(Courbariaux, Hubara 2016)
Two innovations:

1. Augment with continuous variable (weights). In forward propagation, take continuous weights and binarize to get binary weights.

Go through

- Matmul or conv+MP
- batch norm (at test time, this is identity if there are only -1's and 1's),
- binarize.

It's more efficient at test time.

Right before softmax layer, don't binarize.

This is a deterministic binarization. It's theoretically nicer to do stochastic binarization, but in practice it isn't better.

Ternary might be better: 0 means no contribution, and can allow sparse connectivity.

Backprop through binarize function is weird.

2. Traditionally in backprop, take derivative. Here we replace forward function with smooth function on backward pass.

Directions are preserved by binarization. Take a random vector (from normal distribution, which is rotationally invariant) and binarize. Angle between vector and binarization is about 37° smaller than between random vectors 90°. 37° is small in high dimensions.

First layer behaves a lot differently—it behaves worse.

Plot distribution between continuous and binarized weights. There's a systematic deviation towards larger angle.

We convert convolution to matrix multiplication. First convolution takes 3x3 patch of 3 channel image (27 dimensions). Next is 3x3x128.

If you train to completion, you get edge filters, etc. You don't get this is you stop when predictions are good enough.