TITLE (BOARD)

BATCH NORMALIZATION

AND WHAT WE KNOW SO FAR

PATH (PRESENTATION)

HTTPS://GITHUB.COM/GVASCONS/BATCH-NORMALIZATION

PAPER (ARTICLE)

HTTPS://DOI.ORG/10.48550/ARXIV.1502.03167

Batch Normalization aims to reduce internal covariate shift, and in doing so aims to accelerate the training of deep neural nets. It accomplishes this via a normalization step that fixes the means and variances of layer inputs. Batch Normalization also has a beneficial effect on the gradient flow through the network, by reducing the dependence of gradients on the scale of the parameters or of their initial values. This allows for use of much higher learning rates without the risk of divergence. Furthermore, batch normalization regularizes the model and reduces the need for Dropout.

We apply a batch normalization layer as follows for a minibatch $\mathcal{B}$:

$$\mu_{\mathcal{B}} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_{\mathcal{B}}^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2$$

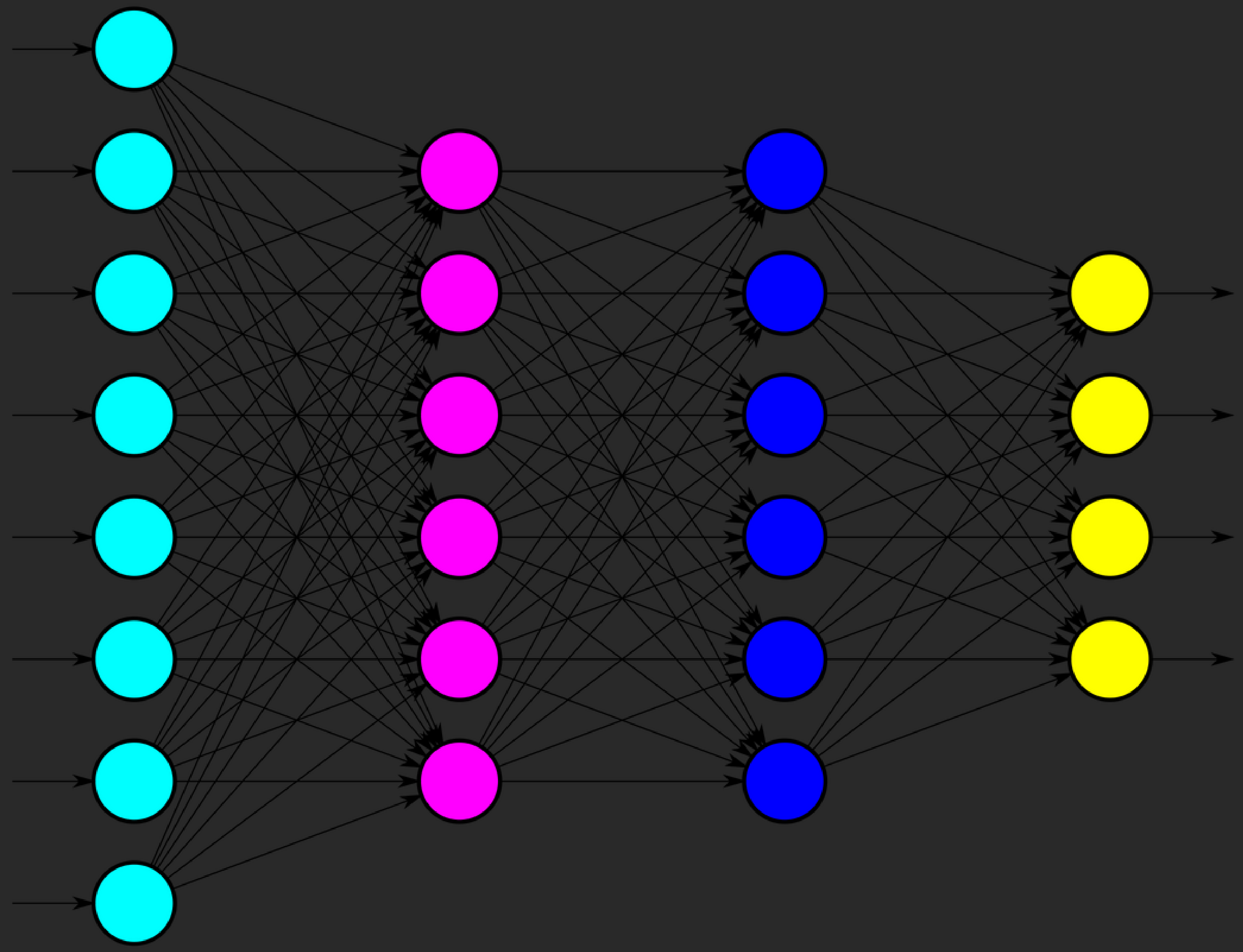
$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

$$y_i = \gamma\hat{x}_i + \beta = \mathrm{BN}_{\gamma,\beta}(x_i)$$

Where γ and β are learnable parameters.

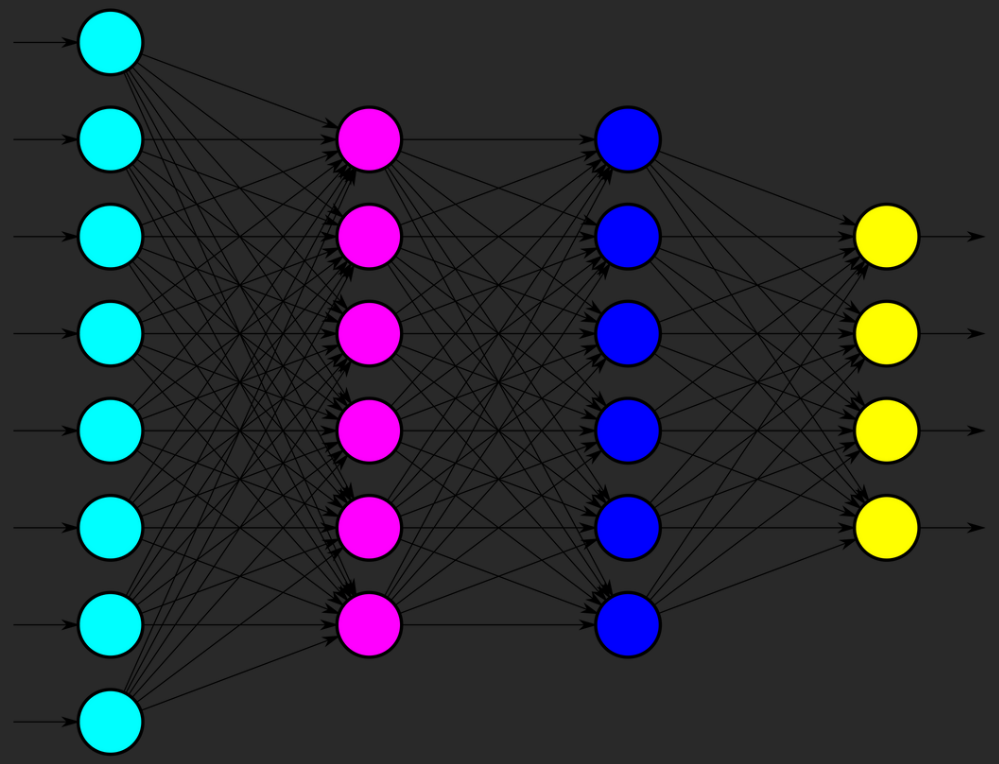
CONTEXT

- DEEP NEURAL NETWORKS
  - HAS A CHANGE OF DISTRIBUTION ON EACH LAYER'S INPUT IN IT'S ARCHITECTURE



OVERVIEW

- BATCH NORMALIZATION



- RAW SIGNAL
- HIGH INTERDEPENDANCY BETWEEN DISTRIBUTIONS
- SLOW AND UNSTABLE TRAINING

- NORMALIZED SIGNAL
- MITIGATED INTERDEPENDANCY BETWEEN DISTRIBUTIONS
- FAST AND STABLE TRAINING

PRINCIPLE

- DIFFERENT COMPUTATIONS ON TRAINING AND EVALUATION

PRINCIPLE

- DIFFERENT COMPUTATIONS ON TRAINING AND EVALUATION

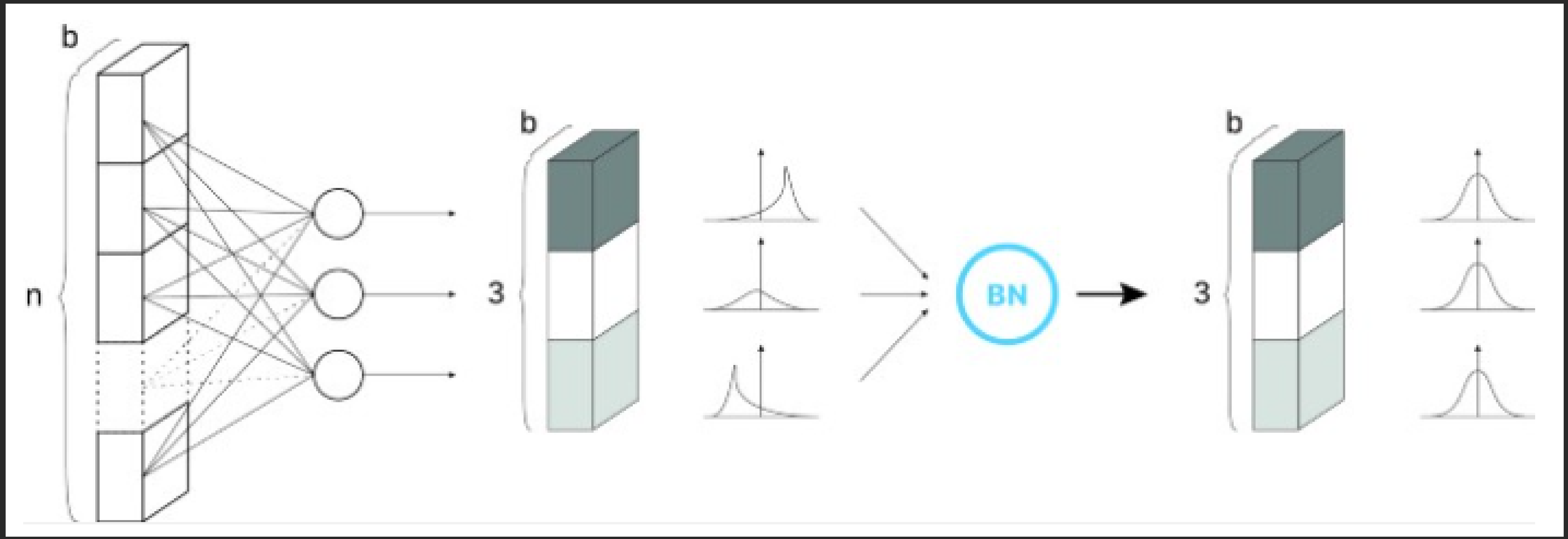
TRAINING

$$(1)\quad \mu = \frac{1}{n}\sum_i Z^{(i)}$$

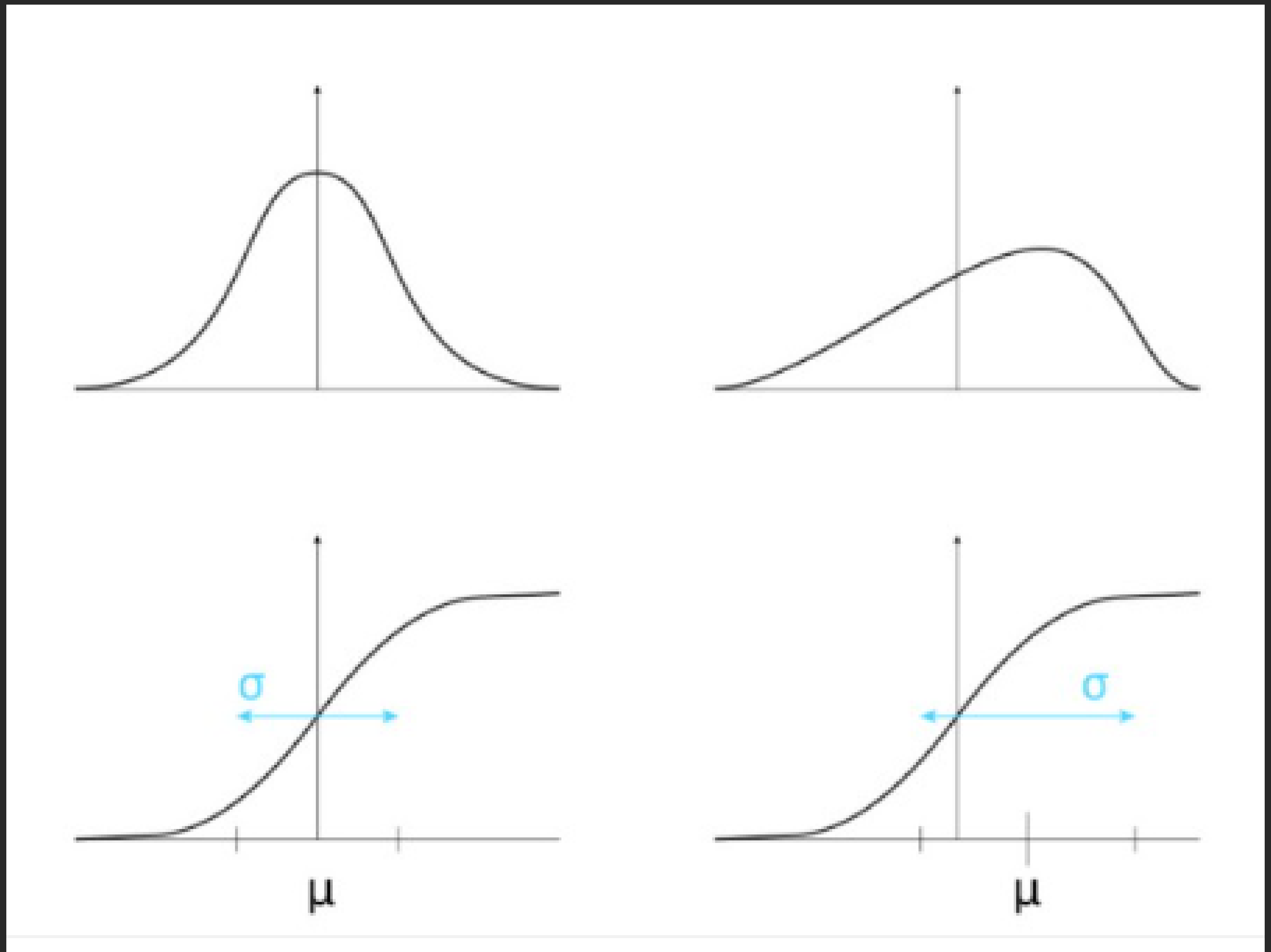
$$(2)\quad \sigma^2 = \frac{1}{n}\sum_i (Z^{(i)} - \mu)^2$$

$$(3)\quad Z_{norm}^{(i)} = \frac{Z^{(i)} - \mu}{\sqrt{\sigma^2 - \epsilon}}$$

$$(4)\quad \breve{Z} = \gamma * Z_{norm}^{(i)} + \beta$$



BATCH NORMALIZATION FIRST STEP. EXAMPLE OF A 3-NEURONS HIDDEN LAYER, WITH A BATCH OF SIZE B. EACH NEURON FOLLOWS A STANDARD NORMAL DISTRIBUTION.



BENEFITS OF γ AND β PARAMETERS. MODIFYING THE DISTRIBUTION (ON THE TOP) ALLOWS US TO USE DIFFERENT REGIMES OF THE NONLINEAR FUNCTIONS (ON THE BOTTOM).

TRAINING PRINCIPLE

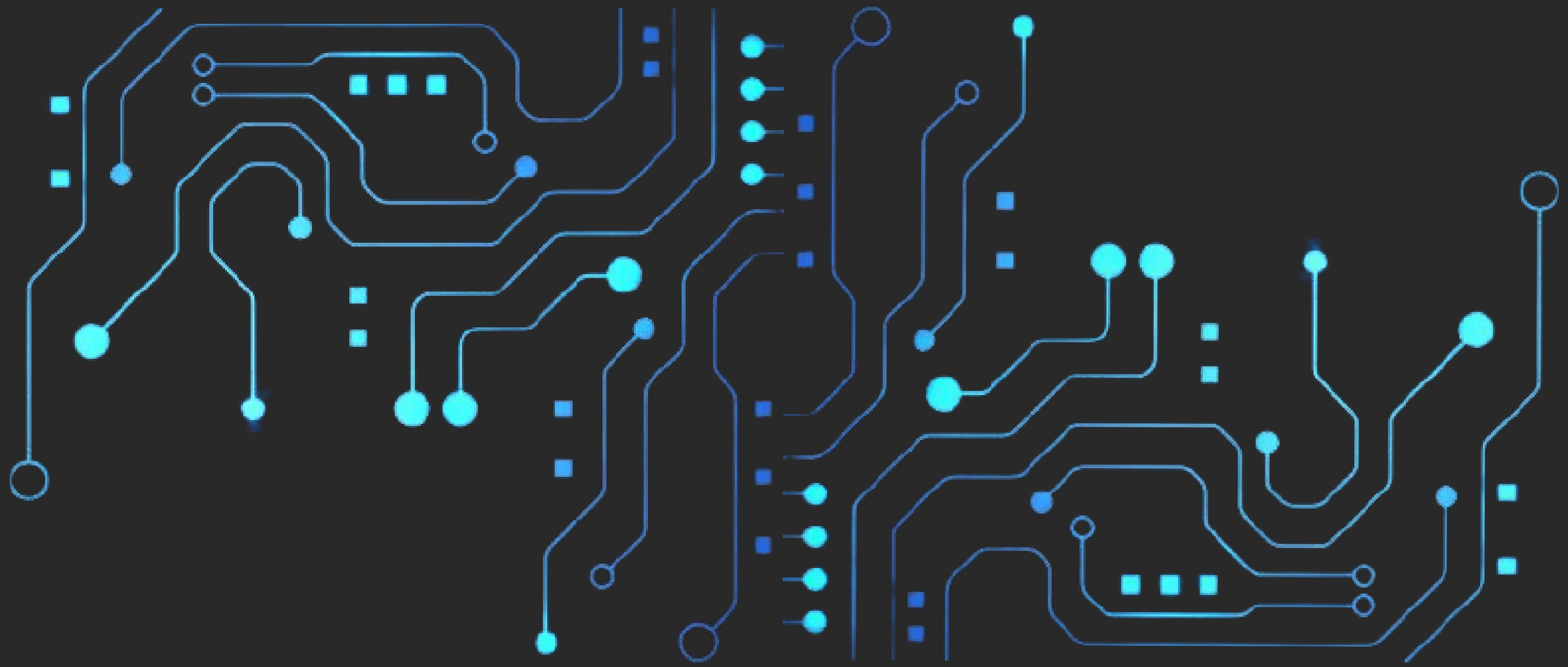
- DIFFERENT COMPUTATIONS ON TRAINING AND EVALUATION

PRINCIPLE

- DIFFERENT COMPUTATIONS ON TRAINING AND EVALUATION

EVALUATION

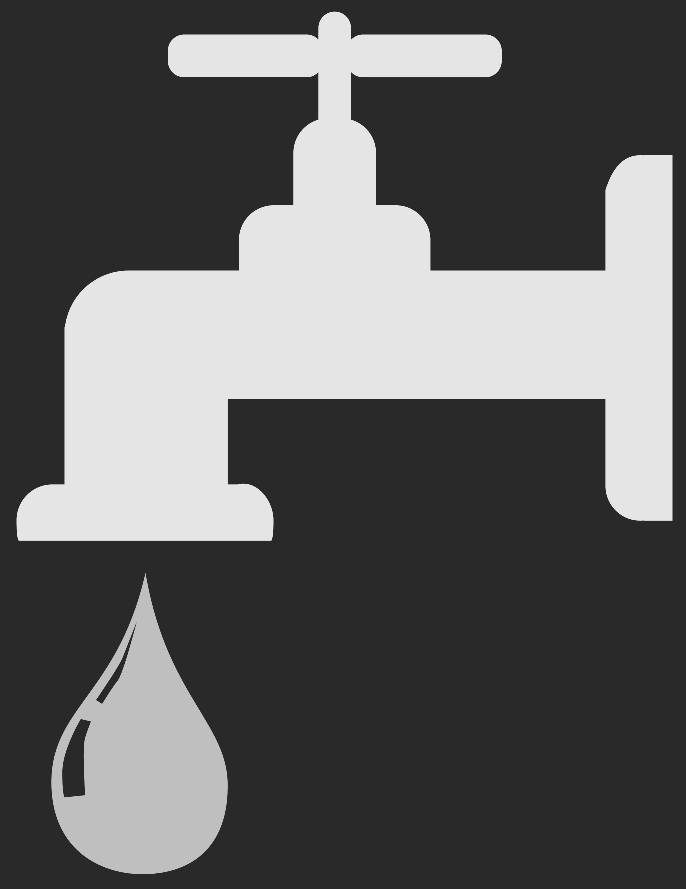
- BASED ON ESTIMATION
- DETERMINED DURING TRAININ
- DIRECTLY FED INTO EQUATION



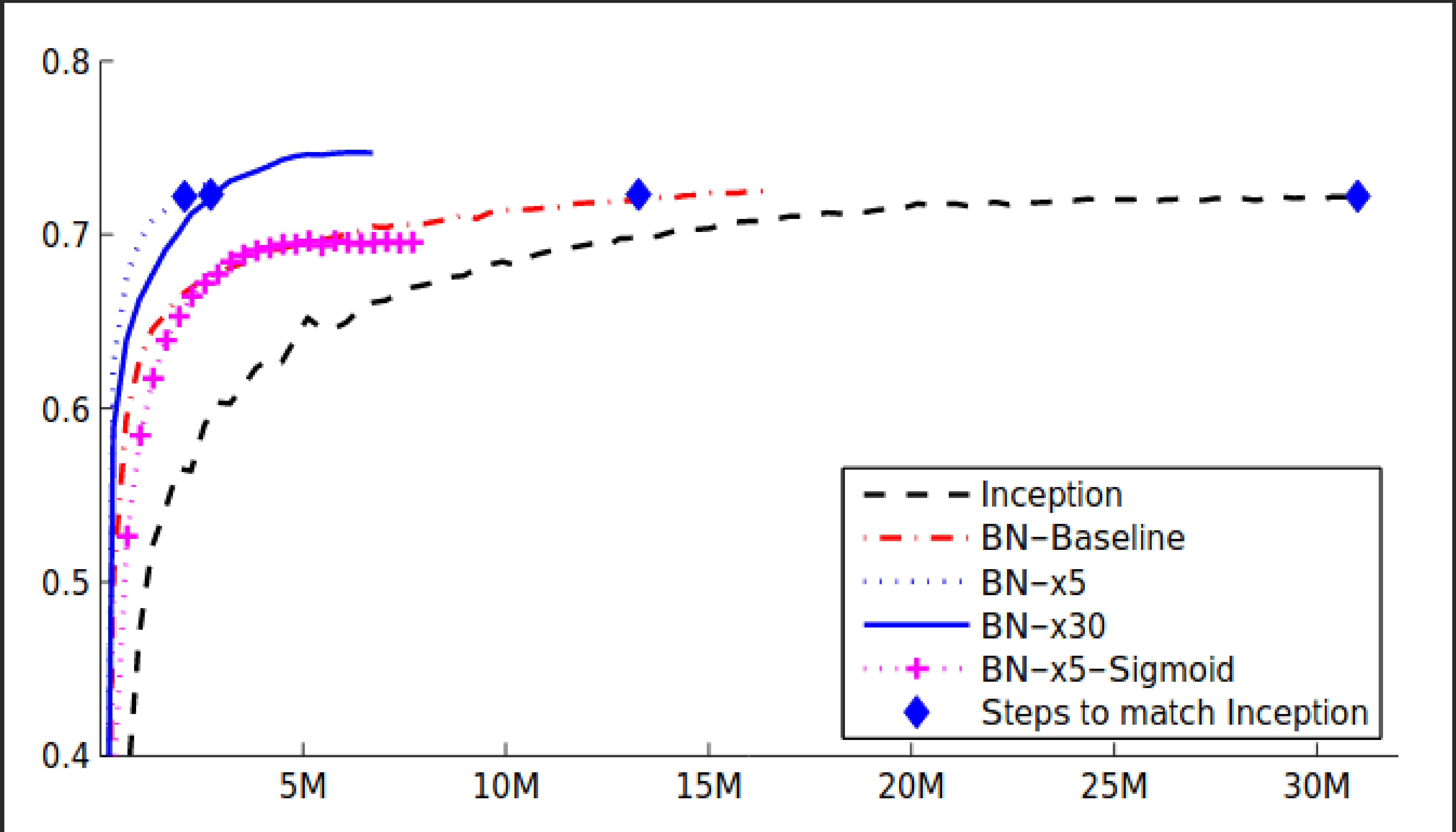
IN PRACTICE

- HOW MANY NEURONS ARE IN THE CURRENT HIDDEN LAYER
- HOW MANY FILTERS ARE IN THE CURRENT HIDDEN LAYER

WHAT'S THE LINK BETWEEN BATCH NORMALIZATION AND THIS PICTURE ?

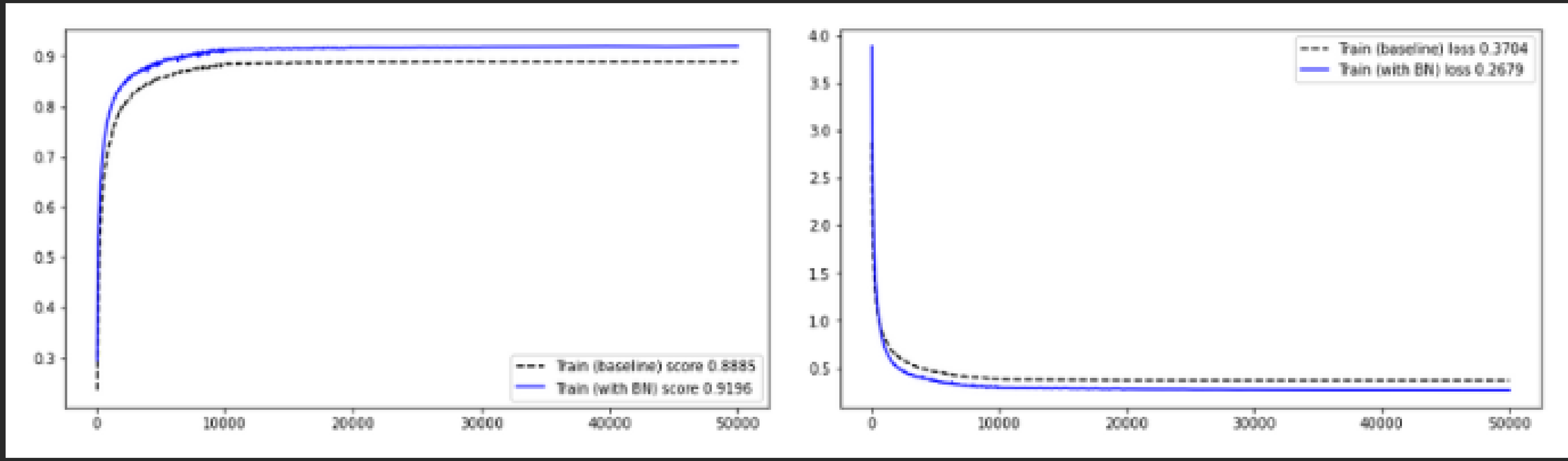


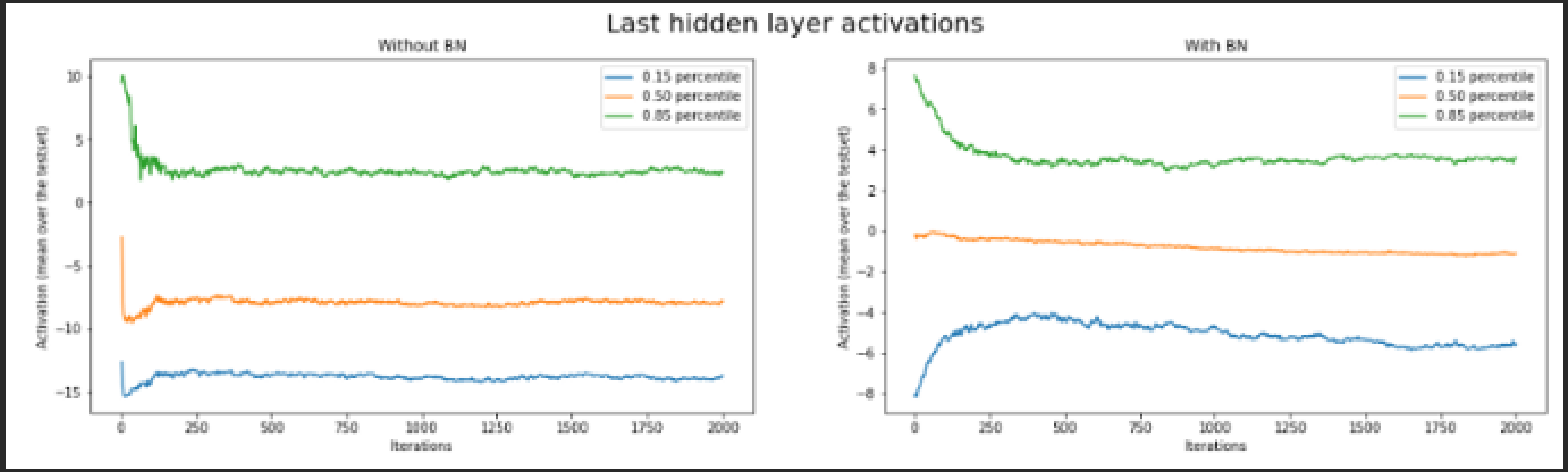
RESULTS OVERVIEW

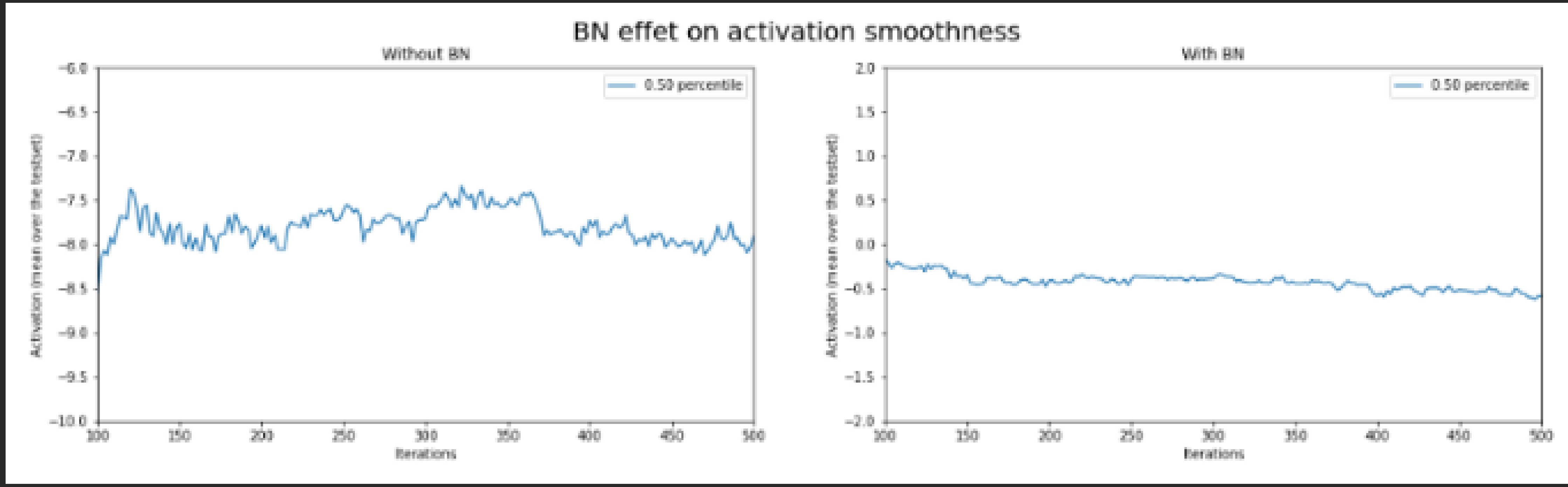


UNDERSTANDING

- BN LAYER IN PRACTICE
  - HOW DOES BN IMPACT TRAINING PERFORMANCES ? WHY IS THIS METHOD SO IMPORTANT IN DEEP LEARNING NOWADAYS ?
  - WHAT ARE THE BN SIDE-EFFECTS WE MUST BE AWARE OF ?
  - WHEN AND HOW SHOULD WE USE BN ?







UNDERSTANDING RESULTS

REGULARIZATION

" PUT SIMPLY, WE SHOULD ALWAYS MAKE SURE THAT ONE MODULE ADDRESSES ONE ISSUE. RELYING ON SEVERAL MODULES TO DEAL WITH DIFFERENT PROBLEMS MAKES THE DEVELOPMENT PROCESS MUCH MORE DIFFICULT THAN NEEDED "

NORMALIZATION ON EVALUATION

- *CASES*
  - *WHEN DOING CROSS-VALIDATION OR TEST, DURING MODEL TRAINING AND DEVELOPMENT ;*
  - *WHEN DEPLOYING THE MODEL.*

LAYER STABILITY

IF THE INPUT DISTRIBUTION VARIES TOO MUCH FROM TRAINING TO EVALUATION, THE MODEL COULD OVEREACT TO SOME SIGNALS, RESULTING IN ACTIVATON DIVERGENCE.

RECURRENT NETWORK

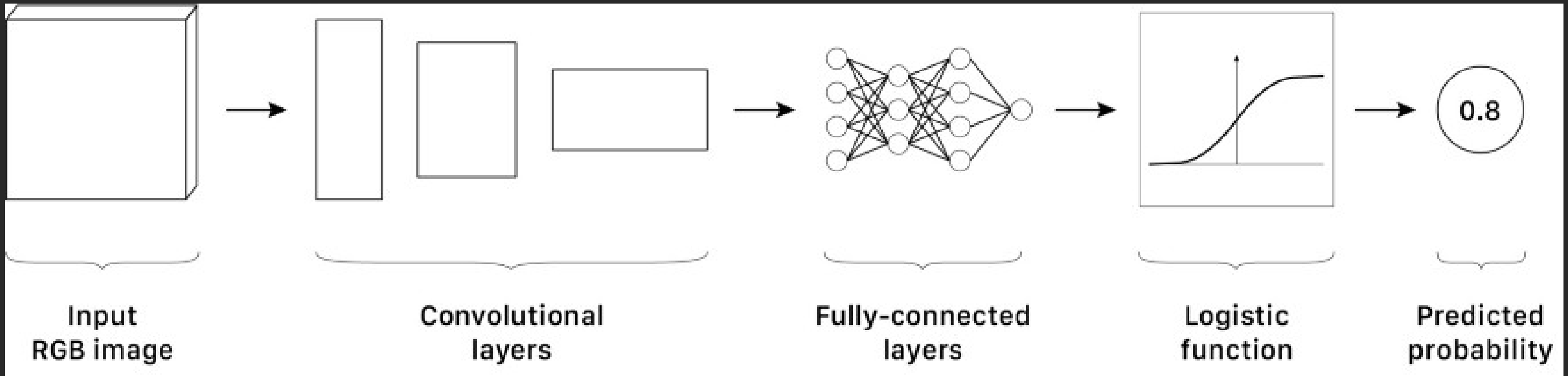
FOR CONVOLUTIONAL NETWORKS (CNN) : BATCH NORMALIZATION (BN) IS BETTER
FOR RECURRENT NETWORK (RNN) : LAYER NORMALIZATION (LN) IS BETTER

BEFORE OR AFTER NONLINEARITY

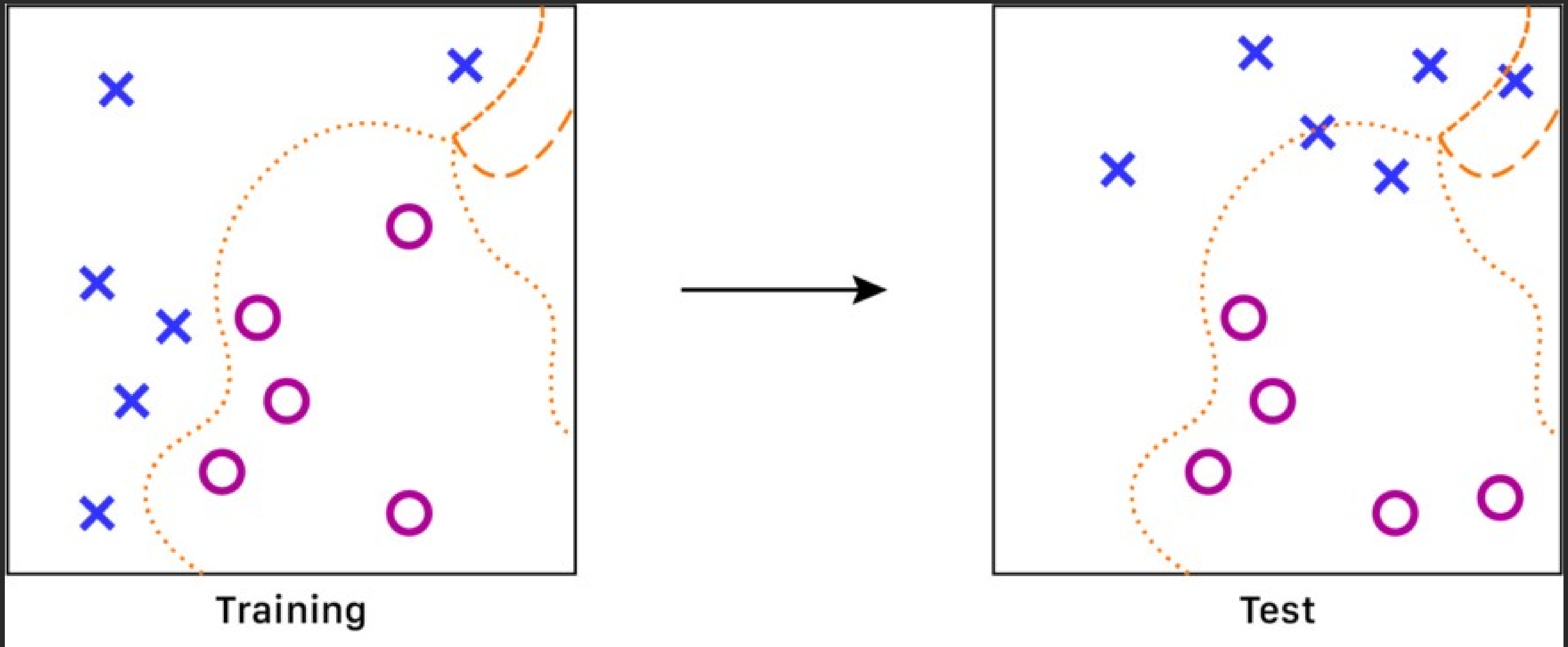
"WE WOULD LIKE TO ENSURE THAT, FOR ANY PARAMETER VALUES, THE NETWORK ALWAYS PRODUCES ACTIVATIONS WITH THE DESIRED DISTRIBUTION."

WHY DOES IT WORK?

BN REDUCES THE INTERNAL COVARIANCE SHIFT (ICS)



THE COVARIATE SHIFT CAN MAKE THE NETWORK ACTIVATIONS DIVERGE (SECTION C.2.4). EVEN IF IT DOESN'T, IT WOULD DETERIORATE OVERALL PERFORMANCES !



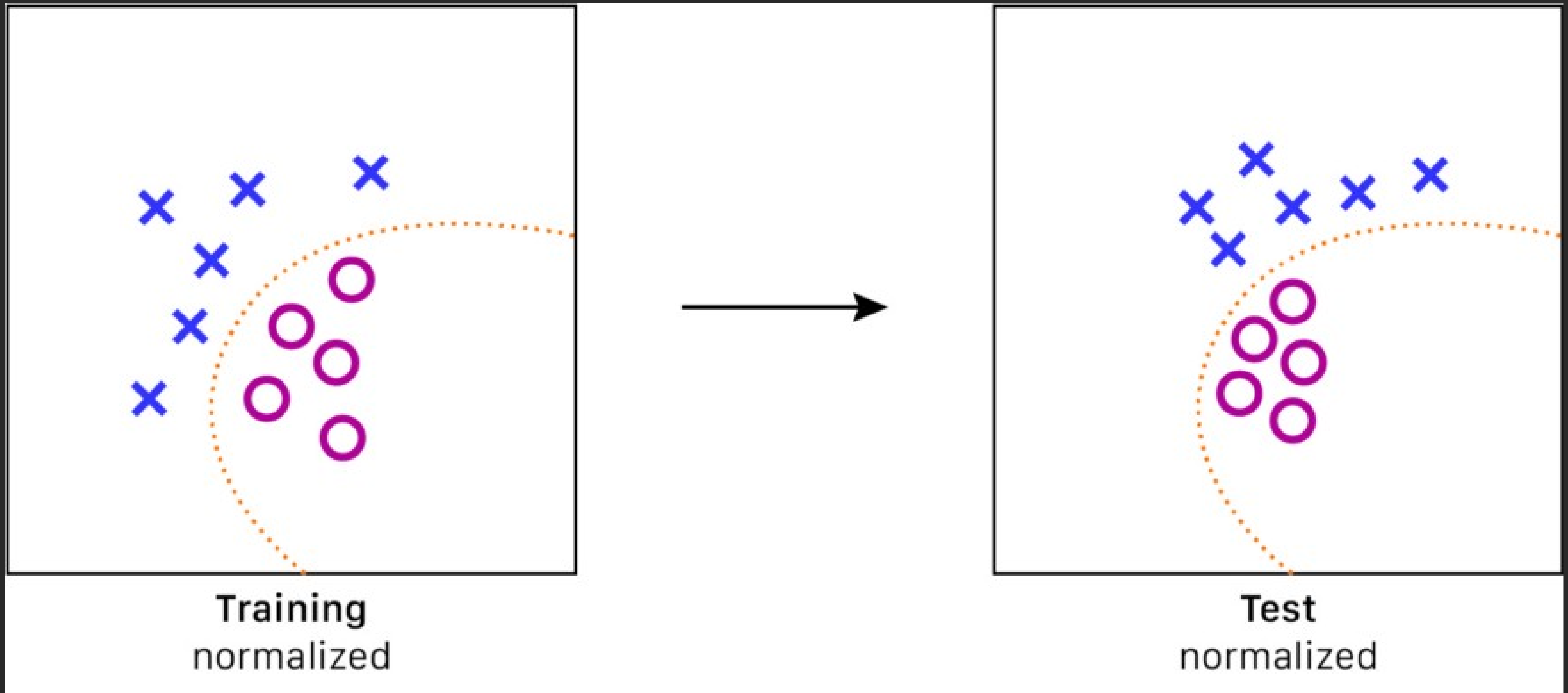
WITHOUT NORMALIZATION, DURING TRAINING, INPUT VALUES ARE SCATTERED : THE APPROXIMATED FUNCTION WILL BE VERY ACCURATE WHERE THERE'S A HIGH DENSITY OF POINTS. ON THE CONTRARY, IT WILL BE INACCURATE AND SUBMITTED TO RANDOMNESS WHERE THE DENSITY OF POINTS IS LOW.

DOES BN REDUCE THE INTERNAL COVARIANCE SHIFT (ICS)?

WHY DOES IT WORK?

WHY DOES IT WORK?

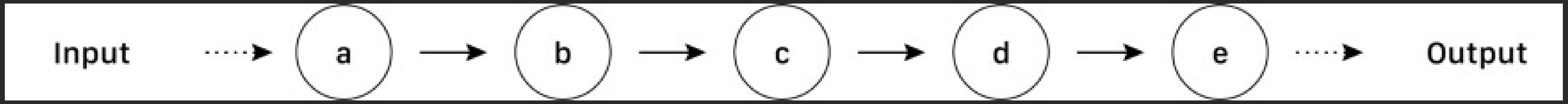
DOES BN REDUCE THE INTERNAL COVARIANCE SHIFT (ICS)?



CASE WITH NORMALIZATION. NORMALIZING THE INPUT SIGNAL MAKES THE POINTS CLOSER TO EACH OTHER IN THE FEATURE SPACE DURING TRAINING : IT IS NOW EASIER TO FIND A WELL GENERALIZING FUNCTION.

WHY DOES IT WORK?

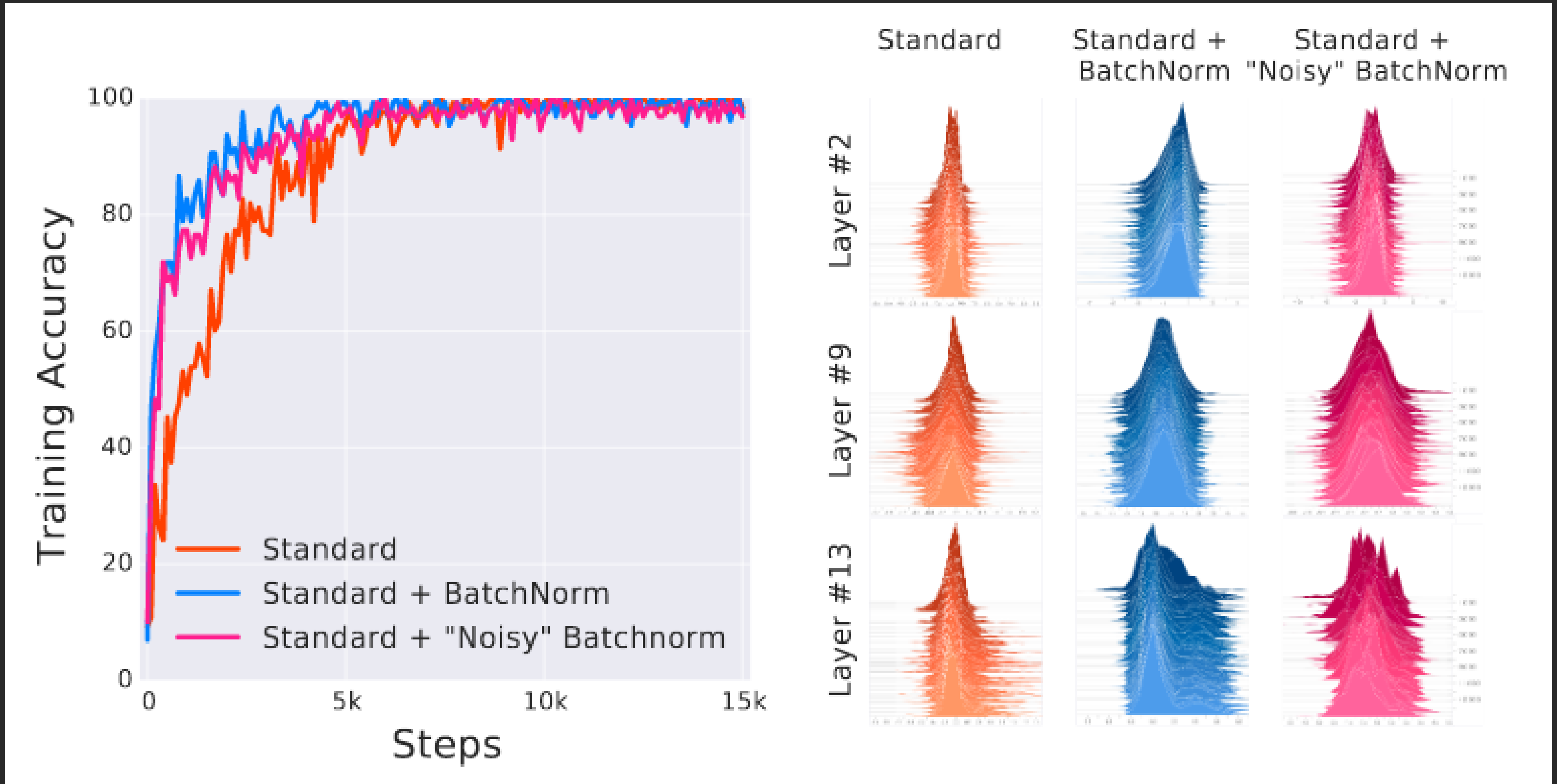
DOES BN MITIGATE INTERDEPENDENCY BETWEEN HIDDEN LAYERS DURING TRAINING?



A SIMPLE DNN, WHICH CONSISTS OF LINEAR TRANSFORMATIONS.

WHY DOES IT WORK?

DOES BN MAKES THE OPTIMISATION LANDSCAPE SMOOTHER?



FYI

HTTPS://TOWARDSDATASCIENCE.COM/BATCH-NORMALIZATION-IN-3-LEVELS-OF-UNDERSTANDING-14C2DA90A338

HTTPS://DOI.ORG/10.48550/ARXIV.1502.03167