Novelty Assessment Report

Paper: Ensembling Pruned Attention Heads For Uncertainty-Aware Efficient Transformers

PDF URL: https://openreview.net/pdf?id=pJZbMECfLP

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Uncertainty quantification (UQ) is essential for deploying deep neural networks in safety-critical settings. Although methods like Deep Ensembles achieve strong UQ performance, their high computational and memory costs hinder scalability to large models. We introduce Hydra Ensembles, an efficient transformer-based ensemble that prunes attention heads to create diverse members and merges them via a new multi-head attention with grouped fully-connected layers. This yields a compact model with inference speed close to a single network, matching or surpassing Deep Ensembles in UQ performance without retraining from scratch. We also provide an in-depth analysis of pruning, showing that naive approaches can harm calibration, whereas Hydra Ensembles preserves robust uncertainty. Experiments on image and text classification tasks, with various architectures, show consistent gains over Deep Ensembles. Remarkably, in zero-shot classification on ImageNet-1k, our approach surpasses state of the art methods, even without requiring additional training.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Uncertainty Quantification in Transformer-Based Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Uncertainty Quantification Methods and Architectures**
- **Natural Language Processing Applications**
- **Computer Vision Applications**
- **Multimodal and Cross-Modal Applications**
- **Time Series and Temporal Modeling**
- **Engineering and Industrial Applications**
- **Robotics and Autonomous Systems**
- **Data Imputation and Missing Value Handling**
- **Graph Neural Networks with Attention**

Complete Taxonomy Tree

- Uncertainty Quantification in Transformer-Based Models Survey Taxonomy
- Uncertainty Quantification Methods and Architectures
  - Probabilistic and Bayesian Transformer Extensions
  - Stochastic Attention Mechanisms (3 papers)
    - [24] Uncertainty-guided probabilistic transformer for complex action recognition (Hongji Guo, 2022) View paper
    - [36] Paying attention to uncertainty: A stochastic multimodal transformers for post-traumatic stress disorder detection using video (Mamadou Dia, 2024) View paper
    - [47] Transformer uncertainty estimation with hierarchical stochastic attention (Cheng Wang, 2022) View paper
  - Bayesian Neural Network Integration (3 papers)
    - [9] Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in transformer (Yiming Xiao, 2023) View paper
    - [19] A reliable and adaptive prediction framework for nuclear power plant system through an improved Transformer model and Bayesian uncertainty analysis (Xiao Xiao, 2025) View paper
    - [37] Uncertainty-aware traffic prediction using attention-based deep hybrid network with bayesian inference (Md. Moshiur Rahman, 2023) View paper
  - Variational and Generative Probabilistic Models (3 papers)
    - [28] Uncertainty-aware deep attention recurrent neural network for heterogeneous time series imputation (Qian, 2024) View paper
    - [31] Long-term Action Forecasting Using Multi-headed Attention-based Variational Recurrent Neural Networks (Siyuan Brandon Loh, 2022) View paper
    - [34] Transformer neural processes: Uncertainty-aware meta learning via sequence modeling (Nguyen, 2022) View paper
  - Ensemble and Aggregation Approaches
  - Pruning-Based and Efficient Ensembles ★ (2 papers)
    - [0] Ensembling Pruned Attention Heads For Uncertainty-Aware Efficient Transformers (Anon et al., 2026) View paper
    - [42] Uncertainty quantification in fine-tuned LLMs using LoRA ensembles (Balabanov, 2024) View paper
  - Training-Based Ensemble Methods (3 papers)
    - [8] Early Uncertainty Quantification Prediction of Lithium-Ion Battery Remaining Useful Life With Transformer Ensemble Model (Jijuan Hu, 2024) View paper

        - [12] Bayesian cooperative probabilistic Transformer for remaining useful life prediction with uncertainty estimation in industrial equipment (Shushuai Xie, 2025) View paper
        - [38] Few-Shot Probabilistic RUL Prediction With Uncertainty Quantification of Slurry Pumps (Yu Wang, 2025) View paper
    - Dropout-Based and Sampling Methods (2 papers)
    - [17] Uncertainty estimation of transformer predictions for misclassification detection (Artem Vazhentsev, 2022) View paper
    - [50] How certain is your Transformer? (Artem Shelmanov, 2021) View paper
    - Score-Based and Alternative Training Objectives (2 papers)
    - [14] SMURF-THP: Score Matching-based UnceRtainty quantiFication for Transformer Hawkes Process (Li, 2023) View paper
    - [39] Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction (Philippe Schwaller, 2019) View paper
- Natural Language Processing Applications
    - Large Language Model Uncertainty (2 papers)
    - [4] A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions (Ola Shorinwa, 2025) View paper
    - [5] Language model uncertainty quantification with attention chain (Li Yinghao, 2025) View paper
    - Text Classification and Sequence Tasks (2 papers)
    - [3] Uncertainty Quantification for Transformer Models for Dark-Pattern Detection (Munoz Javier, 2024) View paper
    - [29] MetaCert: Metabolic Attention Network Utilizing Uncertainty Estimation for Multimodal Aspect-Category-Sentiment Triple Extraction (Haoran Luo, 2025) View paper
    - Topological and Embedding Analysis (1 papers)
    - [43] Uncertainty Estimation of Transformers' Predictions via Topological Analysis of the Attention Matrices (Elizaveta Kostenok, 2023) View paper
- Computer Vision Applications
    - Image Classification and Recognition (2 papers)
    - [7] Uncertainty estimation for time series classification: Exploring predictive uncertainty in transformer-based models for variable stars (G. Cabrera-Vives, 2024) View paper
    - [13] Conformal uncertainty quantification to evaluate predictive fairness of foundation AI model for skin lesion classes across patient demographics (Pal, 2025) View paper
    - Object Detection and Localization (2 papers)
    - [40] Quantifying the reliability of predictions in detection transformers: Object-level calibration and image-level uncertainty (Park Young Jin, 2024) View paper
    - [41] UBTransformer: Uncertainty-based Transformer Model for Complex Scenarios Detection in Autonomous Driving (Ke Wang, 2025) View paper
    - Semantic and Medical Image Segmentation (4 papers)
    - [10] Uacanet: Uncertainty augmented context attention for polyp segmentation (Tae-hun Kim, 2021) View paper
    - [16] Uncertainty-guided transformer reasoning for camouflaged object detection (Fan Yang, 2021) View paper
    - [21] UCTNet: Uncertainty-guided CNN-Transformer hybrid networks for medical image segmentation (Xiayu Guo, 2024) View paper
    - [46] Uncertainty-guided and cross-modality attention network for liver tumor segmentation and quantification via integrating dynamic MRI (Jianfeng Zhao, 2025) View paper
    - Image Enhancement and Restoration (2 papers)
    - [20] Uncertainty-driven mixture convolution and transformer network for remote sensing image super-resolution (Xiaomin Zhang, 2024) View paper
    - [27] Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks (Sixiang Chen, 2023) View paper
- Multimodal and Cross-Modal Applications (1 papers)
    - [18] TMU-Net: A Transformer-Based Multimodal Framework with Uncertainty Quantification for Driver Fatigue Detection (Yaxin Zhang, 2025) View paper
- Time Series and Temporal Modeling
    - Action Recognition and Anticipation (1 papers)
    - [6] Uncertainty-aware action decoupling transformer for action anticipation (Hongji Guo, 2024) View paper
    - Forecasting and Predictive Modeling (3 papers)
    - [15] EEG-based seizure prediction via hybrid vision transformer and data uncertainty learning (Zhi-Wei Deng, 2023) View paper
    - [32] Incorporating uncertainty estimation and interpretability in personalized glucose prediction using the temporal fusion transformer (Antonio J. RodrÃguez-Almeida, 2025) View paper
    - [49] AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting (Ye Yuan, 2021) View paper
- Engineering and Industrial Applications
    - Structural and Geotechnical Analysis (2 papers)
    - [2] Advancing dynamic reliability assessment of reservoir slopes using attention-based neural networks (Lin Wang, 2025) View paper
    - [26] Improving Reliability of Seismic Stratigraphy Prediction: Integration of Uncertainty Quantification in Attention Mechanism Neural Network (C. T. Ang, 2024) View paper
    - Fault Diagnosis and Condition Monitoring (2 papers)
    - [30] Advanced fault diagnosis in milling cutting tools using vision transformers with semi-supervised learning and uncertainty quantification (Muhammad Farooq Siddique, 2025) View paper
    - [48] Uncertainty-aware deep variational attention network: A trustworthy mechanical fault diagnostic model assisted by out-of-distribution detection (Yiming Xiao, 2025) View paper
    - Environmental and Physical System Estimation (3 papers)
    - [11] Estimation of wind turbine responses with attention-based neural network incorporating environmental uncertainties (Dong-Lai Zhu, 2024) View paper
    - [22] Estimation of Sea State Parameters from Ship Motion Responses Using Attention-based Neural Networks (SelimoviÄ, 2023) View paper
    - [25] An interpretable wheat yield estimation model using an attention mechanism-based deep learning framework with multiple remotely sensed variables (Mingqi* LI, 2025) View paper
- Robotics and Autonomous Systems
    - Autonomous Driving and Navigation (2 papers)

    - [23] Uncertainty-aware decision transformer for stochastic driving environments (Li Zenan, 2023) View paper
    - [44] Uncertainty-aware hybrid paradigm of nonlinear MPC and model-based RL for offroad navigation: Exploration of transformers in the predictive model (Faraz Lotfi, 2023) View paper
    - Racing and High-Speed Control (1 papers)
    - [1] Autonomous racing with attention-based neural networks (Resch, 2023) View paper
- Data Imputation and Missing Value Handling (2 papers)
    - [33] Well logging prediction and uncertainty analysis based on recurrent neural network with attention mechanism and Bayesian theory (Lili Zeng, 2022) View paper
    - [45] ST-GIN: An Uncertainty Quantification Approach in Traffic Data Imputation with Spatio-Temporal Graph Attention and Bidirectional Recurrent United Neural Networks (Zepu Wang, 2023) View paper
- Graph Neural Networks with Attention (1 papers)
    - [35] Uag: Uncertainty-aware attention graph neural network for defending adversarial attacks (Ding, 2021) View paper

Narrative

Core task: uncertainty quantification in transformer-based models. The field has grown into a rich landscape organized around both methodological innovations and diverse application domains. At the methodological level, researchers explore ensemble and aggregation approaches, Bayesian and variational techniques, and architectural modifications that embed uncertainty directly into attention mechanisms. Meanwhile, application-oriented branches span natural language processing, computer vision, time series forecasting, robotics, and engineering domains such as fault diagnosis and predictive maintenance. Works like LLM Uncertainty Survey[4] and Attention Chain Uncertainty[5] illustrate the breadth of strategies for capturing epistemic and aleatoric uncertainty, while domain-specific studies—ranging from Autonomous Racing Attention[1] to Battery Life Transformer Ensemble[8]—demonstrate how these methods adapt to real-world constraints and safety-critical settings.

Within the ensemble and aggregation branch, a particularly active line of work focuses on balancing computational efficiency with robust uncertainty estimates. Ensembling Pruned Attention[0] exemplifies this trade-off by combining model pruning with ensemble techniques to reduce inference costs while maintaining reliable uncertainty quantification. This contrasts with approaches like LoRA Ensemble Uncertainty[42], which leverages parameter-efficient fine-tuning to build lightweight ensembles, and with more classical Bayesian methods that impose heavier computational overhead. The original paper sits squarely in this efficiency-focused cluster, addressing the practical challenge of deploying transformer ensembles at scale. By pruning redundant parameters before aggregation, it offers a middle ground between the full expressiveness of large ensembles and the speed required for production systems, a theme that resonates across many engineering and industrial applications where both accuracy and latency matter.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Uncertainty quantification in fine-tuned LLMs using LoRA ensembles

Authors: Balabanov, Oleksandr, Linander, Hampus, Oleksandr Balabanov, et al. (6 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Fine-tuning large language models can improve task specific performance, although a general understanding of what the fine-tuned model has learned, forgotten and how to trust its predictions is still missing. We derive principled uncertainty quantification for fine-tuned LLMs with posterior approximations using computationally efficient low-rank adaptation ensembles. We analyze three common multiple-choice datasets using low-rank adaptation ensembles based on Mistral-7b, and draw quantitative an...

Relationship Analysis

Both papers belong to the Pruning-Based and Efficient Ensembles category, focusing on creating computationally efficient ensemble methods for uncertainty quantification in transformers. They overlap in using parameter-efficient techniques to build diverse ensemble members while reducing inference costs compared to Deep Ensembles. The key difference is that the original paper (Hydra Ensembles) creates diversity through structured pruning of attention heads and merges them into a single model via grouped fully-connected layers, whereas the candidate paper uses LoRA (Low-Rank Adaptation) ensembles with multiple low-rank adapter matrices to approximate Bayesian posteriors during fine-tuning, maintaining separate ensemble members rather than merging into one architecture.

Contributions Analysis

Overall novelty summary. The paper proposes Hydra Ensembles, a framework that creates diverse ensemble members by pruning attention heads and merging them via grouped fully-connected layers in multi-head attention. It sits in the 'Pruning-Based and Efficient Ensembles' leaf, which contains only two papers total. This is a relatively sparse research direction within the broader taxonomy of 50 papers across 24 leaf nodes, suggesting that efficient ensemble construction for transformers remains an underexplored area compared to more crowded branches like Bayesian extensions or application-specific studies.

The taxonomy tree shows that Hydra Ensembles belongs to the 'Ensemble and Aggregation Approaches' branch, which also includes 'Training-Based Ensemble Methods' (three papers on stochastic weight averaging and teacher-student frameworks). Neighboring branches include 'Probabilistic and Bayesian Transformer Extensions' (nine papers on stochastic attention and variational inference) and 'Dropout-Based and Sampling Methods' (two papers on Monte Carlo dropout). The paper diverges from these by avoiding probabilistic modeling or repeated training, instead focusing on structural pruning and parameter sharing to achieve computational efficiency while preserving uncertainty quantification.

Among 25 candidates examined, none clearly refute the three main contributions. The Hydra Ensembles framework itself was assessed against five candidates with zero refutations. The pruning-calibration analysis examined ten candidates, finding no prior work that systematically studies how naive pruning degrades calibration in ensemble settings. The circuit-based head selection strategy also examined ten candidates without encountering overlapping prior art. These statistics suggest that, within the limited search scope, the contributions appear relatively novel, though the small candidate pool (25 total) means the analysis does not cover the full landscape of pruning or ensemble literature.

Based on the top-25 semantic matches and the sparse taxonomy leaf, the work appears to occupy a distinct niche at the intersection of pruning and ensemble uncertainty quantification. However, the limited search scope and the small number of sibling papers in the taxonomy leaf make it difficult to assess whether related ideas exist in adjacent communities (e.g., model compression or neural architecture search). A broader literature review would be needed to confirm the novelty claims more definitively.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Hydra Ensembles framework for efficient transformer-based ensembles

Description: The authors propose Hydra Ensembles, a method that creates diverse ensemble members by pruning attention heads from a single pre-trained transformer and merging them into a compact model using grouped fully-connected layers. This approach achieves inference speed close to a single network while matching or surpassing Deep Ensembles in uncertainty quantification performance without retraining from scratch.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Revisiting Vision Transformer from the View of Path Ensemble

URL: View paper

Brief Assessment

Path Ensemble Vision Transformer[63] focuses on reinterpreting existing ViT architectures as implicit ensemble networks through path analysis and pruning individual paths within a single model, rather than creating diverse ensemble members through attention head pruning and merging multiple models as in the original paper.

2. Towards efficient deep learning for vision and language applications

URL: View paper

Brief Assessment

Efficient Vision Language[61] mentions attention head pruning and token reduction methods but does not describe creating ensemble members through pruning or merging them into compact models for uncertainty quantification, which is the core novelty of Hydra Ensembles.

3. Stochastic Attention Head Removal: A Simple and Effective Method for Improving Transformer Based ASR Models

URL: View paper

Brief Assessment

Stochastic Attention Head Removal[65] focuses on randomly removing attention heads during training for ASR models to improve performance, not on creating efficient ensembles for uncertainty quantification through pruning and merging.

4. Ensemble of winning tickets: pruning bidirectional encoder from the transformers attention heads for enhanced model efficiency

URL: View paper

Brief Assessment

Winning Tickets Attention Pruning[64] focuses on pruning BERT attention heads to reduce model size while maintaining performance on NLP tasks, but does not propose an ensemble framework that merges pruned models into a single compact architecture with grouped fully-connected layers for uncertainty quantification as in Hydra Ensembles.

5. Dha: Learning decoupled-head attention from transformer checkpoints via adaptive heads fusion

URL: View paper

Brief Assessment

Decoupled Head Attention[62] focuses on reducing attention head redundancy through adaptive fusion for efficient inference, not on creating ensemble models for uncertainty quantification.

Contribution 2: Theoretical and empirical analysis of pruning effects on calibration

Description: The authors provide both theoretical analysis (Proposition 1) and empirical evidence showing that commonly used pruning methods can harm calibration and lead to unreliable predictions, despite preserving accuracy. They establish conditions under which pruning degrades uncertainty quantification performance.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Uncertainty Estimation Pseudo-Labels Guided Source-Free Domain Adaptation for Cross-Domain Remaining Useful Life Prediction in IIoT

URL: View paper

Brief Assessment

Cross Domain RUL Prediction[55] focuses on remaining useful life prediction in industrial IoT using domain adaptation, not on pruning effects on calibration or uncertainty quantification in transformers.

2. MOBIUS: Big-to-Mobile Universal Instance Segmentation via Multi-modal Bottleneck Fusion and Calibrated Decoder Pruning

URL: View paper

Brief Assessment

MOBIUS Instance Segmentation[59] focuses on decoder pruning for instance segmentation models with language-guided uncertainty calibration, not on analyzing how pruning degrades calibration in transformers for uncertainty quantification tasks.

3. BaS-former: a trustworthy model of machinery fault diagnosis for quantifying aleatoric uncertainty under noise discrepancy

URL: View paper

Brief Assessment

BaS Machinery Fault[52] focuses on machinery fault diagnosis with uncertainty quantification under noise discrepancy, not on pruning effects on calibration in transformers. The candidate addresses a completely different domain (machinery diagnostics) and does not discuss pruning methods or their impact on model calibration.

4. Likelihood-guided Regularization in Attention Based Models

URL: View paper

Brief Assessment

Likelihood Guided Regularization[60] focuses on Ising-based variational regularization for attention models, not on analyzing how pruning methods degrade calibration in transformers. The candidate does not address pruning's impact on uncertainty quantification.

5. Confident magnitude-based neural network pruning

URL: View paper

Brief Assessment

Confident Magnitude Pruning[56] focuses on uncertainty quantification for pruning decisions using distribution-free methods, not on analyzing how pruning degrades calibration in transformers. The candidate addresses stopping criteria with statistical guarantees, while the original analyzes calibration degradation effects.

6. Platon: Pruning large transformer models with upper confidence bound of weight importance

URL: View paper

Brief Assessment

Platon Upper Confidence Pruning[54] focuses on pruning stability and importance score uncertainty during training, not on calibration or uncertainty quantification performance. The paper does not analyze how pruning affects model calibration metrics.

7. Better Reliability Compression: Model Pruning with Calibrated Uncertainty Estimation for Mobile Deep Learning Applications

URL: View paper

Brief Assessment

Reliability Compression Pruning[57] focuses on structured pruning with uncertainty calibration for mobile deployment, not on transformer attention heads or the theoretical conditions under which pruning degrades uncertainty quantification.

8. Iterative network pruning with uncertainty regularization for lifelong sentiment classification

URL: View paper

Brief Assessment

Lifelong Sentiment Pruning[58] focuses on iterative pruning for lifelong learning in sentiment classification, not on analyzing pruning's effects on calibration or uncertainty quantification in transformers.

9. Self-calibration for language model quantization and pruning

URL: View paper

Brief Assessment

Self Calibration Quantization Pruning[51] focuses on calibration data selection for post-training compression, not on analyzing how pruning methods affect model calibration and uncertainty quantification.

10. Application of dataset pruning and dynamic transfer learning on vision transformers for mgmt prediction on brain mri images

URL: View paper

Brief Assessment

MGMT Vision Transformer Pruning[53] focuses on dataset pruning (removing less informative images) for medical imaging tasks, not model pruning or calibration analysis. The paper does not address uncertainty quantification or calibration degradation from pruning neural network weights.

Contribution 3: Circuit-based head selection strategy for uncertainty quantification

Description: The authors introduce a circuit-based approach for selecting attention heads that preserves useful functionality for uncertainty estimation. This strategy, using methods like Headmap algorithm, extracts subnetworks that remain stable under noise, addressing the limitations of naive pruning for uncertainty quantification tasks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Knowing what you don't know: Estimating the uncertainty of feedforward and feedback inputs with prediction-error circuits

URL: View paper

Brief Assessment

Prediction Error Feedforward Feedback[71] focuses on neural circuit models for estimating uncertainty in sensory prediction-error processing, not on transformer attention head selection or pruning strategies for ensemble uncertainty quantification.

2. Approaches to uncertainty quantification in federated deep learning

URL: View paper

Brief Assessment

Federated Uncertainty Quantification[70] focuses on uncertainty quantification methods (MC-dropout, SWAG, ensembles) in federated learning settings, not on circuit-based subnetwork selection or attention head pruning strategies for transformers.

3. Sequential Bayesian Neural Subnetwork Ensembles

URL: View paper

Brief Assessment

Sequential Bayesian Subnetwork[75] focuses on dynamic sparsity learning in Bayesian neural networks through pruning and regrowing weights during training, not on circuit-based attention head selection for transformers. The candidate's approach uses signal-to-noise ratios and gradient-based criteria for weight pruning in general neural networks, which differs fundamentally from the original paper's circuit-based head selection using methods like Headmap for transformer uncertainty quantification.

4. Bayesian deep learning via subnetwork inference

URL: View paper

Brief Assessment

Bayesian Subnetwork Inference[67] focuses on selecting weight subnetworks for Bayesian inference using variance-based criteria, not circuit-based attention head selection for ensemble uncertainty quantification as in the original paper.

5. U2D2PCB: Uncertainty-Aware Unsupervised Defect Detection on PCB Images Using Reconstructive and Discriminative Models

URL: View paper

Brief Assessment

PCB Defect Detection Uncertainty[69] focuses on uncertainty quantification in PCB defect detection using reconstructive and discriminative U-Net subnetworks, not on circuit-based attention head selection for transformer ensembles.

6. Uncertainty estimation with prediction-error circuits

URL: View paper

Brief Assessment

Prediction Error Uncertainty[66] focuses on prediction-error circuits in neural networks for uncertainty estimation in sensory processing, not on attention head selection in transformers for ensemble-based uncertainty quantification.

7. Expressive yet tractable Bayesian deep learning via subnetwork inference

URL: View paper

Brief Assessment

Expressive Bayesian Subnetwork[74] focuses on subnetwork inference via Laplace approximation for Bayesian deep learning, selecting subnetworks based on marginal variances rather than circuit-based methods. The candidate does not address circuit extraction or attention head selection strategies for uncertainty quantification.

8. Simultaneous Inverse Design and Uncertainty Quantification for Frequency-Selective Rasorber With Tunable and Switchable Abilities by Bayesian Deep Learning

URL: View paper

Brief Assessment

Rasorber Bayesian Deep Learning[68] focuses on electromagnetic device design using Bayesian neural networks for uncertainty quantification in frequency-selective rasorbers, not on circuit-based subnetwork selection for transformer attention heads in ensemble methods.

9. FedSI: Federated Subnetwork Inference for Efficient Uncertainty Quantification

URL: View paper

Brief Assessment

FedSI Subnetwork Inference[73] focuses on federated learning with subnetwork selection based on parameter variance for Bayesian inference, not on circuit-based approaches using mechanistic interpretability methods like Headmap for uncertainty quantification in transformers.

10. Sub-ensembles for fast uncertainty estimation in neural networks

URL: View paper

Brief Assessment

Sub-ensembles for fast uncertainty estimation in neural networks[72] focuses on ensembling only selected layers near the output rather than circuit-based attention head selection. The candidate does not address circuit extraction methods or attention head pruning strategies for uncertainty quantification.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Ensembling Pruned Attention Heads For Uncertainty-Aware Efficient Transformers View paper
- [1] Autonomous racing with attention-based neural networks View paper
- [2] Advancing dynamic reliability assessment of reservoir slopes using attention-based neural networks View paper
- [3] Uncertainty Quantification for Transformer Models for Dark-Pattern Detection View paper
- [4] A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions View paper
- [5] Language model uncertainty quantification with attention chain View paper
- [6] Uncertainty-aware action decoupling transformer for action anticipation View paper
- [7] Uncertainty estimation for time series classification: Exploring predictive uncertainty in transformer-based models for variable stars View paper
- [8] Early Uncertainty Quantification Prediction of Lithium-Ion Battery Remaining Useful Life With Transformer Ensemble Model View paper
- [9] Towards trustworthy rotating machinery fault diagnosis via attention uncertainty in transformer View paper
- [10] Uacanet: Uncertainty augmented context attention for polyp segmentation View paper
- [11] Estimation of wind turbine responses with attention-based neural network incorporating environmental uncertainties View paper
- [12] Bayesian cooperative probabilistic Transformer for remaining useful life prediction with uncertainty estimation in industrial equipment View paper
- [13] Conformal uncertainty quantification to evaluate predictive fairness of foundation AI model for skin lesion classes across patient demographics View paper
- [14] SMURF-THP: Score Matching-based UnceRtainty quantiFication for Transformer Hawkes Process View paper
- [15] EEG-based seizure prediction via hybrid vision transformer and data uncertainty learning View paper
- [16] Uncertainty-guided transformer reasoning for camouflaged object detection View paper
- [17] Uncertainty estimation of transformer predictions for misclassification detection View paper
- [18] TMU-Net: A Transformer-Based Multimodal Framework with Uncertainty Quantification for Driver Fatigue Detection View paper
- [19] A reliable and adaptive prediction framework for nuclear power plant system through an improved Transformer model and Bayesian uncertainty analysis View paper
- [20] Uncertainty-driven mixture convolution and transformer network for remote sensing image super-resolution View paper
- [21] UCTNet: Uncertainty-guided CNN-Transformer hybrid networks for medical image segmentation View paper
- [22] Estimation of Sea State Parameters from Ship Motion Responses Using Attention-based Neural Networks View paper
- [23] Uncertainty-aware decision transformer for stochastic driving environments View paper
- [24] Uncertainty-guided probabilistic transformer for complex action recognition View paper
- [25] An interpretable wheat yield estimation model using an attention mechanism-based deep learning framework with multiple remotely sensed variables View paper
- [26] Improving Reliability of Seismic Stratigraphy Prediction: Integration of Uncertainty Quantification in Attention Mechanism Neural Network View paper
- [27] Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks View paper

- [28] Uncertainty-aware deep attention recurrent neural network for heterogeneous time series imputation View paper
- [29] MetaCert: Metabolic Attention Network Utilizing Uncertainty Estimation for Multimodal Aspect-Category-Sentiment Triple Extraction View paper
- [30] Advanced fault diagnosis in milling cutting tools using vision transformers with semi-supervised learning and uncertainty quantification View paper
- [31] Long-term Action Forecasting Using Multi-headed Attention-based Variational Recurrent Neural Networks View paper
- [32] Incorporating uncertainty estimation and interpretability in personalized glucose prediction using the temporal fusion transformer View paper
- [33] Well logging prediction and uncertainty analysis based on recurrent neural network with attention mechanism and Bayesian theory View paper
- [34] Transformer neural processes: Uncertainty-aware meta learning via sequence modeling View paper
- [35] Uag: Uncertainty-aware attention graph neural network for defending adversarial attacks View paper
- [36] Paying attention to uncertainty: A stochastic multimodal transformers for post-traumatic stress disorder detection using video View paper
- [37] Uncertainty-aware traffic prediction using attention-based deep hybrid network with bayesian inference View paper
- [38] Few-Shot Probabilistic RUL Prediction With Uncertainty Quantification of Slurry Pumps View paper
- [39] Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction View paper
- [40] Quantifying the reliability of predictions in detection transformers: Object-level calibration and image-level uncertainty View paper
- [41] UBTransformer: Uncertainty-based Transformer Model for Complex Scenarios Detection in Autonomous Driving View paper
- [42] Uncertainty quantification in fine-tuned LLMs using LoRA ensembles View paper
- [43] Uncertainty Estimation of Transformers' Predictions via Topological Analysis of the Attention Matrices View paper
- [44] Uncertainty-aware hybrid paradigm of nonlinear MPC and model-based RL for offroad navigation: Exploration of transformers in the predictive model View paper
- [45] ST-GIN: An Uncertainty Quantification Approach in Traffic Data Imputation with Spatio-Temporal Graph Attention and Bidirectional Recurrent United Neural Networks View paper
- [46] Uncertainty-guided and cross-modality attention network for liver tumor segmentation and quantification via integrating dynamic MRI View paper
- [47] Transformer uncertainty estimation with hierarchical stochastic attention View paper
- [48] Uncertainty-aware deep variational attention network: A trustworthy mechanical fault diagnostic model assisted by out-of-distribution detection View paper
- [49] AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting View paper
- [50] How certain is your Transformer? View paper
- [51] Self-calibration for language model quantization and pruning View paper
- [52] BaS-former: a trustworthy model of machinery fault diagnosis for quantifying aleatoric uncertainty under noise discrepancy View paper
- [53] Application of dataset pruning and dynamic transfer learning on vision transformers for mgmt prediction on brain mri images View paper
- [54] Platon: Pruning large transformer models with upper confidence bound of weight importance View paper
- [55] Uncertainty Estimation Pseudo-Labels Guided Source-Free Domain Adaptation for Cross-Domain Remaining Useful Life Prediction in IIoT View paper
- [56] Confident magnitude-based neural network pruning View paper
- [57] Better Reliability Compression: Model Pruning with Calibrated Uncertainty Estimation for Mobile Deep Learning Applications View paper
- [58] Iterative network pruning with uncertainty regularization for lifelong sentiment classification View paper
- [59] MOBIUS: Big-to-Mobile Universal Instance Segmentation via Multi-modal Bottleneck Fusion and Calibrated Decoder Pruning View paper
- [60] Likelihood-guided Regularization in Attention Based Models View paper
- [61] Towards efficient deep learning for vision and language applications View paper
- [62] Dha: Learning decoupled-head attention from transformer checkpoints via adaptive heads fusion View paper
- [63] Revisiting Vision Transformer from the View of Path Ensemble View paper
- [64] Ensemble of winning tickets: pruning bidirectional encoder from the transformers attention heads for enhanced model efficiency View paper
- [65] Stochastic Attention Head Removal: A Simple and Effective Method for Improving Transformer Based ASR Models View paper
- [66] Uncertainty estimation with prediction-error circuits View paper
- [67] Bayesian deep learning via subnetwork inference View paper
- [68] Simultaneous Inverse Design and Uncertainty Quantification for Frequency-Selective Rasorber With Tunable and Switchable Abilities by Bayesian Deep Learning View paper
- [69] U2D2PCB: Uncertainty-Aware Unsupervised Defect Detection on PCB Images Using Reconstructive and Discriminative Models View paper
- [70] Approaches to uncertainty quantification in federated deep learning View paper
- [71] Knowing what you don't know: Estimating the uncertainty of feedforward and feedback inputs with prediction-error circuits View paper
- [72] Sub-ensembles for fast uncertainty estimation in neural networks View paper
- [73] FedSI: Federated Subnetwork Inference for Efficient Uncertainty Quantification View paper
- [74] Expressive yet tractable Bayesian deep learning via subnetwork inference View paper
- [75] Sequential Bayesian Neural Subnetwork Ensembles View paper