Bayesian Sparsification of Deep Complex-valued networks

Ivan Nazarov, Evgeny Burnaev

ADASE Skoltech
Moscow, Russia

Synopsis

Motivation for $\mathbb{C}$-valued neural networks

- perform better for naturally $\mathbb{C}$-valued data
- use half as much storage, but the same number of flops

Synopsis

Motivation for $\mathbb{C}$-valued neural networks

- perform better for naturally $\mathbb{C}$-valued data
- use half as much storage, but the same number of flops

Propose *Sparse Variational Dropout* for $\mathbb{C}$-valued neural networks

- Bayesian sparsification method with $\mathbb{C}$-valued distributions
- empirically explore the compression-performance trade-off

Synopsis

Motivation for $\mathbb{C}$-valued neural networks

- perform better for naturally $\mathbb{C}$-valued data
- use half as much storage, but the same number of flops

Propose *Sparse Variational Dropout* for $\mathbb{C}$-valued neural networks

- Bayesian sparsification method with $\mathbb{C}$-valued distributions
- empirically explore the compression-performance trade-off

Conclusions

- $\mathbb{C}$-valued methods compress similarly to $\mathbb{R}$-valued predecessors
- final performance benefits from fine-tuning sparsified network
- compress a SOTA $\mathbb{C}$VNN on MusicNet by $50 - 100\times$ at a moderate performance penalty

$\mathbb{C}$-valued neural networks: Applications

Data with natural $\mathbb{C}$-valued representation

- radar and satellite imaging

  [Hirose, 2009, Hänsch and Hellwich, 2010, Zhang et al., 2017]

- magnetic resonance imaging

  [Hui and Smith, 1995, Wang et al., 2020]

- radio signal classification

  [Yang et al., 2019, Tarver et al., 2019]

- spectral speech modelling and music transcription

  [Wisdom et al., 2016, Trabelsi et al., 2018, Yang et al., 2019]

$\mathbb{C}$-valued neural networks: Applications

Data with natural $\mathbb{C}$-valued representation

- radar and satellite imaging

  [Hirose, 2009, Hänsch and Hellwich, 2010, Zhang et al., 2017]

- magnetic resonance imaging

  [Hui and Smith, 1995, Wang et al., 2020]

- radio signal classification

  [Yang et al., 2019, Tarver et al., 2019]

- spectral speech modelling and music transcription

  [Wisdom et al., 2016, Trabelsi et al., 2018, Yang et al., 2019]

Exploring benefits beyond $\mathbb{C}$-valued data

- sequence modelling, dynamical system identification

  [Danihelka et al., 2016, Wisdom et al., 2016]

- image classification, road / lane segmentation

  [Popa, 2017, Trabelsi et al., 2018, Gaudet and Maida, 2018]

- unitary transition matrices in recurrent networks

  [Arjovsky et al., 2016, Wisdom et al., 2016]

$\mathbb{C}$-valued neural networks: Implementation

Geometric representation $\mathbb{C} \simeq \mathbb{R}^2$

- $z = \Re z + \jmath \Im z$, $\jmath^2 = -1$
- $\Re z$ and $\Im z$ are **real** and **imaginary** parts of z

An intricate double-$\mathbb{R}$ network that respects $\mathbb{C}$-arithmetic

$$\left[\begin{array}{c:c} W_{11} & W_{12} \\ \hdashline W_{21} & W_{22} \end{array}\right] \times \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \left[\begin{array}{c:c} W_{\mathrm{re}} & -W_{\mathrm{im}} \\ \hdashline W_{\mathrm{im}} & W_{\mathrm{re}} \end{array}\right] \times \begin{bmatrix} \Re z \\ \Im z \end{bmatrix}$$

$\mathbb{R}$VNN linear operation $\mathbb{C}$VNN linear operation

Activations $z \mapsto \sigma(z)$, e.g $re^{\jmath\phi} \mapsto \sigma(r, \phi)$ or $z \mapsto \sigma(\Re z) + \jmath\sigma(\Im z)$.

Sparsity and compression

Improve power, storage or throughput efficiency of deep nets

- Knowledge distillation

  [Hinton et al., 2015, Balasubramanian, 2016]

- Network pruning

  [LeCun et al., 1990, Seide et al., 2011, Zhu and Gupta, 2018]

- Low-rank matrix / tensor decomposition

  [Denton et al., 2014, Novikov et al., 2015]

- Quantization and fixed point arithmetic

  [Courbariaux et al., 2015, Han et al., 2016, Chen et al., 2017]

Applications to $\mathbb{C}$VNN:

- $\mathbb{C}$ modulus pruning, quantization with k-means in $\mathbb{R}^2$,

  [Wu et al., 2019]

- ℓ_1 regularization for hyper-complex-valued networks,

  [Vecchi et al., 2020]

Sparse Variational Dropout

[Molchanov et al., 2017]

Variational Inference with automatic relevance determination effect

$$\underset{q \in \mathcal{Q}}{\text{maximize}} \quad \underbrace{\mathbb{E}_{w \sim q} \log p(D \mid w)}_{\text{data model likelihood}} \quad - \quad \underbrace{KL(q \| \pi)}_{\text{variational regularization}} \qquad \text{(ELBO)}$$

prior π $\rightarrow$ data model likelihood $\rightarrow$ posterior q (close to $p(w \mid D)$)

Sparse Variational Dropout

[Molchanov et al., 2017]

Variational Inference with automatic relevance determination effect

$$\max_{q \in \mathcal{Q}} \quad \underbrace{\mathbb{E}_{w \sim q} \log p(D \mid w)}_{\text{data model likelihood}} \quad - \quad \underbrace{KL(q \| \pi)}_{\text{variational regularization}} \qquad \text{(ELBO)}$$

prior π $\to$ data model likelihood $\to$ posterior q (close to $p(w \mid D)$)

Factorized Gaussian dropout posterior family $\mathcal{Q}$

- $w_{ij} \sim q(w_{ij}) = \mathcal{N}(w_{ij} \mid \mu_{ij}, \alpha_{ij} \mu_{ij}^2)$, $\alpha_{ij} > 0$, and $\mu_{ij} \in \mathbb{R}$

Factorized prior

- (VD) $\pi(w_{ij}) \propto \frac{1}{|w_{ij}|}$ [Molchanov et al., 2017]
- (ARD) $\pi(w_{ij}) = \mathcal{N}(w_{ij} \mid 0, \frac{1}{\tau_{ij}})$ [Kharitonov et al., 2018]

$\mathbb{C}$-valued Variational Dropout

Our proposal

Factorized complex-valued posterior $q(w) = \prod q(w_{ij})$

- w_{ij} are *independent* $\mathcal{CN}(w \,\big|\, \mu, \sigma^2, \sigma^2\xi)$, $\sigma^2 = \alpha|\mu|^2$, $|\xi| \leq 1$

$$\begin{pmatrix} \Re w \\ \Im w \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \Re\mu \\ \Im\mu \end{pmatrix}, \frac{\sigma^2}{2} \begin{pmatrix} 1 + \Re\xi & \Im\xi \\ \Im\xi & 1 - \Re\xi \end{pmatrix} \right)$$

$\mathbb{C}$-valued Variational Dropout

Our proposal

Factorized complex-valued posterior $q(w) = \prod q(w_{ij})$

- w_{ij} are *independent* $\mathcal{CN}(w \,\big|\, \mu, \sigma^2, \sigma^2 \xi)$, $\sigma^2 = \alpha |\mu|^2$, $|\xi| \leq 1$

$$\begin{pmatrix} \Re w \\ \Im w \end{pmatrix} \sim \mathcal{N}_2\Biggl(\begin{pmatrix} \Re \mu \\ \Im \mu \end{pmatrix}, \frac{\sigma^2}{2} \begin{pmatrix} 1 + \Re \xi & \Im \xi \\ \Im \xi & 1 - \Re \xi \end{pmatrix} \Biggr)$$

- w_{ij} are *circularly symmetric* about μ_{ij} ($\xi_{ij} = 0$)
- relevance $\propto \frac{1}{\alpha_{ij}}$ and $\frac{2|w_{ij} - \mu_{ij}|^2}{\alpha_{ij} |\mu_{ij}|^2}$ is χ^2_2

$\mathbb{C}$-valued Variational Dropout

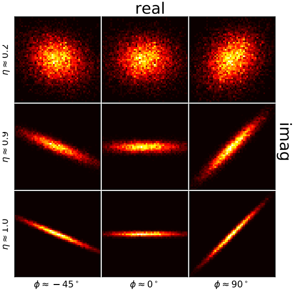
Our proposal

Factorized complex-valued posterior $q(w) = \prod q(w_{ij})$

- w_{ij} are *independent* $\mathcal{CN}(w \,|\, \mu, \sigma^2, \sigma^2\xi)$, $\sigma^2 = \alpha|\mu|^2$, $|\xi| \leq 1$

$$\begin{pmatrix} \Re w \\ \Im w \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \Re\mu \\ \Im\mu \end{pmatrix}, \frac{\sigma^2}{2} \begin{pmatrix} 1 + \Re\xi & \Im\xi \\ \Im\xi & 1 - \Re\xi \end{pmatrix} \right)$$

- w_{ij} are *circularly symmetric* about μ_{ij} $(\xi_{ij} = 0)$
- relevance $\propto \frac{1}{\alpha_{ij}}$ and $\frac{2|w_{ij} - \mu_{ij}|^2}{\alpha_{ij}|\mu_{ij}|^2}$ is χ^2_2



$\mathcal{CN}(0, 1, \eta e^{\jmath\phi})$, $|\eta| \leq 1$

Factorized complex-valued priors π

- ($\mathbb{C}$-VD) $\pi(w_{ij}) \propto |w_{ij}|^{-\rho}$, $\rho \geq 1$
- ($\mathbb{C}$-ARD) $\pi(w_{ij}) = \mathcal{CN}(0, \frac{1}{\tau_{ij}}, 0)$

$\mathbb{C}$-valued Variational Dropout

$KL(q\|\pi)$ term in (ELBO)

$$KL(q\|\pi) = \sum_{ij} KL(q(w_{ij})\|\pi(w_{ij}))$$

($\mathbb{C}$-VD) improper prior

$$\begin{aligned} KL_{ij} &\propto \tfrac{\rho-2}{2}\log|\mu_{ij}|^2 + \log\tfrac{1}{\alpha_{ij}} - \tfrac{\rho}{2}Ei(-\tfrac{1}{\alpha_{ij}}) \\ Ei(x) &= \int_{-\infty}^{x} e^t t^{-1} dt \end{aligned}$$

($\mathbb{C}$-ARD) prior is optimized w.r.t. τ_{ij} in empirical Bayes

$$\begin{aligned} KL_{ij} &= -1 - \log\sigma_{ij}^2\tau_{ij} + \tau_{ij}(\sigma_{ij}^2 + |\mu_{ij}|^2) \\ \min_{\tau_{ij}} KL_{ij} &= \log\big(1 + \tfrac{1}{\alpha_{ij}}\big) \end{aligned}$$

Experiments: Goals and Setup

We conduct numerous experiments on various datasets to

- validate the proposed $\mathbb{C}$-valued sparsification methods
- explore the compression-performance profiles
- compare to the $\mathbb{R}$-valued Sparse Variational Dropout

Experiments: Goals and Setup

We conduct numerous experiments on various datasets to

- validate the proposed $\mathbb{C}$-valued sparsification methods
- explore the compression-performance profiles
- compare to the $\mathbb{R}$-valued Sparse Variational Dropout

'pre-train' $\to$ 'compress' $\to$ 'fine-tune'

- 'compress' with $\mathbb{R}/\mathbb{C}$-Variational Dropout layers
- 'fine-tune' pruned network ($\log \alpha_{ij} \leq -\frac{1}{2}$)

Experiments: Goals and Setup

We conduct numerous experiments on various datasets to

- validate the proposed $\mathbb{C}$-valued sparsification methods
- explore the compression-performance profiles
- compare to the $\mathbb{R}$-valued Sparse Variational Dropout

'pre-train' $\rightarrow$ 'compress' $\rightarrow$ 'fine-tune'

- 'compress' with $\mathbb{R}/\mathbb{C}$-Variational Dropout layers
- 'fine-tune' pruned network ($\log \alpha_{ij} \leq -\frac{1}{2}$)

$$\max_{q} \mathbb{E}_{w \sim q} \log p(D \mid w) - \beta \, KL(q \| \pi) \qquad (\beta\text{-ELBO})$$

Experiments: Datasets

Four MNIST-like datasets

- channel features ($\mathbb{R} \hookrightarrow \mathbb{C}$) or 2d Fourier features
- fixed random subset of $10k$ train samples
- simple dense and convolutional nets

Experiments: Datasets

Four MNIST-like datasets

- channel features ($\mathbb{R} \hookrightarrow \mathbb{C}$) or 2d Fourier features
- fixed random subset of $10k$ train samples
- simple dense and convolutional nets

CIFAR10 dataset ($\mathbb{R}^3 \hookrightarrow \mathbb{C}^3$)

- random cropping and horizontal flipping
- $\mathbb{C}$-valued variant of VGG16 [Simonyan and Zisserman, 2015]

Experiments: Datasets

Four MNIST-like datasets

- channel features ($\mathbb{R} \hookrightarrow \mathbb{C}$) or 2d Fourier features
- fixed random subset of $10k$ train samples
- simple dense and convolutional nets

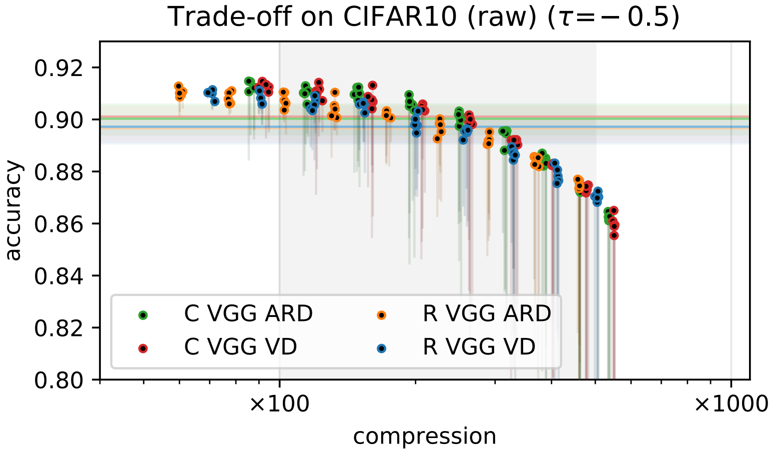
CIFAR10 dataset ($\mathbb{R}^3 \hookrightarrow \mathbb{C}^3$)

- random cropping and horizontal flipping
- $\mathbb{C}$-valued variant of VGG16 [Simonyan and Zisserman, 2015]

Music transcription on MusicNet [Thickstun et al., 2017]

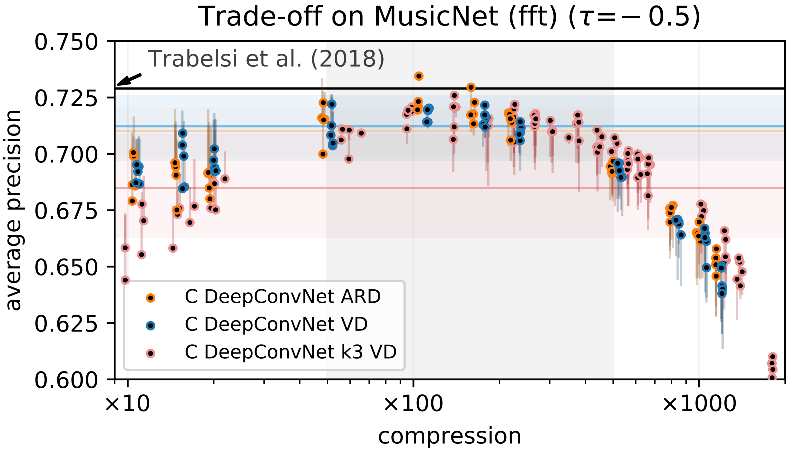
- audio dataset of 330 annotated musical compositions
- use power spectrum to tell which piano keys are pressed
- compress deep $\mathbb{C}$VNN proposed by [Trabelsi et al., 2018]

Results: CIFAR10



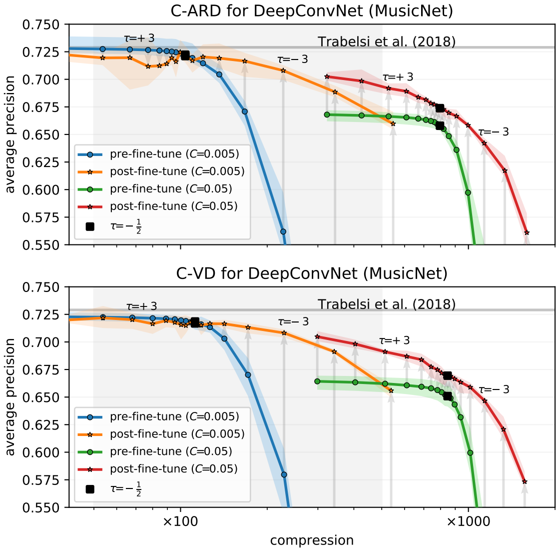
$\mathbb{C}$-valued version of VGG16 [Simonyan and Zisserman, 2015]

Results: MusicNet



The $\mathbb{C}$VNN of Trabelsi et al. [2018]

MusicNet: Effects of pruning threshold



Effect of threshold on the $\mathbb{C}$VNN of Trabelsi et al. [2018]

Summary: Results

Bayesian sparsification of $\mathbb{C}$-valued networks

- proposed $\mathbb{C}$-VD and $\mathbb{C}$-ARD methods

  `https://github.com/ivannz/cplxmodule`

- investigated performance-compression trade-off
- compressed the $\mathbb{C}$VNN of Trabelsi et al. [2018] by $50 - 100\times$

  `https://github.com/ivannz/complex_paper`

Summary: Results

Bayesian sparsification of $\mathbb{C}$-valued networks

- proposed $\mathbb{C}$-VD and $\mathbb{C}$-ARD methods
  `https://github.com/ivannz/cplxmodule`
- investigated performance-compression trade-off
- compressed the $\mathbb{C}$VNN of Trabelsi et al. [2018] by $50 - 100\times$
  `https://github.com/ivannz/complex_paper`

Experiments

- $\mathbb{C}$-VD and $\mathbb{C}$-ARD have trade-off similar to $\mathbb{R}$ methods
- $\mathbb{R}$-networks tend to compress better than $\mathbb{C}$-nets
- fine-tuning improves performance in high compression regime
- β in β-ELBO influences compression stronger than threshold

Summary: Limitations

Circular symmetry of the posterior $q(w_{ij})$ about μ_{ij} implies *independence* of $\Re$ and $\Im$

- modelling $corr(w_{ij}, \overline{w}_{ij})$ gives better variational approximation

Factorized q implies parameter independence

- structured sparsity is desirable for fast computations and hardware implementations

Extensions: PolARD dropout

Modelling correlation between weight and its conjugate

Use allow learnable relation $\xi = \eta e^{\jmath\phi}$ in factorized q

- w_{ij} are *independent* $\mathcal{CN}(w \,|\, \mu, \sigma^2, \sigma^2\xi)$, $\sigma^2 = \alpha|\mu|^2$
- $|\eta| \leq 1$, $\phi \in [-\frac{\pi}{2}, +\frac{\pi}{2}]$, and $\alpha = \frac{\sigma^2}{|\mu|^2}$

Extensions: PolARD dropout

Modelling correlation between weight and its conjugate

Use allow learnable relation $\xi = \eta e^{\jmath\phi}$ in factorized q

- w_{ij} are *independent* $\mathcal{CN}(w \,\big|\, \mu, \sigma^2, \sigma^2\xi)$, $\sigma^2 = \alpha|\mu|^2$
- $|\eta| \leq 1$, $\phi \in [-\frac{\pi}{2}, +\frac{\pi}{2}]$, and $\alpha = \frac{\sigma^2}{|\mu|^2}$

$\mathbb{C}$-ARD Prior $\pi = \mathcal{CN}(w \,\big|\, 0, \tau^{-1}, 0)$

$$\begin{aligned} KL(q\|\pi) &= -1 - \log \sigma^2\tau + \tau(\sigma^2 + |\mu|^2) - \frac{1}{2}\log\big(1 - |\eta|^2\big)\,, \\ \min_{\tau} KL &= \log\Big(1 + \frac{1}{\alpha}\Big) - \frac{1}{2}\log\big(1 - |\eta|^2\big)\,. \end{aligned}$$

Extensions: PolARD dropout

For $y = w^\top x + b$ with $w \in \mathbb{C}^{n\times m}$, $w \sim q$ implies

$$y_i \sim \mathcal{CN}\left(b_i + \sum_j \mu_{ij} x_j, \sum_j |x_{ij}|^2 \sigma_{ij}^2, \sum_j x_{ij}^2 \sigma_{ij}^2 \eta_{ij} e^{\jmath \phi_{ij}} \right).$$

Extensions: PolARD dropout

For $y = w^\top x + b$ with $w \in \mathbb{C}^{n \times m}$, $w \sim q$ implies

$$y_i \sim \mathcal{CN}\biggl(b_i + \sum_j \mu_{ij} x_j, \sum_j |x_{ij}|^2 \sigma_{ij}^2, \sum_j x_{ij}^2 \sigma_{ij}^2 \eta_{ij} e^{\jmath \phi_{ij}}\biggr) .$$

For $\xi \in \mathbb{C}$, $|\xi| \leq 1$, $\rho = \sqrt{1 - |\xi|^2}$,

$$R = \frac{\sigma}{2\sqrt{1+\rho}} \begin{pmatrix} (1+\rho) + \Re\xi & \Im\xi \\ \Im\xi & (1+\rho) - \Re\xi \end{pmatrix} ,$$

and $\varepsilon \sim \mathcal{N}_2(0, I_2)$ we have

$$(R_{11}\varepsilon_1 + R_{12}\varepsilon_2) + \jmath(R_{21}\varepsilon_1 + R_{22}\varepsilon_2) \overset{\mathcal{D}}{\sim} \mathcal{CN}(0, \sigma^2, \sigma^2\xi) .$$