Uncertainty Quantification

in Vision Transformer

Likelihood-guided Regularization in Attention Based Models

2026.04.06 조건우

Table of Content

01

Introduction

Problem Motivation

Vision Transformers (ViTs):

Provide strong representation power via self-attention

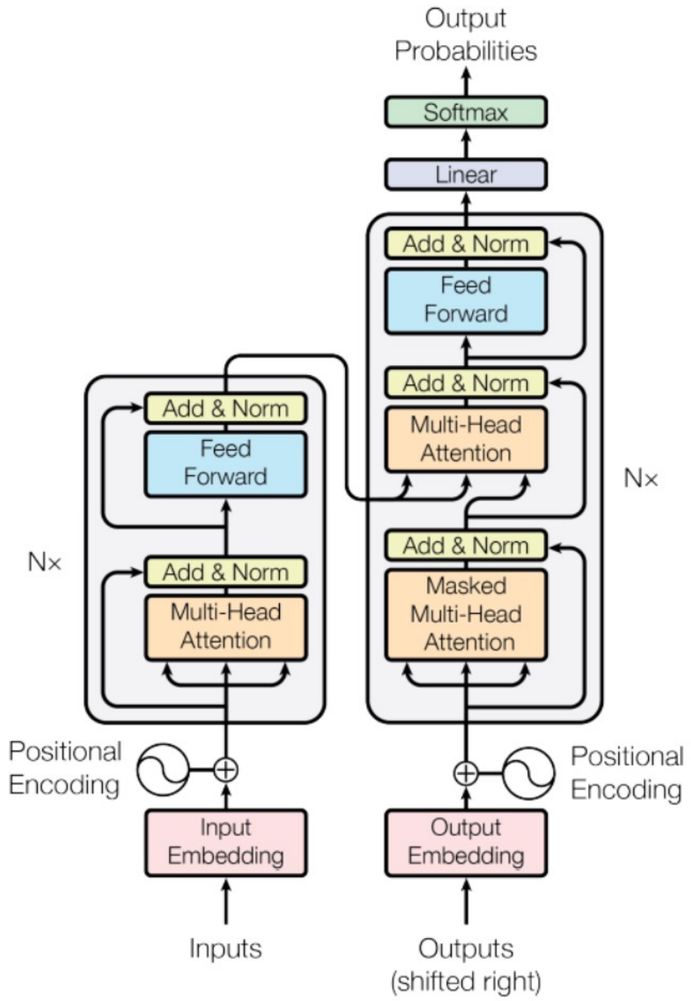
But:

Require large-scale data

Are prone to overfitting

Are inherently dense architectures

Therefore, effective regularization is essential



01

Introduction

Existing Methods and Limitations

(1) Dropout-based methods (Gal and Ghahramani [2016])

Randomly deactivate neurons

Limitations:

Unstructured sparsity

Not data-adaptive (fixed behavior)

(2) Bayesian sparsification (Graves [2011], Titsias and L´azaroGredilla [2014])

Uses priors like spike-and-slab

Limitations:

Fixed sparsity structure

Limited flexibility

01

Introduction

Key Contributes

(1) Variational Ising-based Regularization

Performs structured pruning and model selection simultaneously

Learns the model structure rather than simply removing weights

(2) Uncertainty in Attention

Extends uncertainty quantification to attention mechanisms

Enables uncertainty-aware feature selection and improved interpretability

(3) Computational Efficiency

Linear complexity in the number of parameters

Easily integrates into standard transformer training pipelines

02

Methods

Model Definition

(1)Dataset

$$D = \{(y_n, X_n)\}_{n=1}^{N}$$

y_n: binary label

X_n: input image

(2)Model Definition

$$P(y_n = 1 \mid X, W, \xi) = f_{y,W,\xi}(X)$$

(3)Components

- W: All model parameters
- ξ (key component): Binary dropout mask applied to weights
- $\xi_{ij} = 0$: weight removed, $\xi_{ij} = 1$: weight retained

02

Methods

Variational Bayes for Weights

(1) Bayesian Setup

$$W \sim p(W)$$

- Standard: Gaussian prior

- Spike-and-slab: sharp spike + wide slab

$$p(W) = \pi \mathcal{N}(W; 0, \sigma_1^2) + (1-\pi)\mathcal{N}(W; 0, \sigma_2^2), \quad \sigma_1^2 \ll \sigma_2^2$$

- Limitation: fixed mixture

(2) Proposed Prior

$$p(W \mid \xi) = \prod_{j,j'} \left[\xi_{j,j'} \, \mathcal{N}\left(W_{j,j'}; 0, \sigma_1^2\right) + \left(1 - \xi_{j,j'}\right) \mathcal{N}\left(W_{j,j'}; 0, \sigma_2^2\right) \right]$$

02

Methods

Variational Bayes for Weights

(3) Target Posterior

$$p(W, \xi \mid X, y)$$

- Intractable

(4) Variational Inference

$$q(W, \xi) \approx p(W, \xi \mid X, y)$$

(5) Objective

- To minimize KL Divergence

$$\mathrm{L} = -\sum_{n=1}^{N} \log p(y_n \mid X_n, W, \xi) + \mathrm{KL}\big(q(W \mid \xi) \mid p(W \mid \xi)\big) + \mathrm{KL}\big(q(\xi) \mid p(\xi)\big)$$

Methods

Variational Bayes for Weights

(6) Monte Carlo Sampling

- sample W, ξ

$$min - \sum_{n} \log p(y_n | X_n, \mathrm{W}, \xi) + \mathrm{KL}(q(W|\xi) | p(W|\xi)) + \mathrm{KL}(q(\xi) | p(\xi))$$

- cross entropy + KL regularization

(7) Variational Design

$$q_M(W|\xi) = \prod_{j,j'} \left[\xi_{j,j'} \mathcal{N}\left(W; m_{j,j'}, \sigma^2\right) + \left(1 - \xi_{j,j'}\right) \mathcal{N}(W; 0, \sigma^2)\right]$$

- Mask ξ: depends on likelihood difference

(8) Ising-style Update of ξ

- We update the binary mask variable ξ using a likelihood-guided variational formulation:

$$q\left(\xi_{j',j}^{(l)}=1\right)=\left[1+exp\left(-\frac{1}{2}\frac{\sum_{j''}w_{j'',j'}^{2}\ \mathbb{E}\left[\xi_{j'',j'}^{(l+1)}\right]}{\sum_{j''}w_{j'',j'}^{2}}-\left(L_{j}^{+}-L_{j}^{-}\right)\right)\right]^{-1}$$

- This update resembles an Ising model:
  - Neighboring variables ξ are not independent
  - Encourages structured sparsity
- Adjacent connections tend to be activated or deactivated together

Methods

Variational Bayes for Weights

(9) Likelihood Difference Term

$$L_j^+ = \sum_{n=1}^{N} \log p\left(y_n \mid X_n, W, \xi_{j',j} = 1 \right)$$

$$L_j^- = \sum_{n=1}^{N} \log p\left(y_n \mid X_n, W, \xi_{j',j} = 0 \right)$$

- $L_j^+ - L_j^-$ measures the impact of removing a weight on the predictive likelihood

- quantifies how much the likelihood decreases if the weight is removed

Methods

Variational Bayes for Weights

(10) Computational Challenge

- Exact computation requires: Forward pass for each configuration of ξ, Complexity: $O(2^{|\xi|})$
- Intractable

(11) Second-order Taylor Approximation (Similar to Laplace Approximation)

- To avoid exhaustive computation, we approximate:

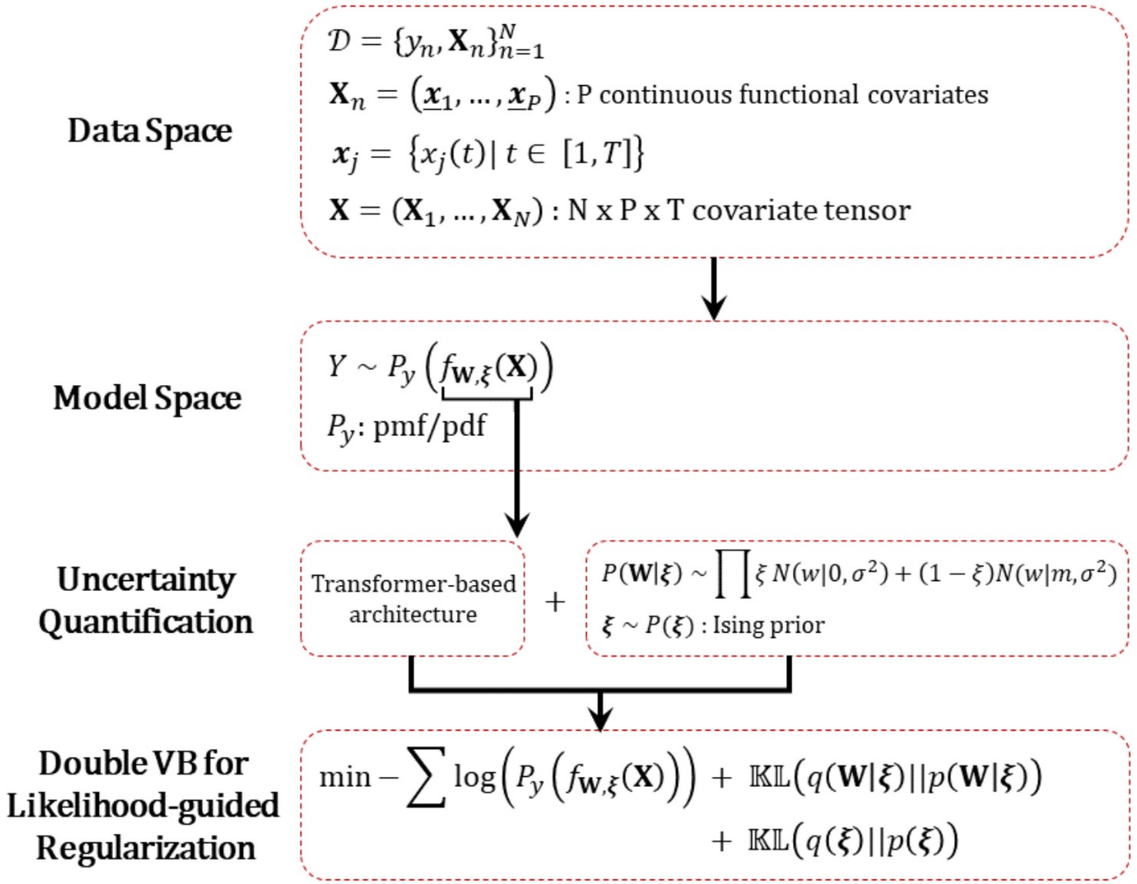
$$L_j^+ - L_j^- \approx \frac{\partial^2 L(q)}{\partial w_{j',j}^2}$$

or equivalently,

$$L_j^+ - L_j^- \approx \frac{\partial^2 L(q)}{\partial a_{j'}^2} x_j^2$$

02 Methods

Flow Chart



03 Algorithm

Algorithm 1 Model fitting via Likelihood-guided Regularization

1: Given the architecture and its initializing hyperparameters

2: Train the model by backpropagating the negative log likelihood,

$$-\log[p(\mathbf{y_n}|\mathbf{X_n}, \widehat{\mathbf{W}}, \hat{\boldsymbol{\xi}}))]$$

onto the means, $\mathbf{M}$ of the proposed variational posterior on $\mathbf{W}$ while setting $\boldsymbol{\xi} = \mathbf{1}$ until a reasonable minimum of the loss function is obtained

3: For each new epoch compute the diagonal of the Hessian for each element of $\mathbf{M}$

4: Recover the saliency approximation for each parameter in $\mathbf{M}$ as

$$L_j^+ - L_j^- \approx \frac{\partial^2 L(q)}{\partial w_{j',j}^2},$$

5: Sample $\boldsymbol{\xi}$ from the proposed variational posterior

$$q(\xi_{j',\,j}^{(l)} = 1) = \left[1 + \exp\left\{ -2\frac{\sum_{j'' \in\,(l+1)} w_{j'',j'}^2 E_q[\xi_{j'',j'}^{(l+1)}]}{\sum_{j'' \in\,(l+1)} w_{j'',j'}^2} - (L_j^+ - L_j^-) \right\}\right]^{-1},$$

6: Apply the sampled mask, do a forward pass then backpropagate the error

7: Sample $\mathbf{W}$ from the updated variational posterior

$$q_{\mathbf{M}}(\mathbf{W}|\boldsymbol{\xi}) = \prod_{j,j'} \xi_{j,j'} \mathcal{N}(\mathbf{W}; 0, \sigma^2) + (1 - \xi_{j,j'}) \mathcal{N}(\mathbf{W}; m_{j,j'}, \sigma^2),$$

8: Iterate through steps 3-6 until a reasonable minimizer is attained

04 Properties

Propositions

Proposition 4.1 *For any input* $\mathbf{X}$*, we have the following properties of our likelihood-guided regularization approach:*

(a) Computing the model output $f_{\mathbf{y},\mathbf{W},\boldsymbol{\xi}}(\mathbf{X})$ *is equivalent to performing a forward pass on the input* $\mathbf{X}$ *and applying the drop with probability* $pr(\xi_{j,j'} = 1)$ *in likelihood-guided regularization of attention based model.*

(b) The posterior $p(y, \mathbf{X}) \approx \frac{1}{T}\sum_{t=1}^{T} f_{\mathbf{y},\widehat{\mathbf{W}}_\mathbf{t},\hat{\boldsymbol{\xi}}_\mathbf{t}}(\mathbf{X})$*, where* $f_{\mathbf{y},\widehat{\mathbf{W}}_\mathbf{t},\hat{\boldsymbol{\xi}}_\mathbf{t}}(\mathbf{X})$ *is the output of the* t^{th} *forward pass through the likelihood-guided regularization approach.*

(c) The prediction uncertainty over the prediction $p(\mathbf{y}^*|\mathbf{X}^*)$ *can be approximated by the Bayesian lower and upper credible intervals of the* T *forward passes.*

This proposition describes how our approach can provide uncertain quantification effectively.

Proposition 4.2 *If* $pr(\xi_{j,j'} = 1) = \pi$ *for all* (j, j') *and* $q(\boldsymbol{\xi}) = O(1)$*, our likelihood-guided regularization of attention based model aligns with BayesFormer[Sankararaman et al., 2022].*

This property describes our approach is a generalization of existing BayesFormer.

05

Experiments

Setup

(1) Models

- Ising-ViT (proposed)
- Bayesian ViT (fixed dropout)
- Bayesian ViT (fixed dropconnect)

(2) Datasets

- MNIST , FashionMNIST, CIFAR-10, CIFAR-100

(3) Settings

- 3 training sizes

(4) Metrics

- Accuracy, Recall, Precision, FPR, F1

05

Experiments

Results

(1) Performance

- Dropconnect performs well in small data
- Ising excels in complex + low data settings

(2) Uncertainty - Dropout → overconfident

- Dropconnect → moderate
- Ising → best calibrated

06

Conclusion

Conclusion

(1) Key mechanism

- Learns dropout probabilities from data
- No manual tuning
- Produces posterior over weights, network structure

(2)Results

- Comparable accuracy
- Better:
  - entropy structure
  - calibration

06

Conclusion

Limitation and Future Work

(3) Limitation

- Conservative predictions (underconfidence)

(4) Future work

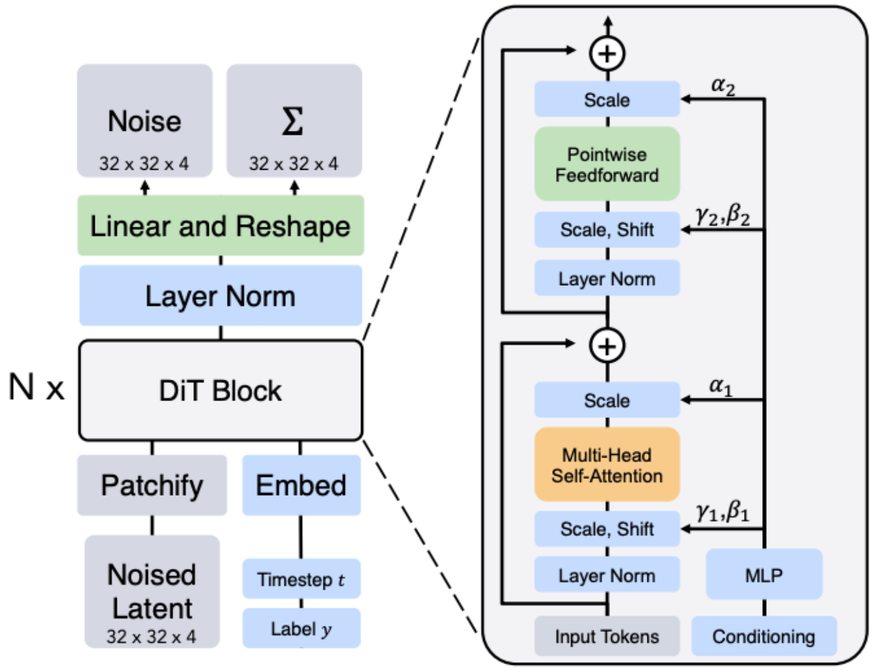
- Improve priors (lower prior weight)
- Better balance confidence & calibration

06

Conclusion

Application to My Research

Apply to Diffusion Transformer



THANK YOU