Quantizing Visual Features

- self-supervised models, like BEiTv2 can be linear probed to downsteam classification tasks, like ImageNet-1K
- result is not perfect, but still quite good -> BEiTv2 reaches 80.1% top-1 accuracy on ImageNet-1K
- consequently, given a function f that maps an image to a feature vector, there must exist a function g

that maps the feature vector to the correct class label (largest object in the image)

- function is approximated by a linear classifier, trained on labeled data
- function f is linear -> $f(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$ (*linear* probing)
- index with highest value (logit) is the index of the predicted class