Paper review

The paper provides with info on the cuDNN library. It states the necessity to speed-up convolution computations when training Deep Neural Networks. It also shows different approaches to accelerate convolutions.

Key insights:

1. There are three ways to increase the speed of convolutions:

* Lower them to matrix multiplication.
* Use Fast Fourier Transform.
* Direct computation approaches.

1. cuDNN lowers convolutions to matrix multiplication, since matrix multiplication is highly optimized.
2. Solid explanation of Spatial Convolution computation.

The article gives detailed explanation of cuDNN library, how the library optimizes convolutions and why its optimization approach was chosen. It also tells that the library is independent from GPU architecture and is integrated in some useful frameworks.

We think that the article is useful and accept it.