interdigital

Low Rank Based End-to-End Deep Neural Network Compression

Swayambhoo Jain, Shahab Hamidi-Rad, and Fabien Racape
InterDigital AI Research Lab
Los Altos, CA, 94022, USA
DCC 2021

Why compress Neural Networks

- **Deep Neural Networks Require:**
  - High Computation Power (Machines equipped with multiple GPU cards)
  - Lots of memory (Both to store the model and to load it for inference)
  - High bandwidth to transfer models to target machines.

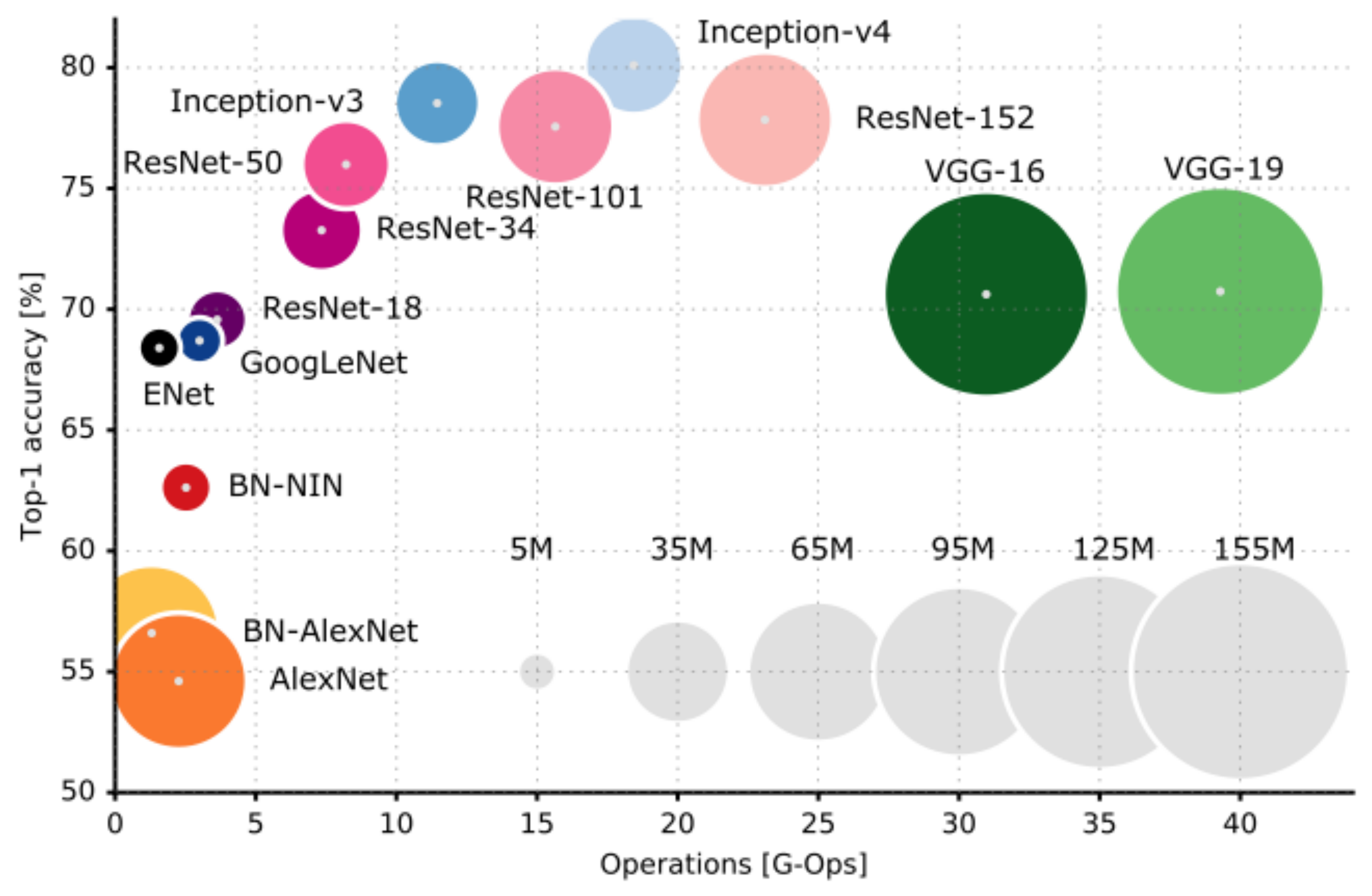
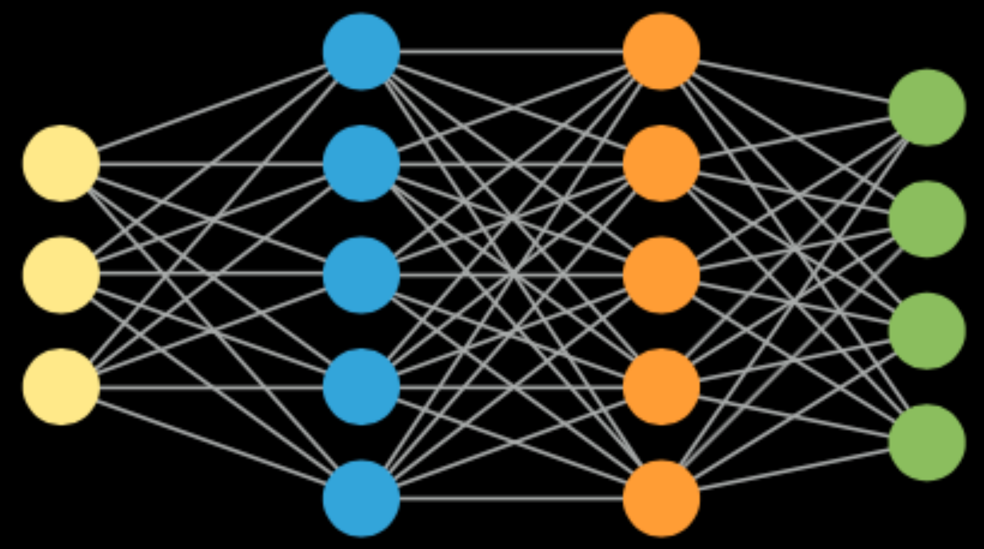


Image Source: [AN ANALYSIS OF DEEP NEURAL NETWORK MODELS FOR PRACTICAL APPLICATIONS] - Canziani, A., Paszke, A., & Culurciello, E. (2016).

Overview of DNN Compression Pipeline

Original Model



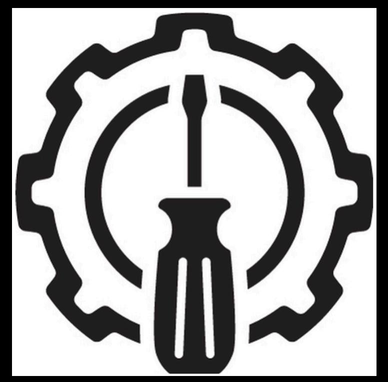
Reduce Number of Parameters

Structural Approximation using
Low Rank Decomposition

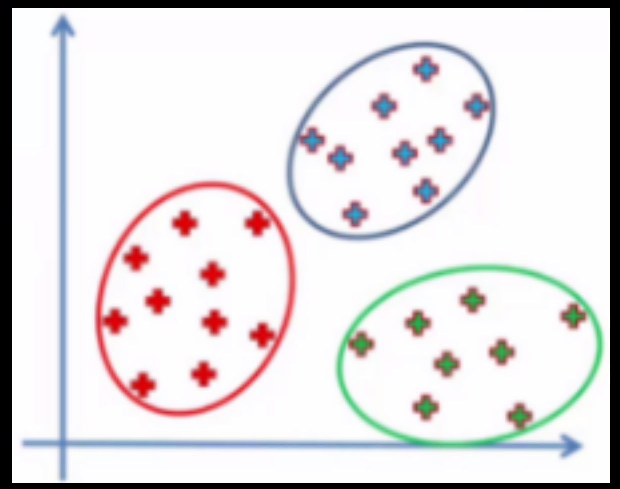
$$W_{mxn} = G_{mxr} \, H^{T}_{rxn}$$

Fine-Tuning

(Retraining the whole model)

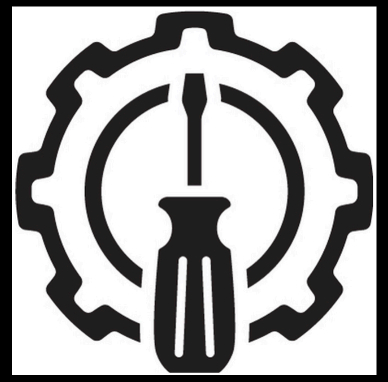


K-Means Quantization



Fine-Tuning

(Retraining the Quantized model)



Lossless compression

(Arithmetic Coding)

$$H = -\sum p(x)\log p(x)$$

Problem Definition

$$\mathcal{W}^{\text{pre}} = \left\{ \mathbf{W}_l^{\text{pre}} \in \mathbb{R}^{n_l \times n_{l-1}}, \mathbf{b}_l^{\text{pre}} \in \mathbb{R}^{n_l} \right\}_{l=1}^{L}$$

$$\frac{1}{N^{\text{tr}}} \sum_{i=1}^{N^{\text{tr}}} \ell\left(\mathbf{y}_i, f\left(\mathbf{x}_i; \mathcal{W}\right)\right)$$

$$\min_{\mathcal{W}=\{\mathbf{W}_l, \mathbf{b}_l\}_{l=1}^{L}} \text{size}\left(\mathcal{W}\right), \ \text{s.t.} \ \mathbb{E}\left[\ell\left(f\left(\mathbf{x}; \mathcal{W}\right), \mathbf{y}\right)\right] \leq \epsilon$$

Structural Approximation (Low-Rank Decomposition)

$$\min_{\mathbf{U}_l \in \mathbb{R}^{n^l \times r^l}, \mathbf{V}_l \in \mathbb{R}^{r^l \times n^{l-1}}} \|\mathbf{W}_l^{\text{pre}} - \mathbf{U}_l \mathbf{V}_l\|_F^2$$

$$n^l . n^{l-1} \quad \rightarrow \quad r^l(n^l + n^{l-1})$$

Low-Rank Decomposition

$$\underbrace{\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n-1} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n-1} \\ \vdots & \vdots & \vdots & \vdots \\ w_{m-1,0} & w_{m-1,1} & \cdots & w_{m-1,n-1} \end{bmatrix}}_{\mathrm{W_{mxn}}} = \underbrace{\begin{bmatrix} u_{0,0} & u_{0,1} & \cdots & u_{0,m-1} \\ u_{1,0} & u_{1,1} & \cdots & u_{1,m-1} \\ \vdots & \vdots & \vdots & \vdots \\ u_{m-1,0} & u_{m-1,1} & \cdots & u_{m-1,m-1} \end{bmatrix}}_{\mathrm{U_{mxm}}} \underbrace{\begin{bmatrix} \sigma_0 & 0 & . & . & . & 0 \\ 0 & \sigma_1 & . & . & . & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{m-1} & \cdots & 0 \end{bmatrix}}_{\Sigma_{\mathrm{mxn}}} \underbrace{\begin{bmatrix} v_{0,0} & v_{0,1} & \cdots & v_{0,n-1} \\ v_{1,0} & v_{1,1} & \cdots & v_{1,n-1} \\ \vdots & \vdots & \vdots & \vdots \\ v_{n-1,0} & v_{n-1,1} & \cdots & v_{n-1,n-1} \end{bmatrix}}_{\mathrm{V^T{}_{nxn}}}$$

$$\underbrace{\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n-1} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n-1} \\ \vdots & \vdots & \vdots & \vdots \\ w_{m-1,0} & w_{m-1,1} & \cdots & w_{m-1,n-1} \end{bmatrix}}_{\mathrm{W_{mxn}}} \approx \underbrace{\begin{bmatrix} u_{0,0} & u_{0,1} & \cdots & u_{0,r-1} \\ u_{1,0} & u_{1,1} & \cdots & u_{1,r-1} \\ \vdots & \vdots & \vdots & \vdots \\ u_{m-1,0} & u_{m-1,1} & \cdots & u_{m-1,r-1} \end{bmatrix}}_{\mathrm{U_{mxr}}} \underbrace{\begin{bmatrix} \sigma_0 & 0 & . & . \\ 0 & \sigma_1 & . & . \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{r-1} \end{bmatrix}}_{\Sigma_{\mathrm{rxr}}} \underbrace{\begin{bmatrix} v_{0,0} & v_{0,1} & \cdots & v_{0,n-1} \\ v_{1,0} & v_{1,1} & \cdots & v_{1,n-1} \\ \vdots & \vdots & \vdots & \vdots \\ v_{r-1,0} & v_{r-1,1} & \cdots & v_{r-1,n-1} \end{bmatrix}}_{\mathrm{V^T{}_{rxn}}}$$

Low-Rank Decomposition

$$\Sigma_{\mathrm{rxr}}$$

$$\begin{bmatrix} \sigma_0 & 0 & . & . \\ 0 & \sigma_1 & . & . \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma_{r-1} \end{bmatrix}$$

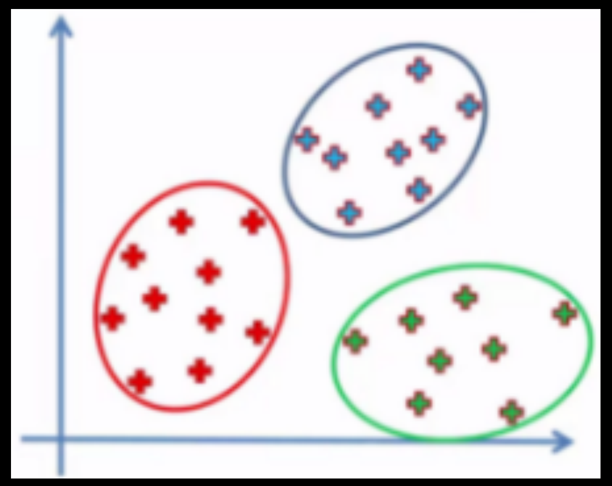
$$\begin{bmatrix} w_{0,0} & w_{0,1} & \dots & w_{0,n-1} \\ w_{1,0} & w_{1,1} & \dots & w_{1,n-1} \\ \vdots & \vdots & \vdots & \vdots \\ w_{m-1,0} & w_{m-1,1} & \dots & w_{m-1,n-1} \end{bmatrix} \approx \underbrace{\begin{bmatrix} u_{0,0} & u_{0,1} & \dots & u_{0,r-1} \\ u_{1,0} & u_{1,1} & \dots & u_{1,r-1} \\ \vdots & \vdots & \vdots & \vdots \\ u_{m-1,0} & u_{m-1,1} & \dots & u_{m-1,r-1} \end{bmatrix} \begin{bmatrix} \sqrt{\sigma_0} & 0 & \dots & 0 \\ 0 & \sqrt{\sigma_1} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sqrt{\sigma_{r-1}} \end{bmatrix}}_{\hat{\mathrm{U}}_\ell} \cdot \underbrace{\begin{bmatrix} \sqrt{\sigma_0} & 0 & \dots & 0 \\ 0 & \sqrt{\sigma_1} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sqrt{\sigma_{r-1}} \end{bmatrix} \begin{bmatrix} v_{0,0} & v_{0,1} & \dots & v_{0,n-1} \\ v_{1,0} & v_{1,1} & \dots & v_{1,n-1} \\ \vdots & \vdots & \vdots & \vdots \\ v_{r-1,0} & v_{r-1,1} & \dots & v_{r-1,n-1} \end{bmatrix}}_{\hat{\mathrm{V}}_\ell}$$

$$\mathrm{W}_\ell^{\mathrm{pre}} \approx \hat{\mathrm{U}}_\ell \cdot \hat{\mathrm{V}}_\ell$$

Codebook Quantization

$$\begin{bmatrix} 4.3 & 3.6 & 1.8 \\ 3.3 & 2.2 & 4.7 \\ 1.4 & 1.2 & 4.5 \end{bmatrix}$$

K-Means Quantization



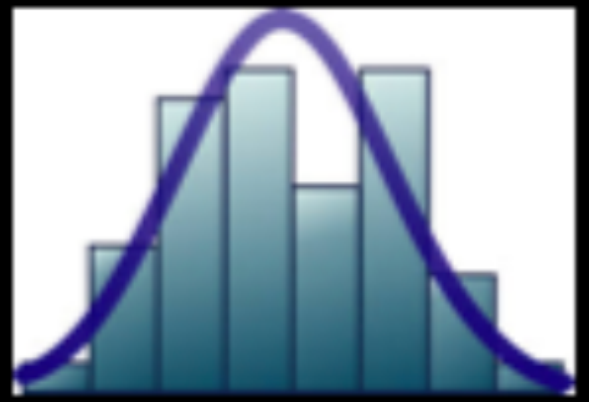
Indexes (Symbols)

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 0 & 2 \\ 0 & 0 & 2 \end{bmatrix}$$

Codebook

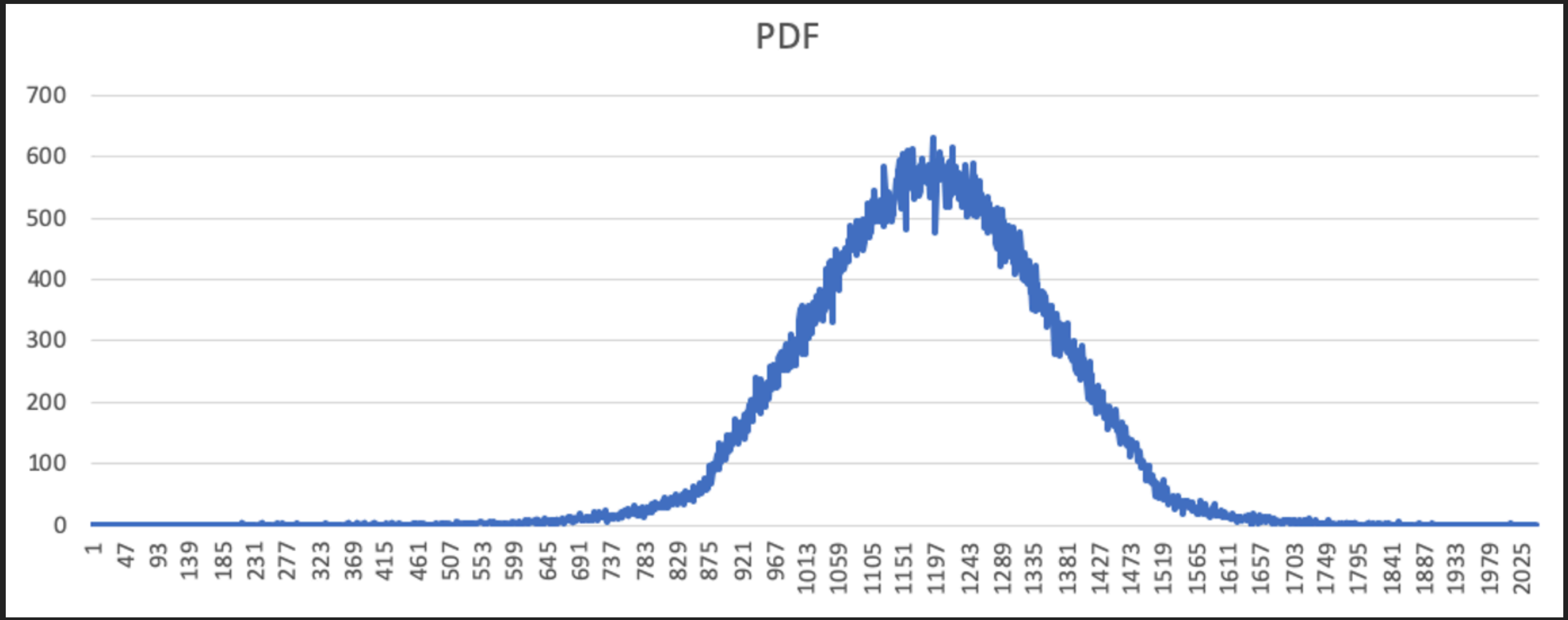
$$\begin{bmatrix} 1.7 \\ 3.6 \\ 4.5 \end{bmatrix}$$

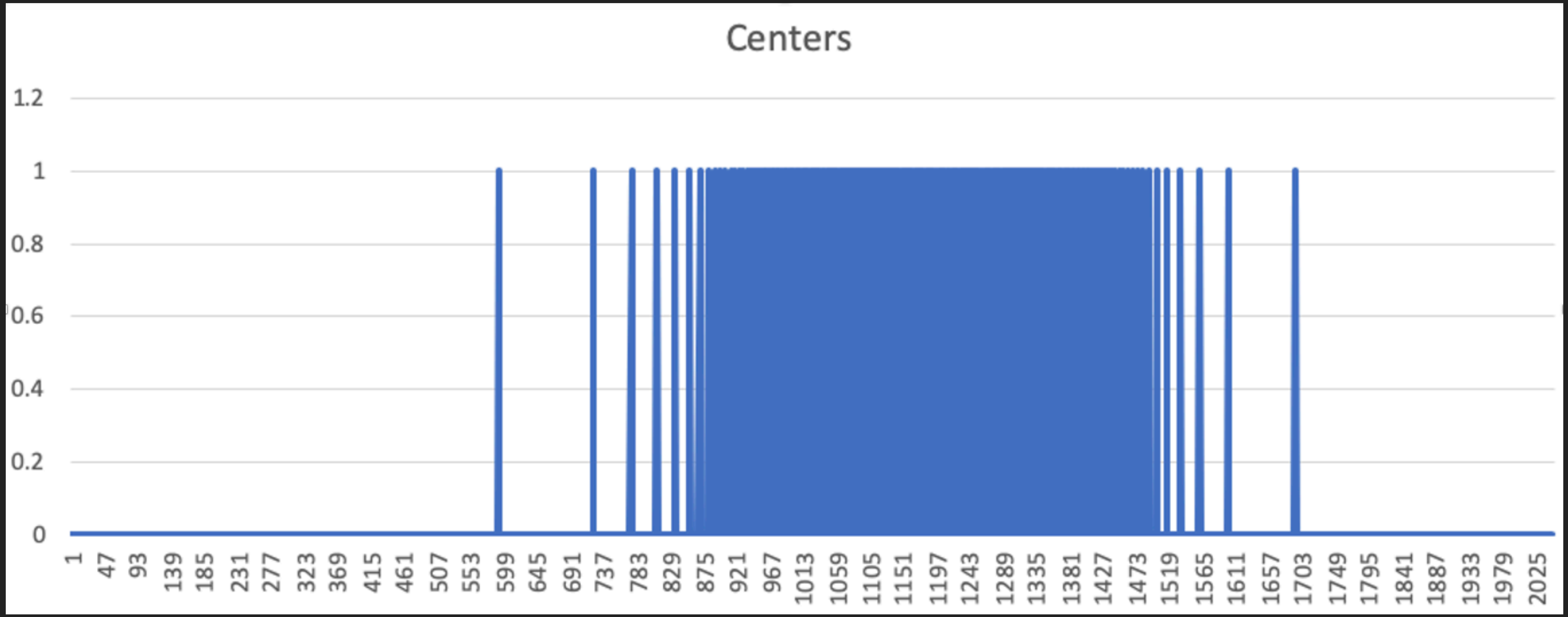
Probability Model



$$\begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}$$

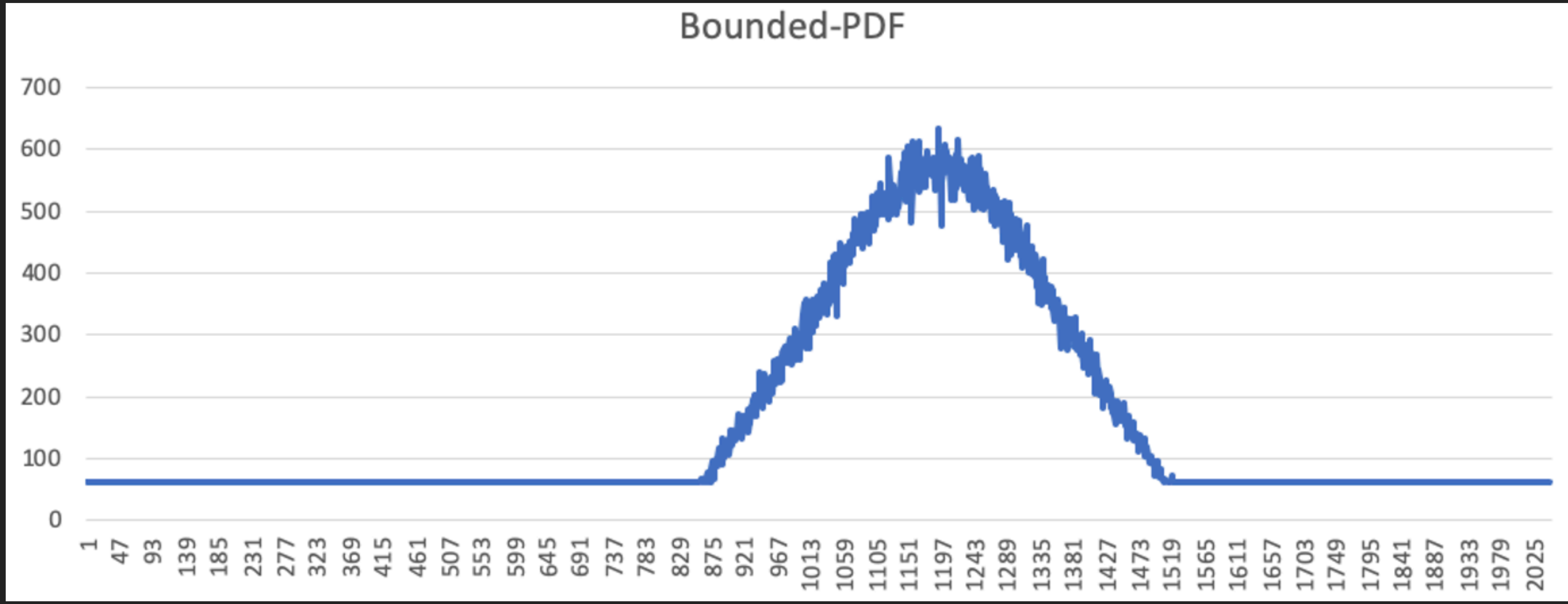
Codebook Quantization (K-Means Initialization)

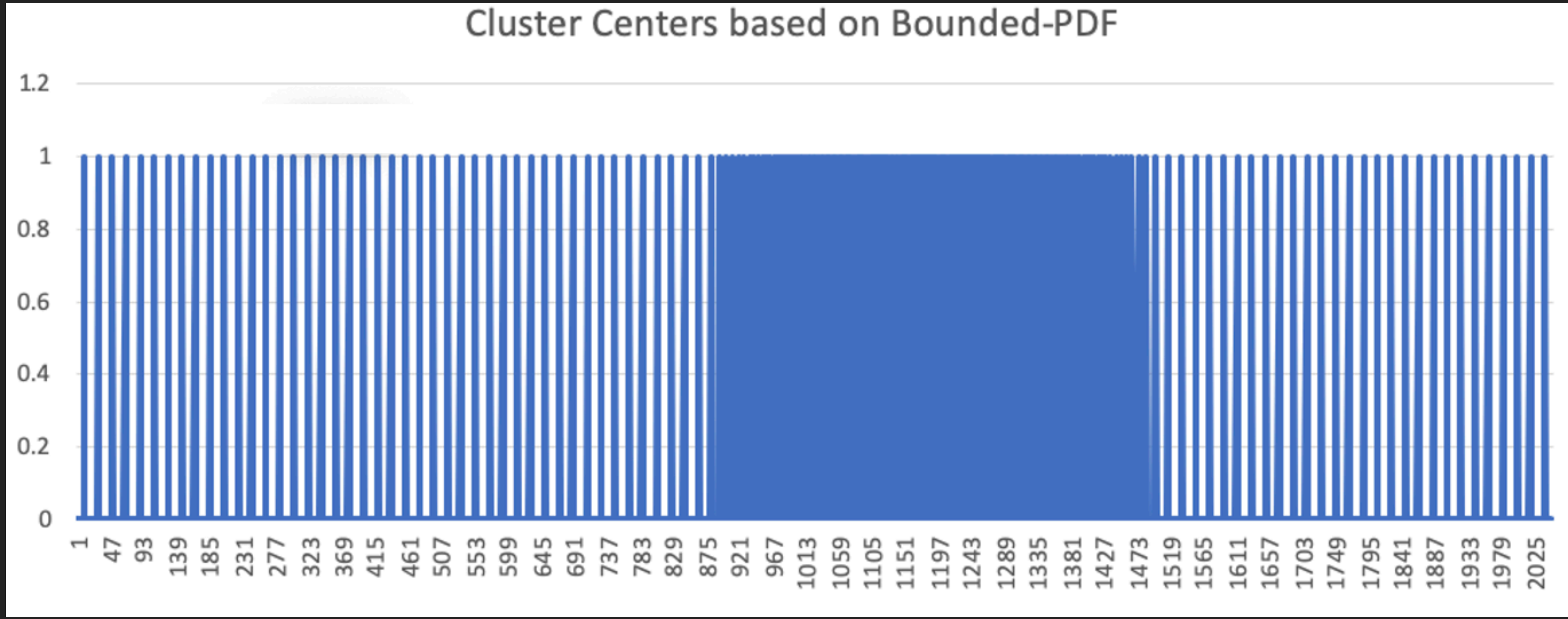




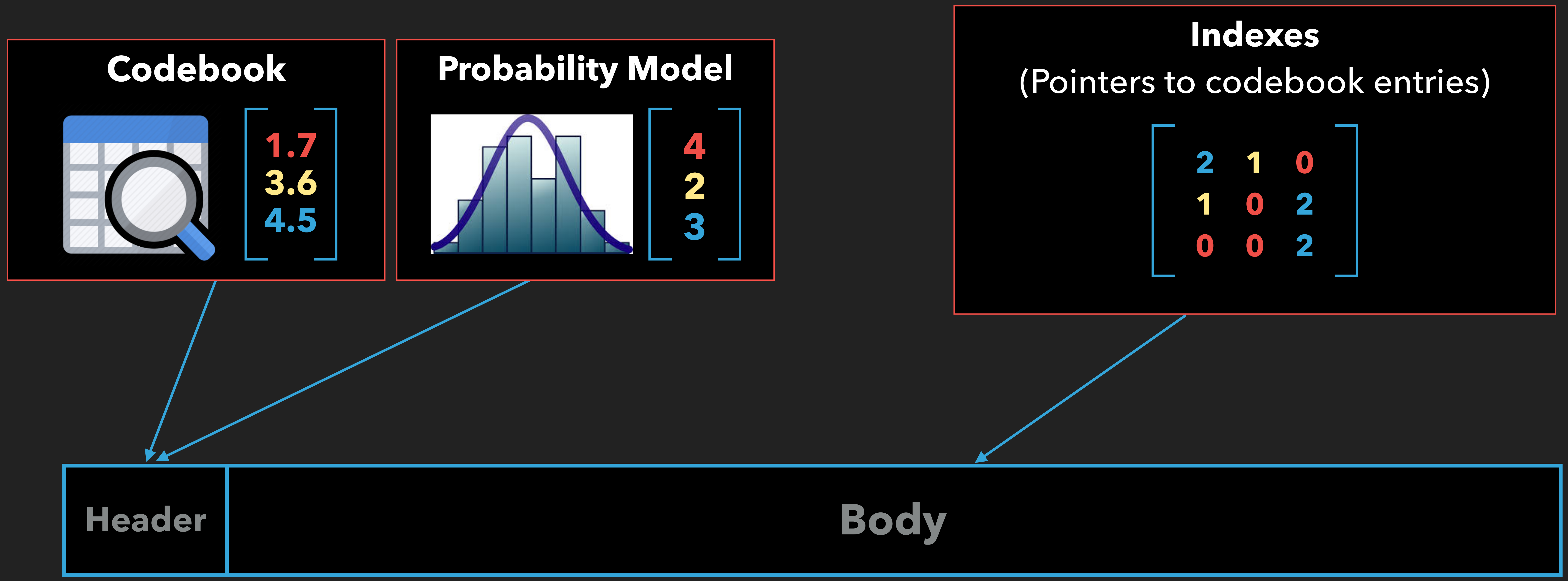
interdigital

Codebook Quantization (K-Means Initialization)

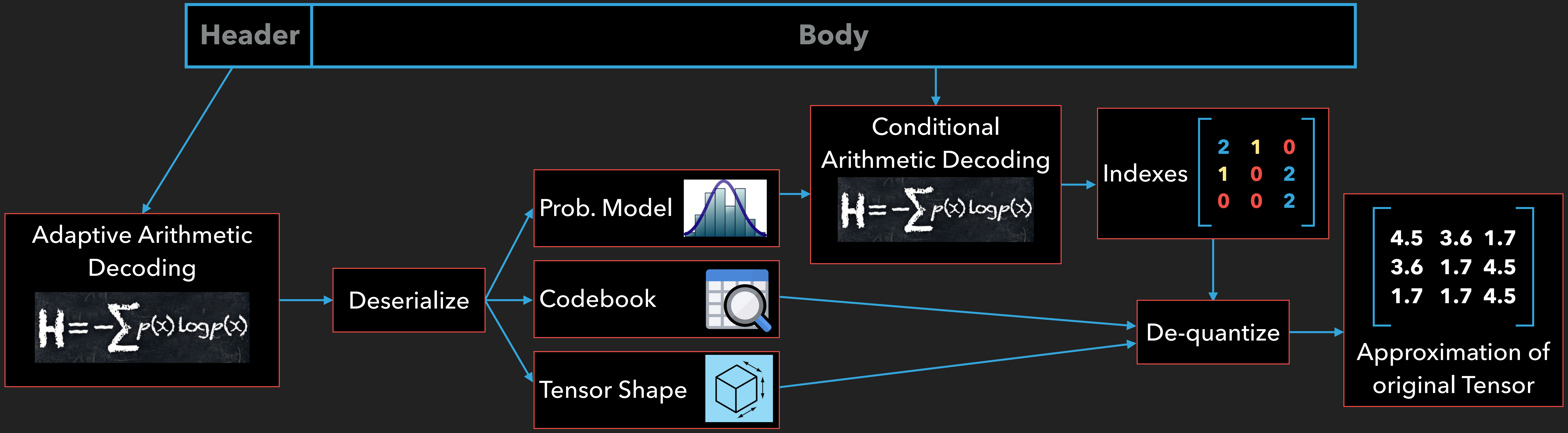




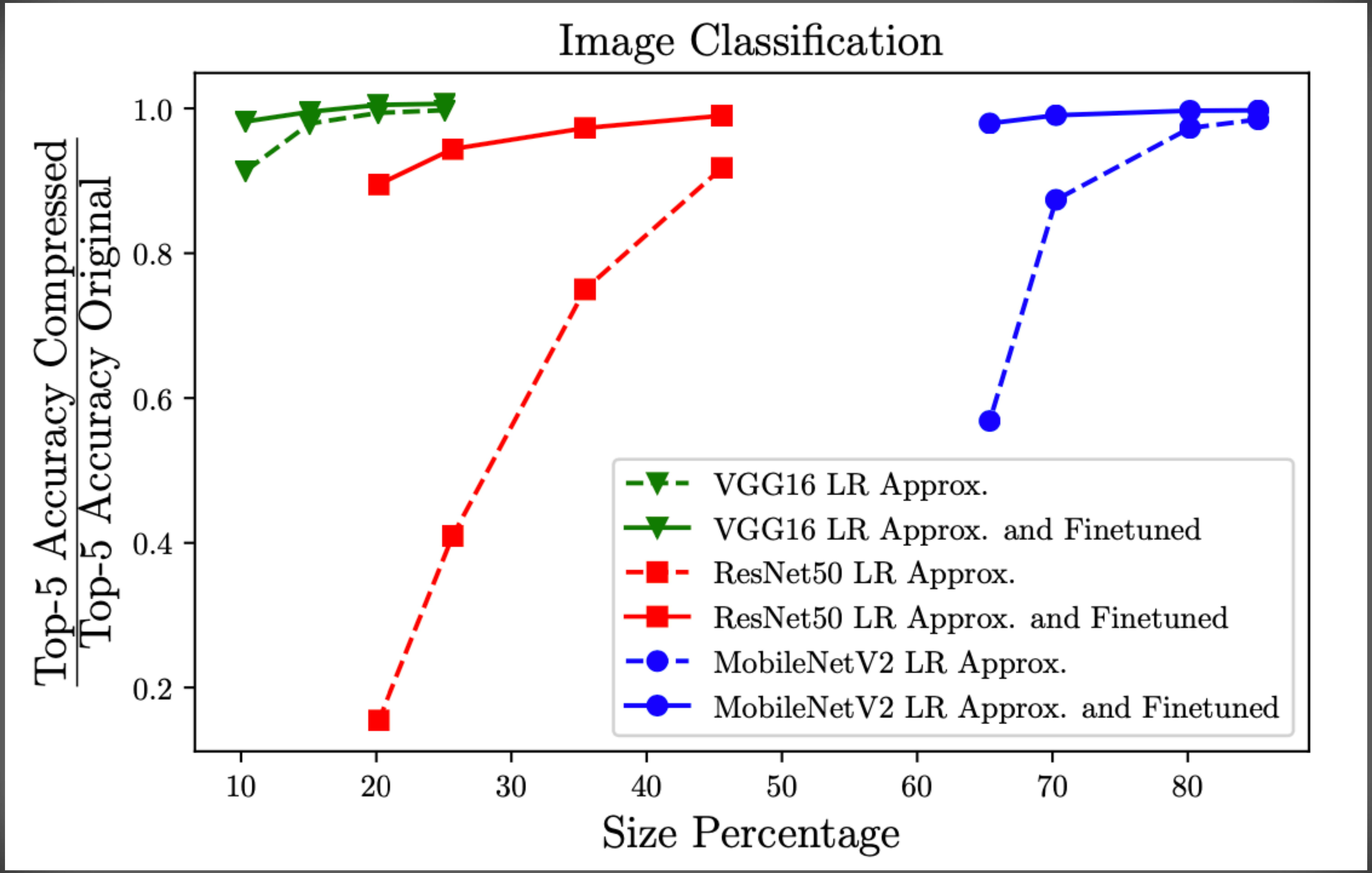
Serialization and Entropy Coding



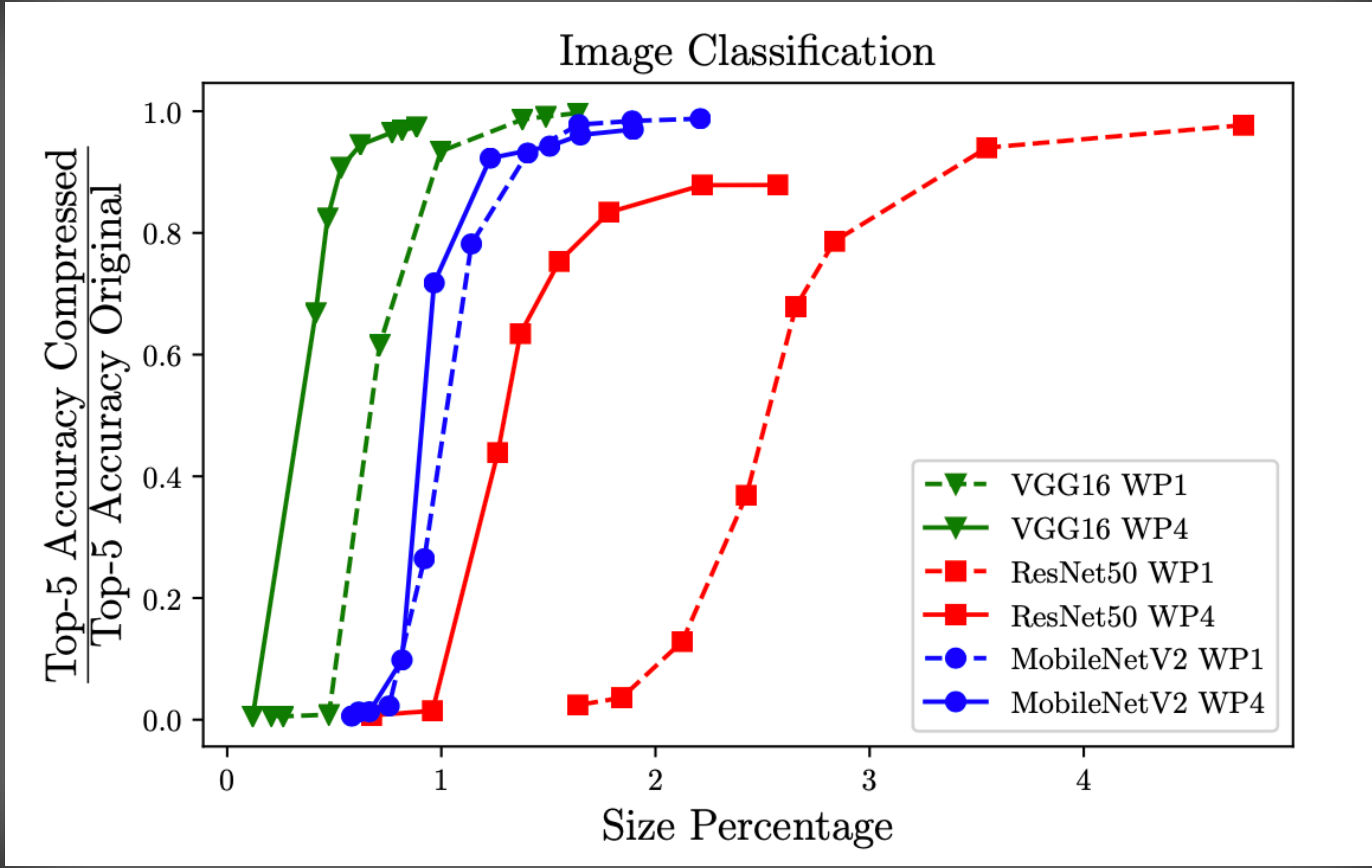
Decoding process



Results – Structural Approximation



Results – The whole compression pipeline



Thank you!

www.interdigital.com/artificial-intelligence