पेटेंट कार्यालय
शासकीय जर्नल

OFFICIAL JOURNAL
OF
THE PATENT OFFICE

निर्गमन सं. 20/2026	शुक्रवार	दिनांक: 15/05/2026
ISSUE NO. 20/2026	FRIDAY	DATE: 15/05/2026

पेटेंट कार्यालय का एक प्रकाशन
PUBLICATION OF THE PATENT OFFICE

(12) PATENT APPLICATION PUBLICATION
(19) INDIA
(22) Date of filing of Application :02/04/2026
(21) Application No.202611042202 A
(43) Publication Date : 15/05/2026

(54) Title of the invention : A SYSTEM AND METHOD FOR DATA-FREE COMPRESSION OF NEURAL NETWORK MODELS USING ENTROPY-OPTIMIZED QUANTIZATION

(51) International classification	:G06N 3/04, G06N 3/08, G06N 3/063, H03M 7/40, H03M 7/30
(31) Priority Document No	:NA
(32) Priority Date	:NA
(33) Name of priority country	:NA
(86) International Application No Filing Date	: :01/01/1900
(87) International Publication No	: NA
(61) Patent of Addition to Application Number Filing Date	:NA :NA
(62) Divisional to Application Number Filing Date	:NA :NA

(71)**Name of Applicant :**
1)Noida Institute of Engineering and Technology (NIET)
Address of Applicant :19, Institutional Area, Knowledge Park II, Greater Noida, Uttar Pradesh 201310 Uttar Pradesh India
(72)**Name of Inventor :**
1)Aman Prasad
2)Dr. Kanika Singhal

(57) Abstract :
The present invention relates to a hardware-software integrated system (100) and method for compressing neural network models without requiring calibration data. The system comprises a processor unit (101) with dedicated arithmetic logic units, a memory controller (102), an entropy optimization module (103), a quantization engine (104), and an entropy coding unit (105) implementing Asymmetric Numeral Systems. The method decouples numerical precision from storage cost by maintaining Float8 or Int8 representation while optimizing weight matrices for minimum entropy using L1-norm regularization as a differentiable entropy proxy in a rate-distortion optimization framework. The processing module (106) performs channel-wise scaling with optimized scale parameters computed via L-BFGS optimization. The decompression buffer (107) enables on-the-fly inference-time decoding. The system achieves effective compression to approximately two bits per parameter while retaining representational precision of thirty-four or more unique values, enabling deployment of large language models exceeding seventy billion parameters on consumer-grade graphics processing units.

No. of Pages : 20 No. of Claims : 10