Is it efficient to use matrix multiplication to implement convolution operations? Why?

Yes, it is. Because there are may very efficient implementations of Matrix Multiplication we can take advantage of, such as BALS API, what’s more, nowadays, matrix multiplication can be implemented efficiently in modern GPUs.