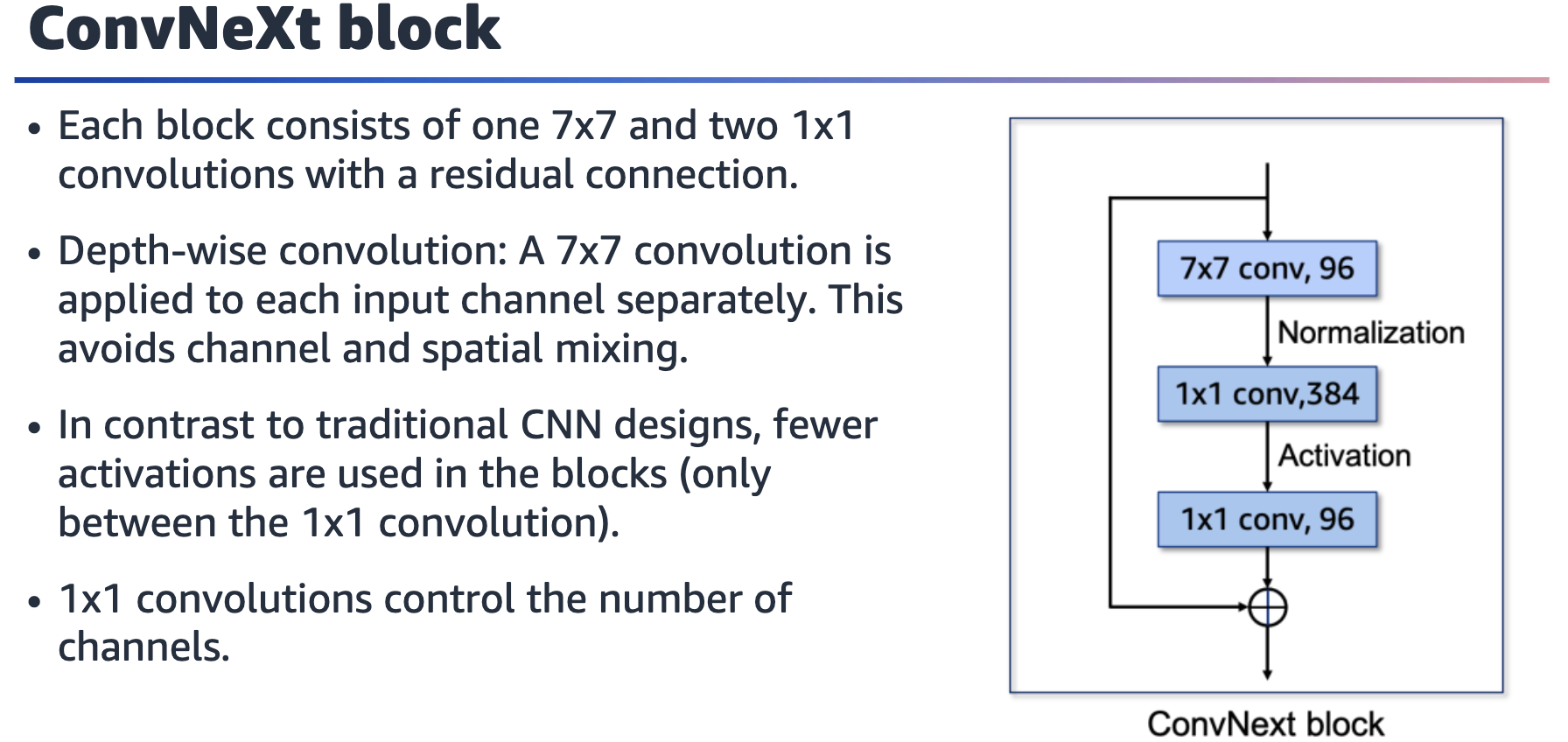
Figure showing nonoverlapping convolution. An 8x8 input turns into a 2x2 output after the convolution. A 4X4 convolution with stride 4 reduces the output to one-fourth of the original size.

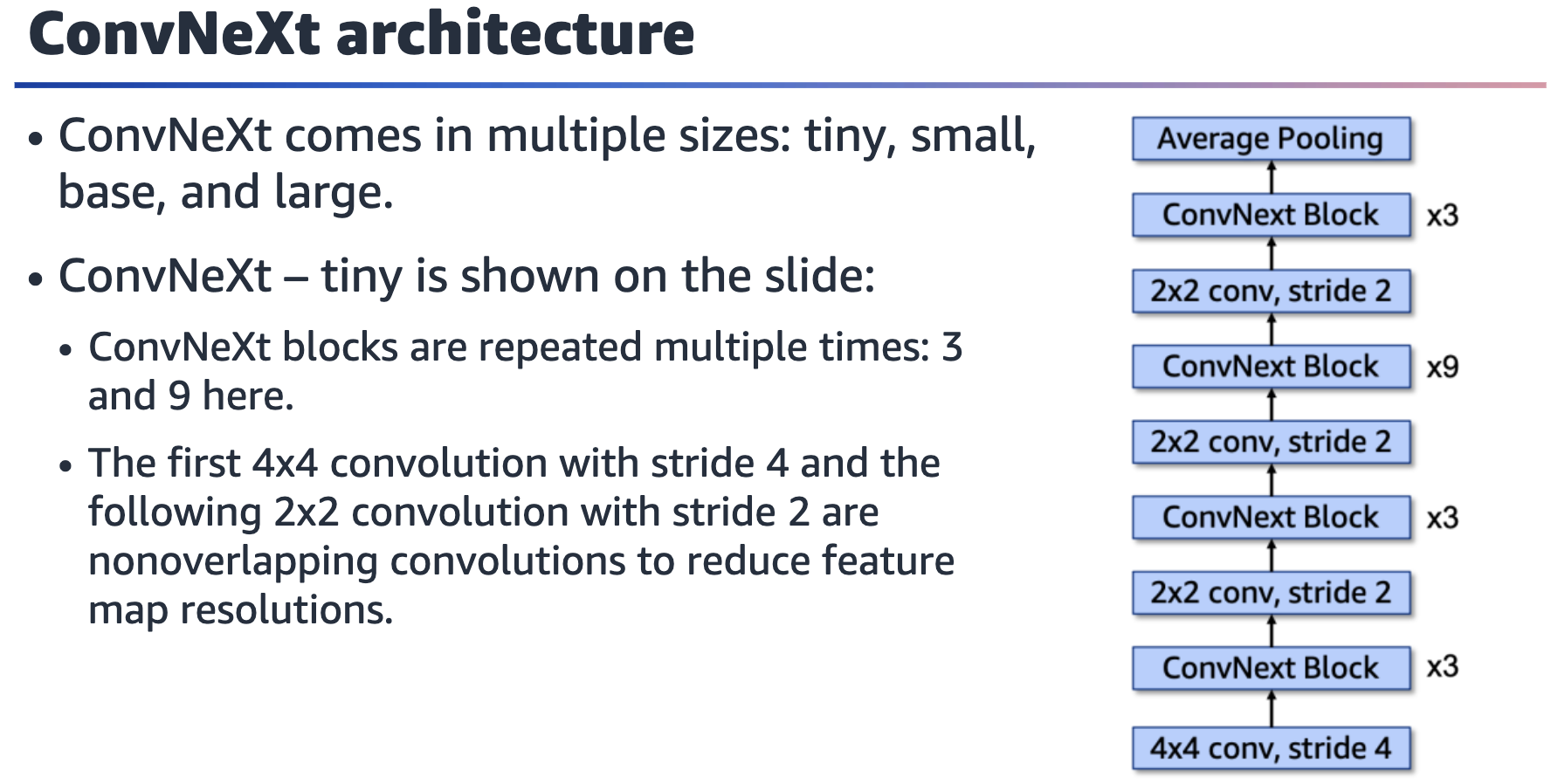
Benefits of the nonoverlapping convolution:

• 4x4 saves some weights compared to 7x7 convolutions of ResNet.

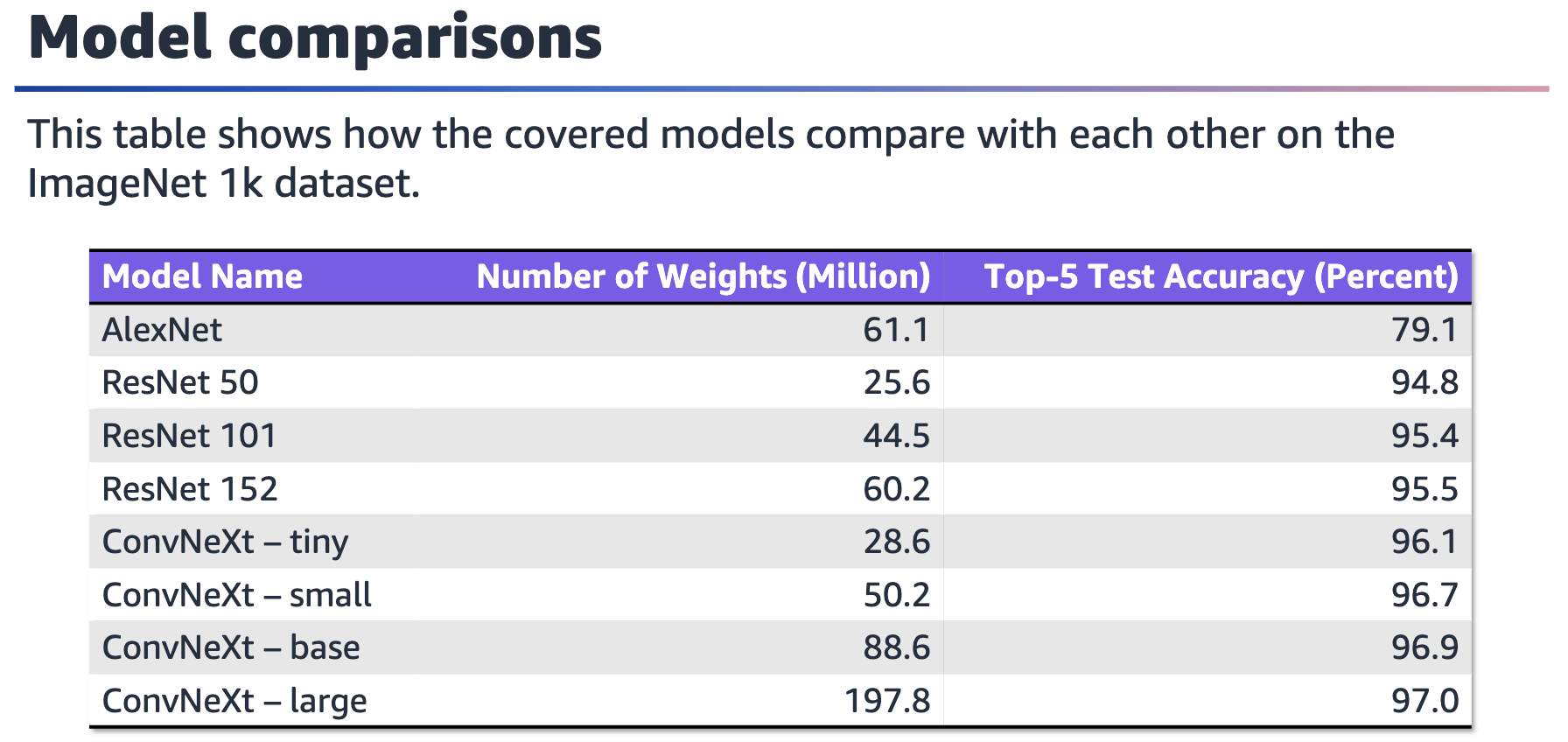
• Stride 4 is more efficient than stride 2, and the convolution operation is applied

fewer times.





A ConvNeXt model has multiple blocks connected in series. The layers from start to finish are a 4x4 convolution layer, a ConvNeXt block repeated 3 times, a 2x2 convolution layer, a ConvNeXt block repeated 3 times, a 2x2 convolution layer, a ConvNeXt block repeated 9 times, a 2x2 convolution layer, a ConvNeXt block repeated 3 times, and an average pooling layer.



Let’s compare the models that have been covered so far. From one architecture to

the next, a lot of progress occurs in terms of the reduction in the number of weights

that are used to achieve the same or better performance.

The top-5 test accuracy scores are reported on the ImageNet 1K dataset (test split).

Top-5 accuracy is measured by comparing the labels of the images with the top 5

predictions from the models.

