ADAPTIVE INITIALIZATION FOR RECURRENT PHOTONIC NETWORKS USING SIGMOIDAL ACTIVATIONS

Nikolaos Passalis, George Mourgias-Alexandris, Nikos Pleros, and Anastasios Tefas

`{passalis, mourgias, npleros, tefas}@csd.auth.gr`

Narrator: Nikolaos Passalis

Dept. of Informatics, Aristotle University of Thessaloniki, Greece

2020 IEEE International Symposium on Circuits and Systems

Virtual, October 10-21, 2020

STRUCTURE

- Introduction to Photonic Deep Learning
- Proposed Method
- Experimental Evaluation
- Conclusions

- **Deep Learning (DL)** provided state-of-the-art solutions to many challenging problems
    - … but **DL models are especially complex**
    - … **powerful hardware is needed both for training and deploying DL models**
- Several **hardware accelerators** have been developed
    - Graphics Processing Units (GPUs)
    - Tensor Processing Units (TPUs)
    - …
- Neuromorphic solutions are especially promising providing **fast and energy efficient DL** accelerators by directly providing the functionality of neurons

- Photonic DL accelerators use **light to represent signals**
- These signals can be then **appropriately manipulated,** using either purely **optical** components, or a combination of **electro-optical** components, to perform computations
- Several advantages
  - Information is **propagated near to the speed of light**
  - **Enormous bandwidth** that provides a **massive parallelism potential**
  - Photonic neurons can operate at **extremely high frequencies**

LIMITATIONS

- Photonic neuromorphic platforms currently face **several important limitations**
- Among them is that many DL-oriented **activation functions cannot be precisely implemented using photonic hardware**
- **DL models must be retrained** using photonic-compliant activation functions before deployment
- Using such functions is not straightforward, e.g., saturable functions can lead to vanishing gradients phenomena, slowing down or even **stopping the learning process**

LIMITATIONS

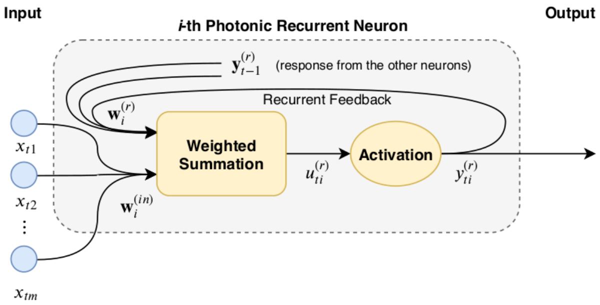
- In “traditional” DL this problem was solved by developing **appropriate activation functions,** e.g., ReLU, along with appropriate **initialization** schemes
- This is not always possible for photonic DL
- Developing the appropriate initialization scheme to ensure that the models will be initialized into a region that **allows for information to be propagated**
- This is even more **critical for recurrent architectures**, where **both vanishing and exploding gradient phenomena** can be occur

CONTRIBUTION

- We propose an **adaptive data-driven initialization method** that can overcome these limitations
  - The proposed method is **activation-agnostic** (i.e., can be used with any activation function)
  - **Does not require manually and analytically calculating the initialization variance** for different activation functions[1]
  - Takes into account both the **actual distribution** of the data used to train the network and the **task at hand**
  - **Simple and easy to implement!**
- **Solid step toward the effective training of photonic DL models**, overcoming many limitations of existing variance-preserving initialization methods

[1] Passalis, Nikolaos, et al. "Training deep photonic convolutional neural networks with sinusoidal activations." IEEE Transactions on Emerging Topics in Computational Intelligence (2019).

- This work focuses on **training deep recurrent neural networks using a sigmoid-based activation** function that can be implemented using a recently proposed all-optical activation mechanism[2]



[2]G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos, and N. Pleros, "An all-optical neuron with sigmoid activation function," Optics express, vol. 27, no. 7, pp. 9620–9630, 2019

- The proposed method exploits the **effectiveness in training shallow (up to two layers) neural networks** to estimate the optimal initialization variance of a layer
- The proposed method estimates the initialization variance **layer-by-layer** (starting with the input layers)
- Each layer is equipped with a trainable scaling factor α_i that is used to estimate the optimal initialization variance:

$$\mathbf{y}_t^{(i)} = f(|\alpha_i|\mathbf{W}_i\mathbf{y}_{t-1}^{(i-1)} + \mathbf{b}_i), \tag{1}$$

where $f(\cdot)$ is the employed activation function, and $\mathbf{W}_i$ and $\mathbf{b}_i$ the weights and biases of the layer.

- Then, **an auxiliary classification layer** is used on top of each layer of the network and trained using regular gradient descent
- Only the **auxiliary classification layer** and **the scaling factor** are trained (the layer's weights are kept fixed)
- The value of the **scaling factor implicitly provides an estimation for the optimal initialization variance**
- After estimating the variance for a layer, this process is repeated with the next one
- The **scaling factors are discarded** after this process is completed and the **network is re-initialized and can be directly trained**

Algorithm 1 Adaptive Data-driven Initialization

Input: Initial weights $\mathbf{W}_i$, learning rate η, number of iterations N_{est}

Output: Initialization variance for each layer σ_i^2

Initialize the layers $\mathbf{W}_i$ using any initialization scheme
for i=1 **to** n **do**
  Initialize α_i to 1
  for j=1 **to** N_{est} **do**
    Update parameters using gradient descent($\frac{\partial \mathcal{L}}{\partial \alpha_i}, \frac{\partial \mathcal{L}}{\partial \mathbf{W}_i^{class}}$)
  end for
  Calculate the variance as $\sigma_i^2 = (\alpha_i \sigma_{init})^2$
end for
return Estimated values for σ_i

EXPERIMENTAL SETUP

- **Two different time-series datasets**, suitable for recurrent neural architectures, were used:
  - a **high-frequency limit order book dataset** (abbreviated as “FI-2010”), and
  - a **household power consumption forecasting dataset** (abbreviated as “HPCF”)
- A **recurrent neural networks with 32 recurrent units**, followed by **two fully connected layers** with 512 units and N_C output neurons (number of classes)
- The **RMSprop optimization algorithm** was used for all the conducted experiments.
- The optimization ran for **20 epochs for the FI-2010** dataset and for **10 epochs for the HPCF dataset**

Model	Initialization	Avg. F1	Cohen's κ
MLP	Xavier	35.27 ± 1.05	0.1058 ± 0.0108
LSTM	Xavier	43.61 ± 1.17	0.1796 ± 0.0142
RNN (sigmoid)	Xavier	40.44 ± 1.77	0.1648 ± 0.0184
Photonic RNN	Xavier	34.46 ± 1.78	0.0928 ± 0.0175
Photonic RNN	He	33.43 ± 0.87	0.0849 ± 0.0098
Photonic RNN	Proposed (Xav.)	41.21 ± 1.78	0.1635 ± 0.0216
Photonic RNN	Proposed (He)	**41.68 ± 2.73**	**0.1693 ± 0.0300**

EXPERIMENTAL RESULTS (HPCF DATASET)

Model	Initialization	Accuracy
MLP	Xavier	60.07%
LSTM	Xavier	75.46%
RNN (sigmoid)	Xavier	69.58%
Photonic RNN	Xavier	63.20%
Photonic RNN	He	57.29%
Photonic RNN	Proposed (Xavier)	73.03%
Photonic RNN	Proposed (He)	**73.63%**

CONCLUSIONS

- An **adaptive data-driven initialization approach** for recurrent **photonic neural networks** was proposed
- Can be directly used with any photonic activation function
- It takes into account the actual distribution of the data used to train the network
- Provides a solid approach for training DL models, that would be otherwise very difficult to train and would **require manually tuning the variance** for each layer or **analytically deriving the optimal layer-wise** initialization variance
- Sample implementation available at `https://github.com/passalis/adaptive_phos`

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871391 (PlasmoniAC). This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains.

www.plasmoniac.eu

Thank You!