

Neural Networks in the Air: How to Train Your Dragon

Deniz Gündüz (Imperial College London)
joint with M. Jankowski, Y. Shao, K. Mikolajczyk, S. C. Liew

Workshop on Machine Learning for Communications (MLCOM)
24 March 2022

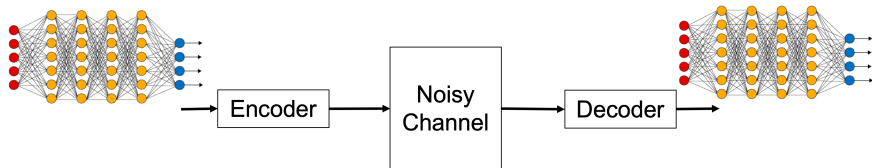
This work received support from European Research Council through grant BEACON
and from the UK EPSRC through CHIST-ERA project CONNECT.

Neural Networks on Demand



- Many deep neural networks (DNNs) are time or location specific
- Users can't store all possible DNNs they may need
- **DNN on demand**: Download DNN at the time of inference
- Federated edge learning: Training over the air
- Storage of DNN parameters in memory
- **DNNs are huge**: VGG-16 has 138 million parameters (>500 MB)

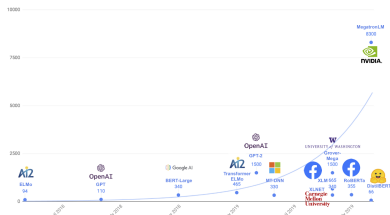
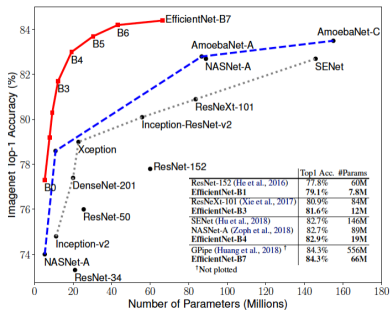
AirNet: Neural Networks in the Air



- Model parameters transmitted over a noisy channel
- Classification done with reconstructed DNN
- A joint source-channel coding problem: Goal is to reconstruct a model with high accuracy

M. Jankowski, D. Gündüz, and K. Mikolajczyk, "AirNet: Neural network transmission over the air," 2021.

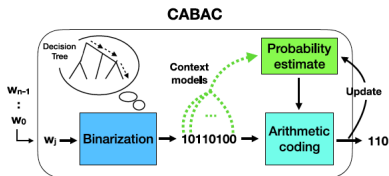
Conventional Approach: Compress and Transmit



- Weight sharing
- Network pruning
- Tensor decomposition
- Knowledge distillation
- Quantization

Figures from Tan and Le (2019), Sanh et al. (2019).

Universal DNN Compression



- **MPEG-7 Part 17**: Compression of Neural Networks for Multimedia Content Description and Analysis
- Context-based Adaptive Binary Arithmetic Coding (CABAC)
- No retraining required

S. Wiedemann et al., "DeepCABAC: A Universal Compression Algorithm for Deep Neural Networks," in IEEE Journal of Selected Topics in Signal Processing, May 2020.

Uncoded/Analog Transmission of DNNs

- **Conventional approach:** Compress NN weights, use channel coding against errors
 - Requires accurate channel estimation
 - Channel encoding/decoding can be time consuming
 - Suffers from the ‘cliff effect’
- **Proposed approach:** Analog transmission of DNNs
 - Recent success of end-to-end joint source-channel coding solutions for image and video transmission
- Challenges:
 - How to do compression in analog domain?
 - How to make model robust against noise?
 - How to introduce error correction?

E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” *IEEE Trans. on Cognitive Comms. and Networking*, Sep. 2019.

T.-Y. Tung and D. Gunduz, “DeepWiVe: Deep-learning-aided wireless video transmission,” submitted, Nov. 2021.

B. Isik, K. Choi, X. Zheng, T. Weissman, S. Ermon, H.-S. P. Wong, A. Alaghi, “Neural Network Compression for Noisy Storage Devices,” 2020.

Uncoded/Analog Transmission of DNNs

- **Conventional approach:** Compress NN weights, use channel coding against errors
 - Requires accurate channel estimation
 - Channel encoding/decoding can be time consuming
 - Suffers from the ‘cliff effect’
- **Proposed approach:** Analog transmission of DNNs
 - Recent success of end-to-end joint source-channel coding solutions for image and video transmission
- **Challenges:**
 - How to do compression in analog domain?
 - How to make model robust against noise?
 - How to introduce error correction?

E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” *IEEE Trans. on Cognitive Comms. and Networking*, Sep. 2019.

T.-Y. Tung and D. Gunduz, “DeepWiVe: Deep-learning-aided wireless video transmission,” submitted, Nov. 2021.

B. Isik, K. Choi, X. Zheng, T. Weissman, S. Ermon, H.-S. P. Wong, A. Alaghi, “Neural Network Compression for Noisy Storage Devices,” 2020.

Uncoded/Analog Transmission of DNNs

- **Conventional approach:** Compress NN weights, use channel coding against errors
 - Requires accurate channel estimation
 - Channel encoding/decoding can be time consuming
 - Suffers from the ‘cliff effect’
- **Proposed approach:** Analog transmission of DNNs
 - Recent success of end-to-end joint source-channel coding solutions for image and video transmission
- **Challenges:**
 - How to do compression in analog domain?
 - How to make model robust against noise?
 - How to introduce error correction?

E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” *IEEE Trans. on Cognitive Comms. and Networking*, Sep. 2019.

T.-Y. Tung and D. Gunduz, “DeepWiVe: Deep-learning-aided wireless video transmission,” submitted, Nov. 2021.

B. Isik, K. Choi, X. Zheng, T. Weissman, S. Ermon, H.-S. P. Wong, A. Alaghi, “Neural Network Compression for Noisy Storage Devices,” 2020.

Uncoded/Analog Transmission of DNNs

- **Conventional approach:** Compress NN weights, use channel coding against errors
 - Requires accurate channel estimation
 - Channel encoding/decoding can be time consuming
 - Suffers from the ‘cliff effect’
- **Proposed approach:** Analog transmission of DNNs
 - Recent success of end-to-end joint source-channel coding solutions for image and video transmission
- **Challenges:**
 - How to do compression in analog domain?
 - How to make model robust against noise?
 - How to introduce error correction?

E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” IEEE Trans. on Cognitive Comms. and Networking, Sep. 2019.

T.-Y. Tung and D. Gunduz, “DeepWiVe: Deep-learning-aided wireless video transmission, submitted, Nov. 2021.

B. Isik, K. Choi, X. Zheng, T. Weissman, S. Ermon, H.-S. P. Wong, A. Alaghi, “Neural Network Compression for Noisy Storage Devices,” 2020.

Noise Injection to Neural Networks

- Deep neural networks (DNNs) are often over-parametrized and suffer from overfitting
- Proper regularization is essential for better generalization
- **Inject noise during training:** dropout
- Proposed approach: **Pruning** (for bandwidth reduction) + **noise injection** during training (for robustness) + **knowledge distillation** (for higher accuracy)

Srivastava et al., "Dropout: A simple way to prevent neural networks from overfitting," The Journal of Machine Learning Research, 2014.

Noh et al., "Regularizing deep neural networks by noise: its interpretation and optimization," NeurIPS 2017.

Noise Injection to Neural Networks

- Deep neural networks (DNNs) are often over-parametrized and suffer from overfitting
- Proper regularization is essential for better generalization
- Inject noise during training: dropout
- Proposed approach: Pruning (for bandwidth reduction) + noise injection during training (for robustness) + knowledge distillation (for higher accuracy)

Srivastava et al., “Dropout: A simple way to prevent neural networks from overfitting,” The Journal of Machine Learning Research, 2014.

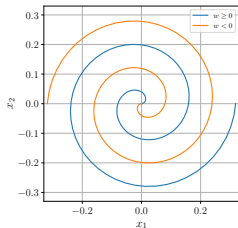
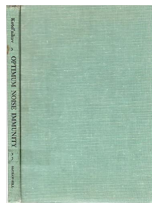
Noh et al., “Regularizing deep neural networks by noise: its interpretation and optimization,” NeurIPS 2017.

Analog Error Correction

- Can we achieve error **reduction** for sensitive/important weights?
- Power / bandwidth allocation
- **Shannon-Kotelnikov mapping**:

$$x_1 = \frac{\Delta}{\pi} w \cos(\gamma w), \quad x_2 = \frac{\Delta}{\pi} w \sin(\gamma w), \quad w \geq 0$$

$$x_1 = -\frac{\Delta}{\pi} w \cos(-\gamma w + \pi), \quad x_2 = -\frac{\Delta}{\pi} w \sin(-\gamma w + \pi), \quad w < 0,$$

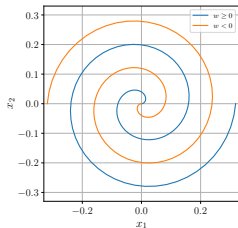
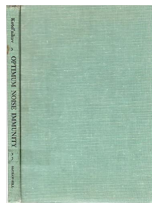


Analog Error Correction

- Can we achieve error **reduction** for sensitive/important weights?
- Power / bandwidth allocation
- **Shannon-Kotelnikov mapping**:

$$x_1 = \frac{\Delta}{\pi} w \cos(\gamma w), \quad x_2 = \frac{\Delta}{\pi} w \sin(\gamma w), \quad w \geq 0$$

$$x_1 = -\frac{\Delta}{\pi} w \cos(-\gamma w + \pi), \quad x_2 = -\frac{\Delta}{\pi} w \sin(-\gamma w + \pi), \quad w < 0,$$

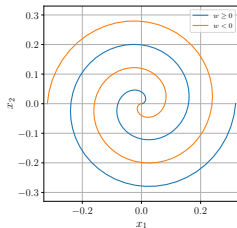
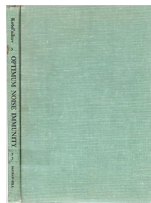


Analog Error Correction

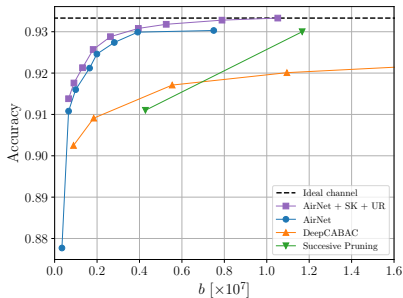
- Can we achieve error **reduction** for sensitive/important weights?
- Power / bandwidth allocation
- **Shannon-Kotelnikov mapping**:

$$x_1 = \frac{\Delta}{\pi} w \cos(\gamma w), \quad x_2 = \frac{\Delta}{\pi} w \sin(\gamma w), \quad w \geq 0$$

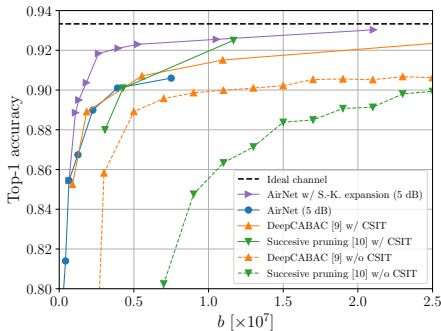
$$x_1 = -\frac{\Delta}{\pi} w \cos(-\gamma w + \pi), \quad x_2 = -\frac{\Delta}{\pi} w \sin(-\gamma w + \pi), \quad w < 0$$



DNNs in the Air



AWGN, SNR = 5 dB



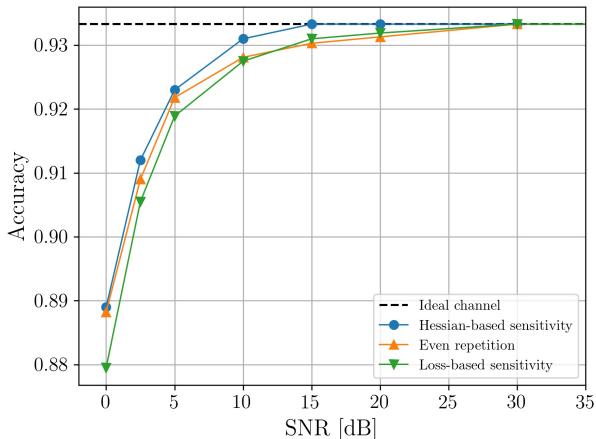
Fading channel, Average SNR = 5 dB

- Small-VGG16 for CIFAR-10 classification
- Observation: Better to prune more, and then introduce redundancy through SK mapping

Unequal Error Protection

- Some weights are more important/ sensitive to noise than others
- **How to do unequal error protection in analog domain?**
 - Choose layers according to noise sensitivity
 - Use second-order information

AirNet Bandwidth Allocation



AWGN, SNR = 5dB, $b \sim 0.65 \times 10^6$.

- Noise during inference
- Noise during training

Let $\mathbf{r} \in \mathbb{R}^d$ be the DNN parameters, received with additive noise (often Gaussian):

$$\mathbf{r} = \mathbf{w} + \mathbf{z}$$

- ML estimation: $\hat{\mathbf{w}}^{\text{ML}} = \mathbf{r}$

- Assume Gaussian prior on \mathbf{w} : $\mathcal{W} \sim \mathcal{N}(\mathcal{W}; \mu_w, \sigma_w^2)$, where μ_w and σ_w^2 are the sample mean and sample variance of $\mathbf{w} = \{w[i] : i = 1, 2, \dots, d\}$:

$$\mu_w = \frac{1}{d} \sum_{i=1}^d w[i], \quad \sigma_w^2 = \frac{1}{d} \sum_{i=1}^d (w[i] - \mu_w)^2.$$

- MMSE estimation:** Given $\mathbf{r} \in \mathbb{R}^d$,

$$\hat{\mathbf{w}}^{\text{MMSE}} = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_z^2} \mathbf{r} + \frac{\mu_w \sigma_z^2}{\sigma_w^2 + \sigma_z^2} \mathbf{e},$$

where \mathbf{e} is an all-ones vector.

- MMSE estimate is also the MAP estimate.
- But, the goal is to maximize inference accuracy, not MSE.

Bayesian Denoiser with Compensators

- Most parameter values in a DNN are very small in magnitude
- Larger parameters matter more than smaller ones for accuracy
- Minimize

$$\text{MSE}_{pb} = \mathbb{E} \left[\left(\hat{\mathcal{W}} - \mathcal{W} \right)^2 e^{\lambda \mathcal{W}^2 + \beta \mathcal{W}} \middle| \mathcal{R} \right]$$

λ, β : temperature parameters

$$\hat{\mathcal{w}}^{\text{MMSE}_{pb}} = \frac{\sigma_w^2}{\sigma_w^2 + (1 - 2\sigma_w^2 \lambda) \sigma_z^2} \mathbf{r} + \frac{\sigma_w^2 \sigma_z^2 \beta}{\sigma_w^2 + (1 - 2\sigma_w^2 \lambda) \sigma_z^2} \mathbf{e},$$

where $0 \leq \lambda < \frac{1}{2\sigma_w^2} + \frac{1}{2\sigma_z^2}$.

Bayesian Denoiser with Compensators

- Most parameter values in a DNN are very small in magnitude
- Larger parameters matter more than smaller ones for accuracy
- Minimize

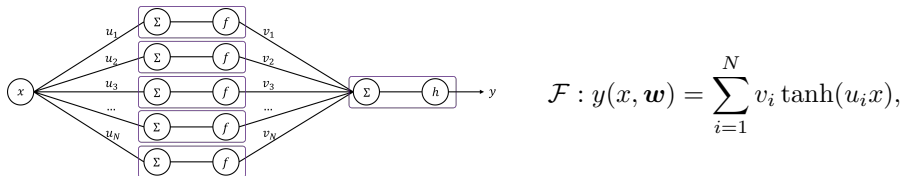
$$\text{MSE}_{pb} = \mathbb{E} \left[\left(\hat{\mathcal{W}} - \mathcal{W} \right)^2 e^{\lambda \mathcal{W}^2 + \beta \mathcal{W}} \middle| \mathcal{R} \right]$$

λ, β : temperature parameters

$$\hat{\mathbf{w}}^{\text{MMSE}_{pb}} = \frac{\sigma_w^2}{\sigma_w^2 + (1 - 2\sigma_w^2 \lambda) \sigma_z^2} \mathbf{r} + \frac{\sigma_w^2 \sigma_z^2 \beta}{\sigma_w^2 + (1 - 2\sigma_w^2 \lambda) \sigma_z^2} \mathbf{e},$$

where $0 \leq \lambda < \frac{1}{2\sigma_w^2} + \frac{1}{2\sigma_z^2}$.

Simple Three-Layer Network



- After denoising, neural network output can be written as

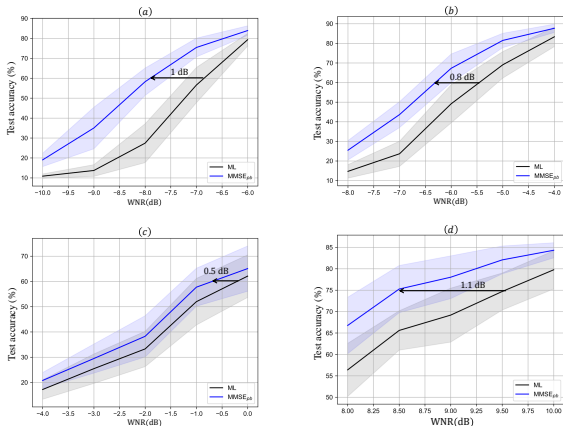
$$\tilde{y}(x, \mathbf{w}, \mathbf{z}, \theta(\lambda), \rho(\lambda, \beta)) = \sum_{i=1}^N [\theta v_i + \theta \Delta v_i + \rho] \tanh(\theta u_i x + \theta \Delta u_i x + \rho x),$$

$\theta(\lambda), \rho(\lambda, \beta)$: multiplicative and additive factors in MMSE_{pb} .

- For $x \sim U[-c, c], u_i, v_i \sim \mathcal{N}(0, \sigma_{\mathbf{w}}^2)$, gain w.r.t. ML estimation in average output error:

$$\frac{\bar{\mathcal{D}}^{\text{ML}} - \bar{\mathcal{D}}^{\text{MMSE}_{pb}}}{\bar{\mathcal{D}}^{\text{ML}}} \approx \frac{2\sigma_{\mathbf{z}}^2}{\sigma_{\mathbf{w}}^2 + 2\sigma_{\mathbf{z}}^2}.$$

Bayesian Denoiser with Compensators

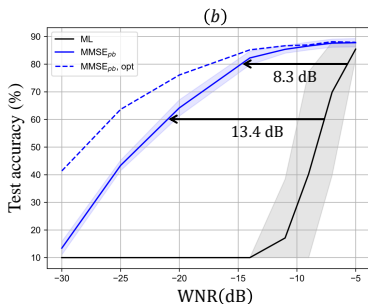
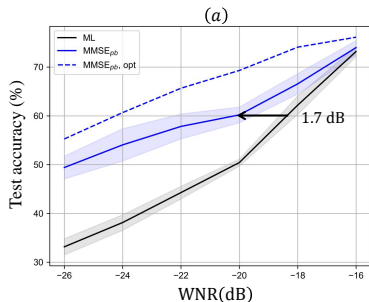


(a) RestNet34 (CIFAR-10); (b) RestNet18 (CIFAR-10); (c) ShuffleNet V2 (CIFAR-10); (d) BERT (SST-2).

WNR: weight variance to noise power ratio

Y. Shao, S. C. Liew, D. Gunduz, "Denoising noisy neural networks: A Bayesian approach with compensation," arXiv 2021.

Training with a Bayesian Denoiser



- Federated edge learning across 20 wireless devices
- CIFAR-10 training with ShuffleNet V2 (a) and ResNet18 (b)

Y. Shao, S. C. Liew, D. Gunduz, "Denoising noisy neural networks: A Bayesian approach with compensation," arXiv 2021.

Thank You!

For more information:

www.imperial.ac.uk/ipc-lab

Call for papers:

IEEE Journal on Selected Areas in Communications , Special Issue on
“Beyond Transmitting Bits: Context, Semantics and Task-Oriented
Communications”

Deadline April 1st, 2022