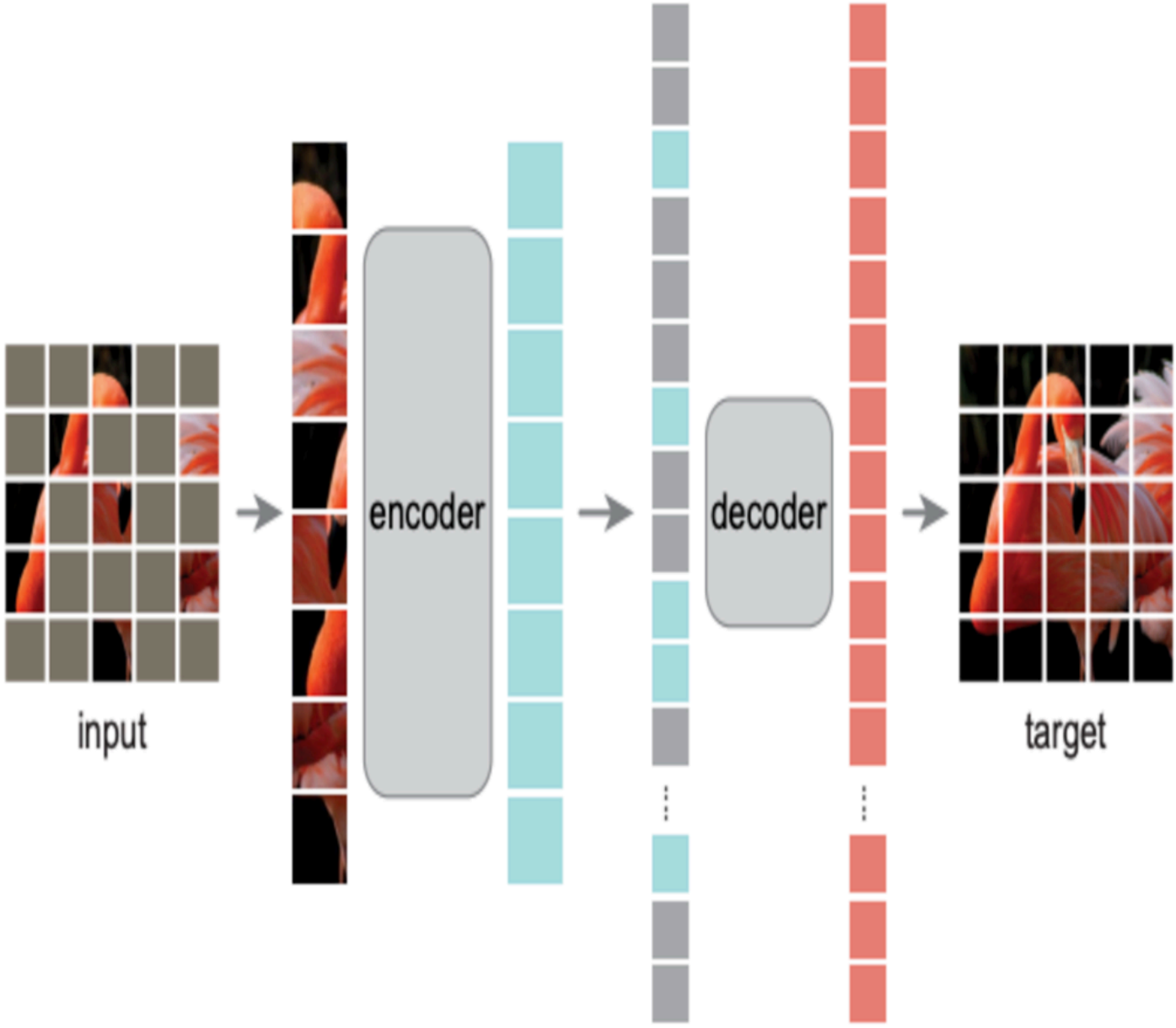
BOLSTER

Blog / Research Labs
Research

VIT-MAE: Scalable Learning for Vision Transformers

February 13, 2024 | 6 min read

Adithya Singh



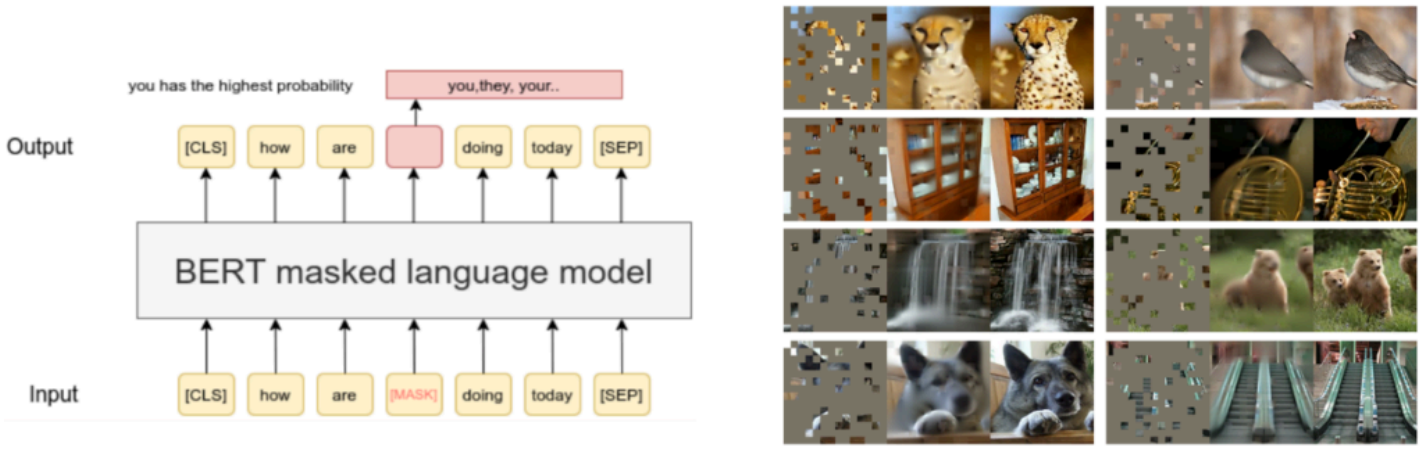
Unprecedented gains in hardware capabilities and model sizes have compelled researchers to devise scalable learning algorithms. State of the art models today have the ability to overfit even on a dataset of a million images.

This trend indicates an increasing data appetite for deep learning models.

Advent of transformer based models like BERT and GPT have posed a highly demanding requirement of data. Transformers which were initially implemented for the NLP tasks have been able to address the data need successfully with self-supervised learning.

Autoregressive language modeling tasks like masked autoencoding in Natural Language Processing have enabled models to learn and generalize, containing even billions of learnable parameters. The idea behind masked autoencoding is to randomly remove some parts of the data before feeding it to the network and training the model in such a way that it can learn to predict the missing parts of the input.

This tasks doesn't need explicit labels, as the learning target is generated from the input itself. This allows utilization of vast amounts of unlabeled data available on the internet.



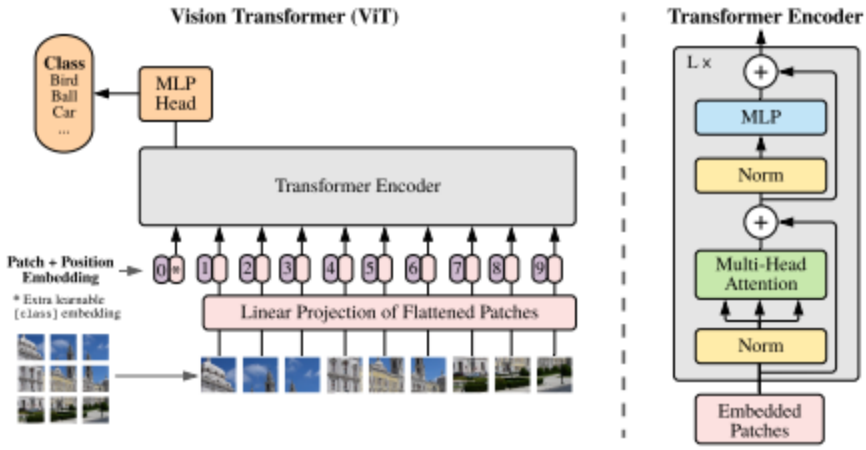
Masked Autoencoding in NLP and Computer Vision

On the side of vision however, things have moved a bit slower compared to NLP when it comes to scalable training. The challenge of translating the masked autoencoding paradigm to computer vision was posed mainly by two obstacles.

The first is that in computer vision, convolutional neural networks have been the defacto problem solver tool for more than a decade. The masked autoencoding task is not directly transferable to convolution operation.

The second challenge is that, it is impossible to train a model in NLP task with very minimal masking of input sequence. This is because of highly dense nature of language data, which forces model to gain a deep understanding of language semantics.

However, this doesn't immediately translate to vision problems, as predicting a missing patch given neighboring patches is trivial and does not require deep understanding of image features, objects.



Vision transformer architecture

The first challenge has largely been tackled by the formulation of **vision transformers**. Vision transformers are extremely good at reconstructing full pixel images from latent representations. Accompanying tools to implement masked autoencoding like tokenization and embedding generation is very straightforward for vision transformers.

The second challenge remains to be solved as it is difficult to translate machine learning training for masked autoencoding that forces models to not settle for high level understanding of the network

As the models grow in size, the challenge of gaining satiate the data requirements for these models is also handled differently for NLP and computer vision tasks.

Labeling of large scale datasets under the domain of computer vision is extremely expensive and challenging. Especially for tasks like object detection and image segmentation which require object level and pixel level an notation of data, it is a considerable burden on both time and effort resources.

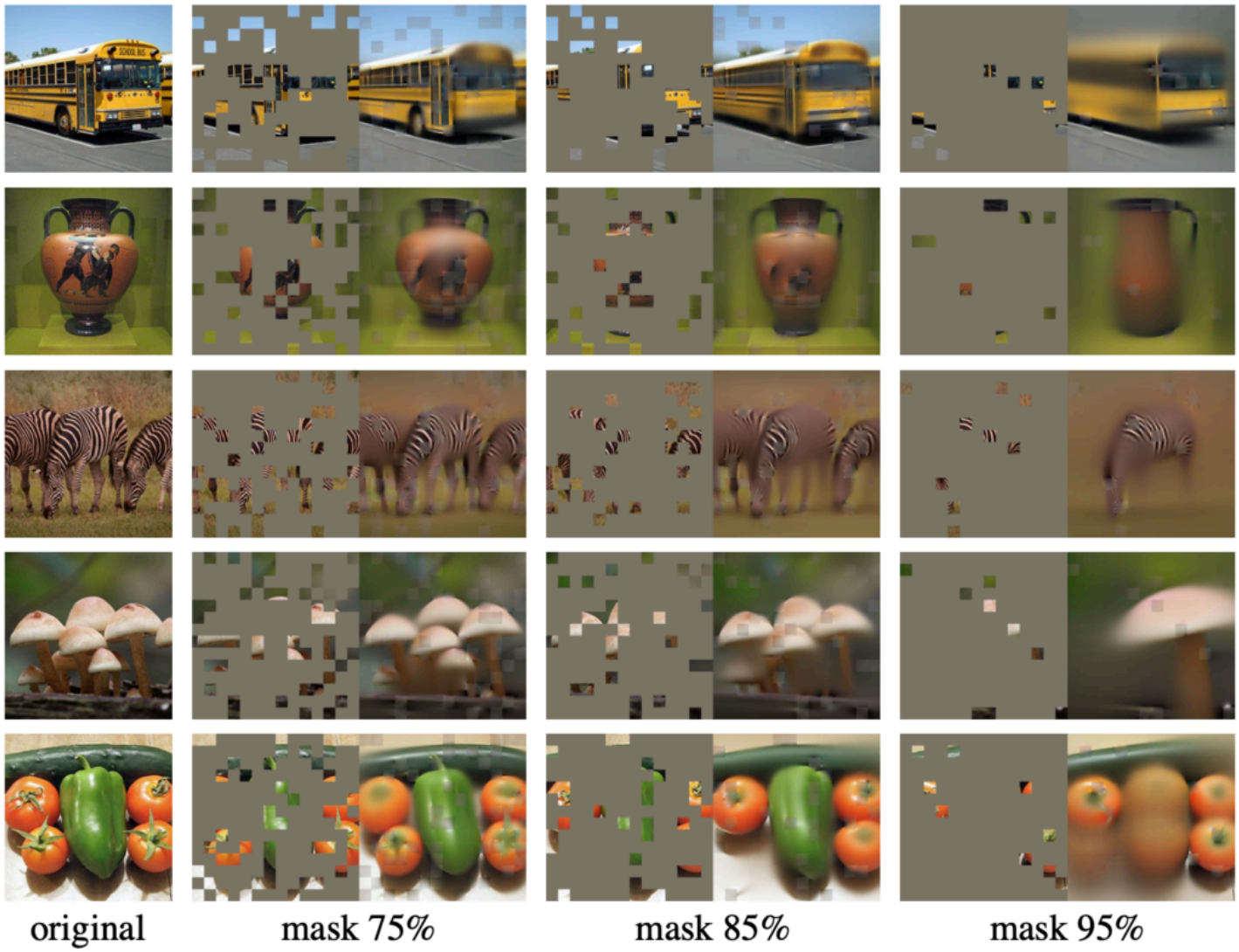
Motivation

ViT-MAE tries to address the challenges for reliably translating masked autoencoding to vision domain and solve the ever increasing need of data for vision transformers by relying on self-supervised training.

Implementation of a transformer-based encoder and decoder structure helps with resolving the challenges surrounding training on masked autoencoding with convolutional networks.

The second challenge is to support a deeper level semantic understanding of features, object and pixel and not just rely on higher level understanding for the reconstruction task. To tackle this, the training pipeline modifies how the data is fed to the network.

The idea is to use a very high masking proportion of the image patches, the ViT-MAE masks upto 75-80% of image patches to force a holistic understanding for the model.



Masking ratios performance variation

The masked autoencoding is a task based on self supervised learning. The self supervised learning paradigm removes the requirement of labeled datasets. This eases out a big bottleneck when it comes to training very large models and makes the learning much more scalable.

The masked autoencoding task removes about 75-80% regions of the image, presents this image as the input and trains the model to reconstruct the masked regions of the image. This kind of learning pipeline allows the usage of unlabled data available very affordably and allows the model to gain very deep understanding of the images.

The self-supervised training step is also known as the pre-training step, and serves as the basis of learning rich features in the image data. This pre-training step also helps in

optimizing and speeding up of the fine tuning on downstream tasks such as image segmentation and object detection.

Methodology

ViT-MAE follows a basic autoencoder architecture, with one modification where it follows an asymmetric structure. This implies that the encoder and decoder are not the typical mirror images of each other.

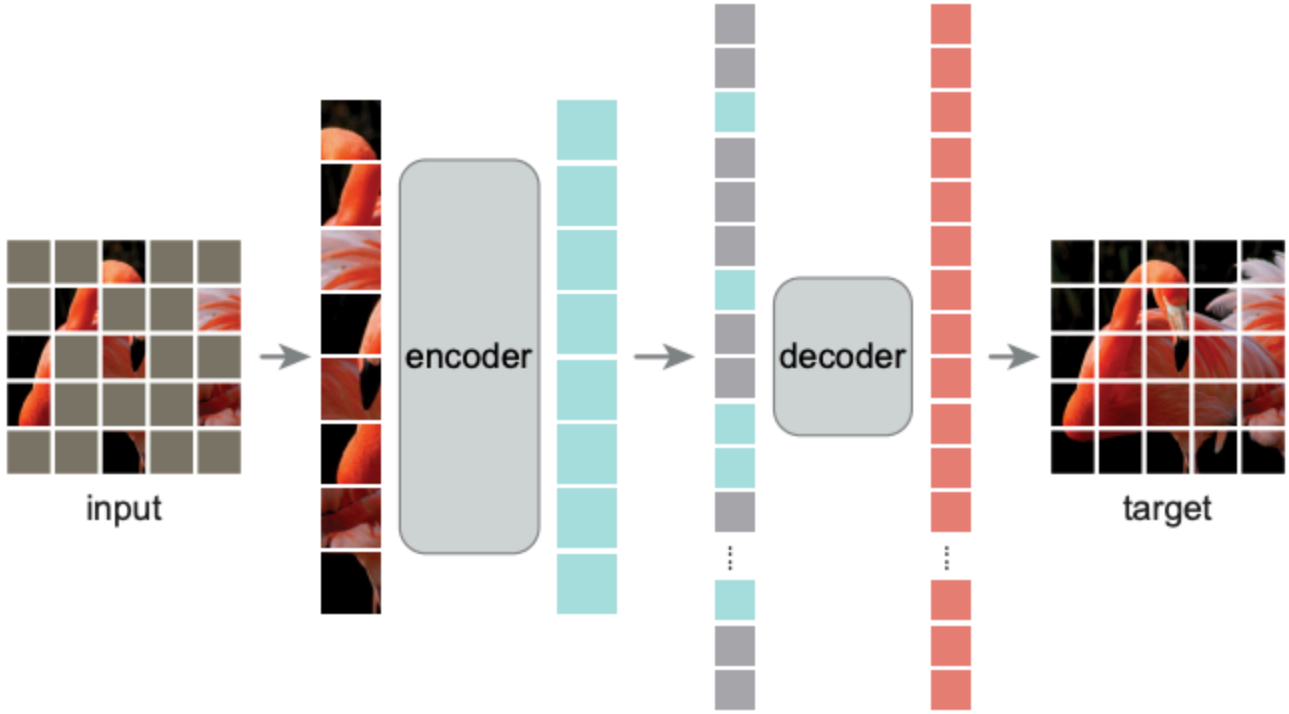
In fact for ViT-MAE, the encoder is bigger in size than the decoder. The decoder is a lightweight network which is responsible for reconstructing the entire image from the latent representation produced by the encoder network.

During pre-training, the image is broken up into a grid of non overlapping patches called input tokens. Among these non overlapping patches, a subset of patches are sampled out and removed or masked.

The masking ratio is kept high which implies a majority of patches are masked and only the minority of visible patches are fed to the encoder module. As explained before, this is done to address the challenge of deep understanding of image features and objects.

The masking ratio is kept at around 75-80% which makes sure that the reconstruction task is not trivial, and challenges the model to learn underlying features. The sampling strategy followed for the masked patches is uniform random sampling.

Uniform sampling also ensures that the patches are masked over the entire image and there’s no bias for the patches to be masked from the center.



ViT-MAE pre-training design

Breaking down the encoder

For the encoder, ViT architecture is used, which takes patches as inputs in form of linearly projected tokens and adds the positional embedding to localize each patch.

These patches are processed through a cascade of transformer blocks. The transformer blocks have the capability to process these tokens in parallel. The transformers deploy the self-attention mechanism which forms the connections between the input patches that emphasize their connecting weights according to their relationship with each other.

The self-attention mechanism trains the model to find the patches with closest relevance with each other.

The key feature of how encoder processes the input tokens in the ViT-MAE architecture is that only the unmasked visible patches are processed by the encoder. The idea is to utilize only a fraction of compute to train a very large encoder module.

The encoder module ten produces an embedding of the input image patches and is a latent representation of the input image as a whole.

The decoder

For the decoder, the encoded embedding produced by the encoder module is now combined with the tokens of the masked patches to form a full set of tokens. All the tokens are combined with positional embedding, which implies even the masked tokens get a positional embedding.

The masked token is a learned vector to indicate missing pixels. Note that the full set of tokens are only used by the decoder to reconstruct the image. Once the pre-training is complete for image reconstruction, decoder is not required any more however it is important to adjust the weights of the encoder reliably.

This allows a flexible design for the architecture of the decoder, and the default design only possesses less than 10% computation power of the encoder.

As an output, the decoder module produces vectors of pixel values. Each element in the decoder output contains same number of pixel values as the number of pixels in a patch. This output is reshaped to for the reconstructed image.

The training is governed by a loss function as in any machine learning pipeline. The loss function for the reconstruction task is Mean Squared Error. The loss is only calculated on the masked patches.

One issue with masked autoencoder is the sparse operations due to masked patches. The ViT-MAE solves this problem by random shuffling of patches. First tokens are generated for every patch in the image using linear projection along with the positional embedding. The tokens are added to a list and this list is then shuffled randomly.

Post shuffling, last tokens are removed or masked according to the masking ratio. This is similar to sampling without replacement. The masked tokens are then appended to the encoded visible patches from the encoder and unshuffling operation is carried out before decoder processes the full token set.

Results and Experiments

After the pre-training step is complete, ViT-MAE is ready to be fine tuned for other downstream tasks with transfer learning. These tasks are cognitive computer vision tasks which are directly applicable in the real world.

These tasks are image classification, object detection and image segmentation. The downstream tasks help verify the veracity of the algorithm by implementing the pre-trained network to solve a real world problem and judging its performance on well established benchmarks.

		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Object Detection performance comparison

For object detection task, ViT backbone is adopted with Feature Pyramid Network (FPN) to execute object detection. Four instances of this network are trained with different training methodologies, supervised, MoCo, BEiT and finally the MAE or Masked Autoencoding, and their results are compared.

MAE outperforms all the methods except BEiT, and is vastly superior over the ViT backbone model trained with supervised learning paradigm. The benchmark dataset

used for comparison is the IN1K data, supervised training obviously uses the labels, however other 3 models trained without labels also perform better.

This proves that self supervised learning is an effective technique to train the vision transformer and MAE methodology holds superior performance in the self supervised segment of models.

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	**48.1**	**53.6**

Image segmentation performance comparison

For segmentation task, ViT-MAE is trained to classify each pixel of an image into a class. As in the case of image segmentation, the ViT backbone is trained with four different training methodologies.

The dataset to benchmark the performance also remains the same with INIK. ViT-MAE outperforms BEiT marginally and beats the supervised training vision transformer by almost 4 points.

dataset	ViT-B	ViT-L	ViT-H	ViT-H_{448}	prev best
iNat 2017	70.5	75.7	79.3	**83.4**	75.4 [50]
iNat 2018	75.4	80.1	83.0	**86.8**	81.2 [49]
iNat 2019	80.5	83.4	85.7	**88.3**	84.1 [49]
Places205	63.9	65.8	65.9	**66.8**	66.0 [19]†
Places365	57.9	59.4	59.8	**60.3**	58.0 [36]‡

Image classification performance comparison

For Image classification task, different benchmarks are considered. Different sizes of MAE models are trained on iNaturalists and Places benchmarks. The largest model ViT-MAE H with the input image size 448, vastly outperforms the previously best model by 5-8 points.

Conclusion

Computer vision research has made strides in catching up to the NLP research in their capability to scale the learning of data. Self-supervised pre-training methods have proven to be very effective in exponentially increasing the learning capacity of models in NLP.

The self supervised learning paradigm allows utilization of unlabeled data and alleviates the need of effort and time resources in labeling data. This is even more critical when it comes to computer vision as data labeling for tasks like object detection and image segmentation is significantly more intensive due to sub pixel and pixel level annotations.

Traditionally, computer vision has relied only on supervised pre-training paradigms, creating a bottleneck due to heavy dependence on data labeling, ViT-MAE tries to address this problem with masked autoencoding training which can scale up for large unlabeled datasets.

Masked autoencoding has unique challenges in computer vision which need to be addressed separately from NLP. The biggest difference is apparent with the masking required to gain effective deep understanding of the image data as compared to language data. With images it is trivial to accomplish the image reconstruction by predicting image patches given neighboring patches.

To learn more about the research Bolster conducts on vision transformers, and to see how our AI Security platform can protect your business, ***contact our team today****.*