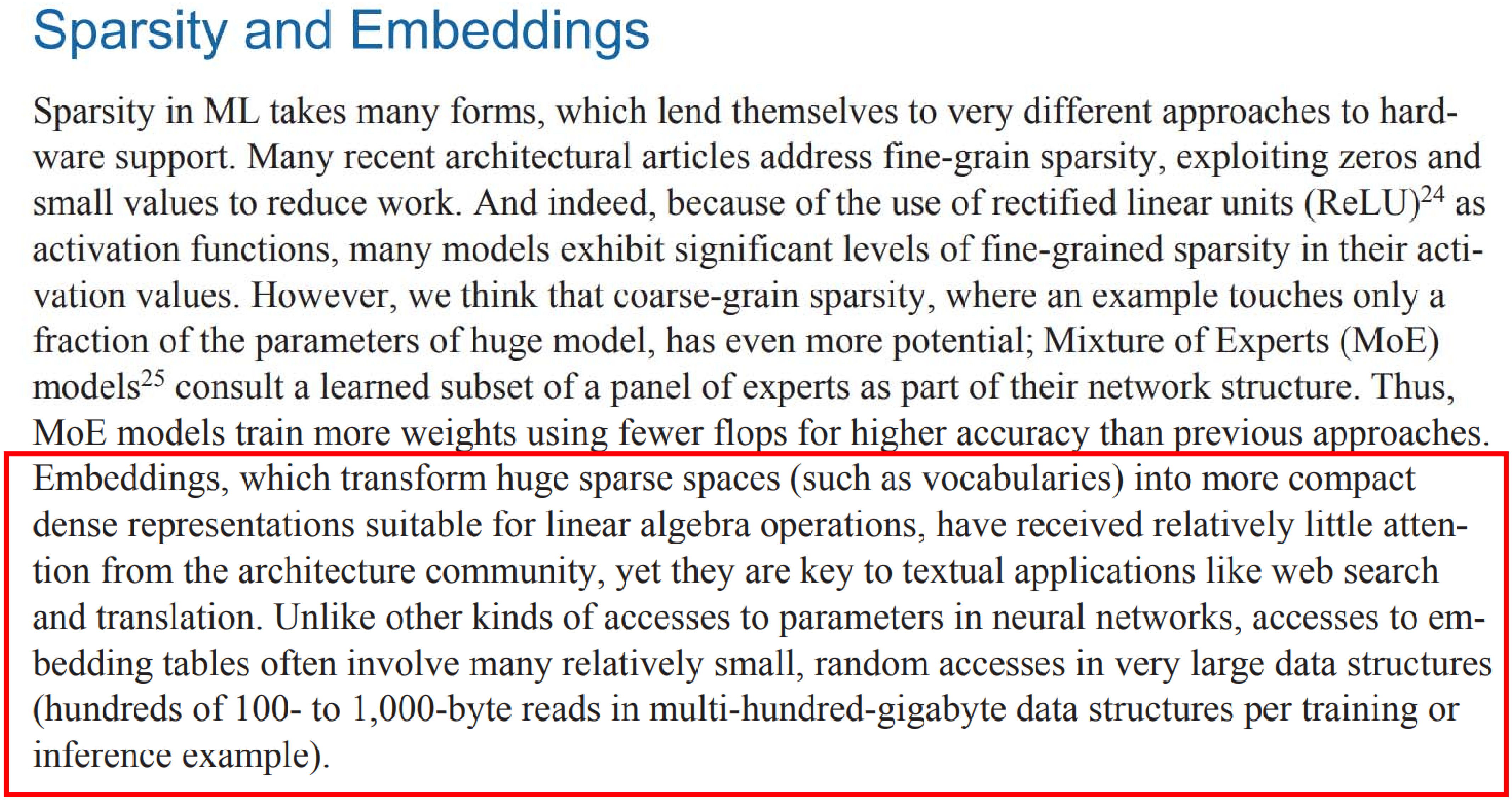
Jeff Dean, David Patterson, and Cliff Young, "A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution," *IEEE Micro* 38.2 (2018): 21-29



Sparsity and Embeddings

Sparsity in ML takes many forms, which lend themselves to very different approaches to hardware support. Many recent architectural articles address fine-grain sparsity, exploiting zeros and small values to reduce work. And indeed, because of the use of rectified linear units (ReLU)[24] as activation functions, many models exhibit significant levels of fine-grained sparsity in their activation values. However, we think that coarse-grain sparsity, where an example touches only a fraction of the parameters of huge model, has even more potential; Mixture of Experts (MoE) models[25] consult a learned subset of a panel of experts as part of their network structure. Thus, MoE models train more weights using fewer flops for higher accuracy than previous approaches. Embeddings, which transform huge sparse spaces (such as vocabularies) into more compact dense representations suitable for linear algebra operations, have received relatively little attention from the architecture community, yet they are key to textual applications like web search and translation. Unlike other kinds of accesses to parameters in neural networks, accesses to embedding tables often involve many relatively small, random accesses in very large data structures (hundreds of 100- to 1,000-byte reads in multi-hundred-gigabyte data structures per training or inference example).