Fine-Tuning Pretrained Models with NVIB for Improved Generalisation

Anonymous ACL submission

Abstract

Fine-tuned pretrained attention-based models often struggle with generalisation, leading to poor performance in scenarios like out-of-domain transfer, distribution shifts, and few-shot learning. This limitation is prevalent across modalities such as speech, text, graphs, and vision. Nonparametric Variational Information Bottleneck (NVIB) is an attention-based information-theoretic regulariser applicable to pretrained models that has been shown to improve generalisation. However, prior work has applied NVIB only to the text modality and without fine-tuning. We investigate whether NVIB's ability to remove information from pretrained embeddings helps the model avoid spurious correlations with noisy and superficial features during fine-tuning. We are the first to integrate NVIB regularisation during fine-tuning across multiple diverse models and modalities. This required modifications to the architecture which enhance adaptability and stability during fine-tuning and simplify the evaluation. We found improved out-of-distribution generalisation in: speech quality assessment and language identification, text with induced attention sparsity, graph-based link prediction, and image-based tasks, including few-shot classification and privacy classification.

1 Introduction

Leveraging pretrained attention-based representations by fine-tuning has become the de facto modelling paradigm due to its wide applicability and significant improvements on the state-of-the-art (Ruder et al., 2019). Applications of pretrained Transformers (Vaswani et al., 2017) are modality agnostic and gained prevalence across: speech processing (Baevski et al., 2020; Radford et al., 2023); natural language processing (Devlin et al., 2019; Raffel et al., 2020; Touvron et al., 2023), graphs (Rong et al., 2020; Li et al., 2021b) and computer vision (Liu et al., 2021; Dosovitskiy et al., 2021; Bao et al., 2022).

The success of pretrained attention-based models is thought to stem from their ability to scale, both in terms of corpora size and the number of parameters, as well as the inductive biases inherent in the attention-based architecture (Henderson, 2020; Zhai et al., 2022; Fedus et al., 2021; Dehghani et al., 2023). Despite their success, these models still exhibit notable limitations during fine-tuning. Due to their large number of parameters and expressivity, they can be prone to overfitting and struggle to generalise in the presence of shortcuts from spurious correlations (Bhargava et al., 2021; Geirhos et al., 2020), distribution shift (Wu et al., 2020a; Kumar et al., 2022). The attention mechanism facilitates expressivity through token interaction, but this also introduces redundant information, which can hinder generalisation (Bian et al., 2021; Bhojanapalli et al., 2021). Introducing sparsity as a form of regularisation into attention has been shown to improve generalisation performance by reducing this redundancy (Child et al., 2019; Behjati et al., 2023; Fehr and Henderson, 2024). However, regularising attention during fine-tuning of pretrained models remains both challenging and unexplored.

Information bottleneck (IB) is an information-theoretic regulariser that learns latent features Z that compress the input X while preserving information for the downstream task Y (Tishby et al., 2000). The variational information bottleneck (VIB) framework, introduced through a variational lower bound to the IB objective (Alemi et al., 2017), enables deep neural representations (Tishby and Zaslavsky, 2015) to be trained using gradient-based optimisation. This framework has been widely applied across speech (Nelus and Martin, 2021; Lian et al., 2022), natural language (McCarthy et al., 2020; mahabadi et al., 2021), graphs (Wu et al., 2020b; Sun et al., 2022) and vision (Han et al., 2020; Chun, 2024). The success of the VIB framework can be attributed to its key properties, including resilience against spurious correlations

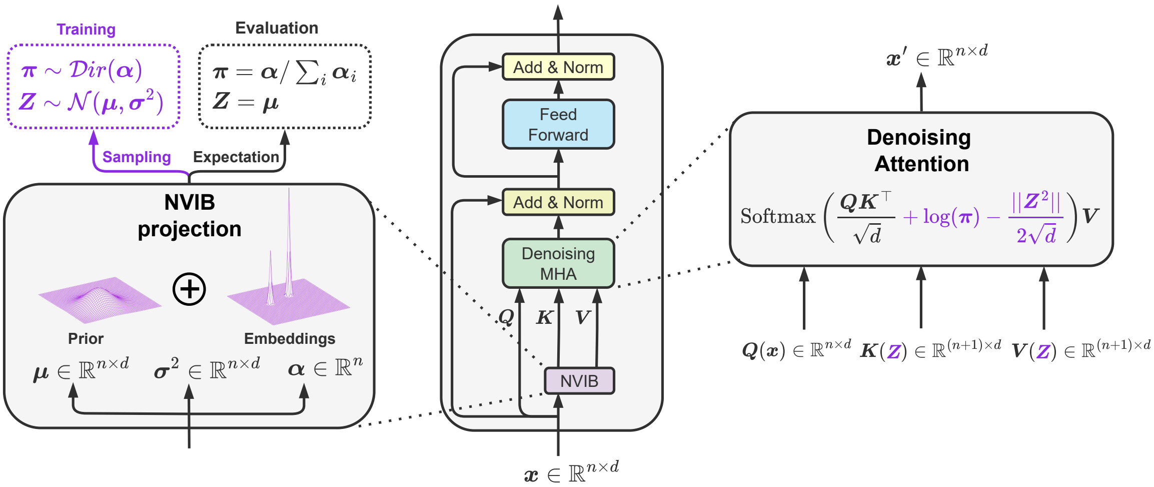
Training
Evaluation
$\pi \sim \mathcal{D}ir(\alpha)$
$Z \sim \mathcal{N}(\mu, \sigma^2)$
$\pi = \alpha / \sum_i \alpha_i$
$Z = \mu$
Sampling
Expectation
NVIB projection
Prior
Embeddings
$\mu \in \mathbb{R}^{n \times d}$
$\sigma^2 \in \mathbb{R}^{n \times d}$
$\alpha \in \mathbb{R}^{n}$
Add & Norm
Feed Forward
Add & Norm
Denoising MHA
Q
K
V
NVIB
$x \in \mathbb{R}^{n \times d}$
$x' \in \mathbb{R}^{n \times d}$
Denoising Attention
$\mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + \log(\pi) - \frac{||Z^2||}{2\sqrt{d}}\right)V$
$Q(x) \in \mathbb{R}^{n \times d}$
$K(Z) \in \mathbb{R}^{(n+1) \times d}$
$V(Z) \in \mathbb{R}^{(n+1) \times d}$

Figure 1: The NVIB module including the NVIB layer (left) and denoising attention (right).

(Chuah et al., 2022) and distribution shift (Li et al., 2021a), robustness (Zhang et al., 2022) and sparsity (Paranjape et al., 2020). Despite this success, VIB regularisation has seen limited exploration in the fine-tuning of pretrained attention-based models. Applying VIB to these pretrained models is difficult due to the complexity of incorporating it into the variable-sized latent representations accessed by attention.

Henderson and Fehr (2023) propose Nonparametric Variational Information Bottleneck (NVIB) as a VIB regulariser for attention layers. NVIB regularises the variable-sized representations accessed by attention by compressing both the information in individual vectors and the number of vectors. Further contributions to NVIB have demonstrated characteristics such as out-of-distribution (OOD) generalisation, robustness and sparsity (Henderson and Fehr, 2023; Behjati et al., 2023; Fehr and Henderson, 2024). Behjati et al. (2023) employ NVIB for representation learning by incorporating the regulariser into the self-attention layers of a Transformer-based encoder, and trains from scratch to progressively learn sparser representations through its layers. Fehr and Henderson (2024) integrated NVIB into pretrained models and achieved improvements in OOD summarisation and translation tasks without further training. Previous work has not applied NVIB regularisation during fine-tuning of pretrained models, nor has it explored generalising nonparametric variational models beyond text to diverse modalities like vision, speech, and graphs with their varying model architectures, data, and tasks.

Contributions. In this paper, we are the first to extend NVIB regularisation methods to fine-tuning, with diverse pretrained models. (1) We propose several novel methods for NVIB fine-tuning, including a learnable prior mean embedding per layer for adaptability, clipped Dirichlet pseudo-counts for stability, and a simplified denoising attention function at evaluation (Section 2). (2) We do the first empirical evaluation of NVIB on diverse modalities such as speech (Section 3.1), text (Section 3.2), graphs (Section 3.3), and vision (Section 3.4 and 3.5). (3) We show improved OOD generalisation in classification and regression tasks, demonstrating NVIB's added value across diverse applications.

2 Fine-tuning with NVIB

Figure 1 depicts an NVIB module, with the NVIB layer (left) and denoising attention function (right). The NVIB layer projects the sequence of vectors $\boldsymbol{x} \in \mathbb{R}^{n \times d}$ from a Transformer embedding to the parameters of a Dirichlet Process. These parameters include the isotropic Gaussian means $\boldsymbol{\mu} \in \mathbb{R}^{(n+1) \times d}$ and variances $\boldsymbol{\sigma}^2 \in \mathbb{R}^{(n+1) \times d}$ specifying the mixture base distribution, and the Dirichlet concentration parameters $\boldsymbol{\alpha} \in \mathbb{R}^{(n+1)}$. Each of the n vectors has an associated mixture component, along with an additional $(n+1)^{\text{th}}$ component that serves as a prior for the embeddings. During training, the NVIB layer samples a mixture distribution, represented as a set of weighted vectors $(\boldsymbol{\pi}, \boldsymbol{Z})$, where $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ and $\boldsymbol{Z_i} \sim \mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\sigma_i}^2)$. During evaluation, the NVIB layer outputs the expectation of these samples, which is the mixture of $n+1$ Gaussians, but can be approximated as $\boldsymbol{Z} = \boldsymbol{\mu}$ and $\boldsymbol{\pi} = \boldsymbol{\alpha} / \sum_i^n \alpha_i$.

Figure 1 (right) depicts how the denoising attention function is a generalisation of standard attention to any nonparametric mixture distribution. In the case of a set of weighted vectors, this involves using the weights $\boldsymbol{\pi}$ as bias terms for the attention weights over keys $K(\boldsymbol{Z})$. We provide a detailed description and pseudocode for denoising attention in Appendix B, and a consolidated overview of prior research on NVIB in Appendix A.

Following from Fehr and Henderson (2024), we reinterpret the pretrained models as nonparametric variational models by including NVIB layers before the attention mechanisms. This layer maps the input vectors $\boldsymbol{x}$ to the DP parameters $(\boldsymbol{\mu}^q, \boldsymbol{\sigma}^q, \boldsymbol{\alpha}^q)$:

$$\boldsymbol{\mu} \;=\; \mu(\boldsymbol{x}) \;= \boldsymbol{x}\boldsymbol{W}^{\mu} + \boldsymbol{b}^{\mu} \qquad (1)$$

$$\boldsymbol{\sigma}^2 = \sigma^2(\boldsymbol{x}) = \exp(\boldsymbol{x}\boldsymbol{W}^{\sigma} + \boldsymbol{b}^{\sigma}) \qquad (2)$$

$$\boldsymbol{\alpha} \;=\; \alpha(\boldsymbol{x}) \;= \exp(\boldsymbol{x}^2\boldsymbol{w}_1^{\alpha} + \boldsymbol{x}\boldsymbol{w}_2^{\alpha} + b^{\alpha}) \qquad (3)$$

The weight and bias parameters are initialised as: $\boldsymbol{W}^{\mu} = \boldsymbol{I}$ and $\boldsymbol{b}^{\mu} = \boldsymbol{0}$, ensuring an identity initialisation for $\boldsymbol{\mu}$ (Equation 1). For $\boldsymbol{\sigma}^2$, we set $\boldsymbol{W}^{\sigma} = \boldsymbol{0}$ and $\boldsymbol{b}^{\sigma} = \log(\tau_{\sigma}^2)$, which initialises the variance (Equation 2). Finally, the parameters for $\boldsymbol{\alpha}$ (Equation 3) are given by $\boldsymbol{w}_1^{\alpha} = \frac{1}{2\sqrt{d/h}} \odot \boldsymbol{1}$, $\boldsymbol{w}_2^{\alpha} = \boldsymbol{0}$, and $b^{\alpha} = \tau_{\alpha}$ which allows for a constant bias term in denoising attention in Figure 1 (right). This initialisation ensures empirical equivalence with the pretrained model, after manual adjustment of the hyperparameters $(\tau_{\sigma}^2, \tau_{\alpha})$ for each model, where d and h denote the projection size and number of attention heads. During fine-tuning, all model parameters are updated, including $\boldsymbol{W}^{\mu}$, $\boldsymbol{b}^{\mu}$, $\boldsymbol{W}^{\sigma}$, $\boldsymbol{b}^{\sigma}$, $\boldsymbol{w}_1^{\alpha}$, $\boldsymbol{w}_2^{\alpha}$, and b^{α}.

To fine-tune with NVIB regularisation, we add Kullback-Leibler (KL) divergence terms to the task loss. As with previous VIB regularisers, information flow is controlled during training by sampling the latent representations. Minimising the KL divergence with the prior tries to maintain this sampling noise and remove information, while the task loss keeps the information needed for the task. The task loss $\mathcal{L}_T$ is computed with the sampled representations. With NVIB, the KL divergence is decomposed into two loss terms: $\mathcal{L}_G$ for the Gaussians and $\mathcal{L}_D$ for the Dirichlet distributions, with hyperparameters λ_G and λ_D controlling their balance. The corresponding equations from Henderson and Fehr (2023) are provided in Appendix A.3. This gives us a total fine-tuning loss of:

$$\mathcal{L} = \mathcal{L}_T \; + \; \lambda_D\mathcal{L}_D \; + \; \lambda_G\mathcal{L}_G \qquad (4)$$

Novel methods for NVIB fine-tuning. Firstly, in contrast to Fehr and Henderson (2024), we simplify the denoising function during evaluation to better align with the training function. The equations used in both training and evaluation are shown in Figure 1 (right) and pseudocode in Appendix B. Secondly, while Fehr and Henderson (2024) estimate the prior parameters from training data, in this work we allow the prior mean $\boldsymbol{\mu}^p$ to be fine-tuned. This allows for flexibility and adaptation to the pretrained model. To maintain the noise in the prior during training, we keep the prior variance $(\boldsymbol{\sigma}^p)^2 = \boldsymbol{1}$ and the prior pseudo-count $\alpha_0^p = 1$ fixed. Thirdly, we stabilise fine-tuning by applying proportional clipping to the Dirichlet sampling parameters $\boldsymbol{\alpha}$. The magnitude of $\boldsymbol{\alpha}$ controls the amount of noise when sampling the weights $\boldsymbol{\pi}$, with larger values reducing noise. The relative values of $\boldsymbol{\alpha}$ determine the expected $\boldsymbol{\pi}$ distribution. Thus, we control the magnitude of $\boldsymbol{\alpha}$ while preserving its relative values using the clipping functions $\max(\epsilon, .)$ and $\min(\omega, .)$ to prevent underflow and overflow, respectively. The parameter ϵ is set small enough to prevent values from vanishing, while ω is chosen to be sufficiently large to avoid distorting the distribution.

$$\boldsymbol{\alpha} = \max\left(\epsilon, \frac{\boldsymbol{\alpha}}{\sum_i \alpha_i}\right) \times \min\left(\omega, \sum_i \alpha_i\right) \quad (5)$$

3 Experiments

To evaluate the NVIB regulariser, we design controlled experiments by fine-tuning pretrained models across modalities, including speech, text, graphs, and vision. We compare to models that are first pretrained and then fine-tuned using empirical risk minimization (ERM) with task-specific loss functions.

Baselines. For simplicity and to maintain uniformity across experiments, we define a set of fine-tuned baselines, avoiding modality-specific alternatives. These baselines include models trained without regularisation and models with dropout regularisation. To ensure consistency with standard practices, we use the predefined dropout rate of 0.1 for all pretrained models. Dropout is a suitable baseline for NVIB regularisation, as it is widely used and effective, seamlessly integrates into pretrained models, and introduces noise into both embeddings and attention mechanisms. All experiments are conducted on a consumer-grade NVIDIA RTX 3090 (24GB) GPU, with smaller Transformer models chosen to reduce computational costs.

Table 1: Speech quality assessment for NISQA (ID) and Tencent (OOD). Average test results (0–1) are reported with standard deviation across 5 seeds.

	NISQA (ID)		Tencent (OOD)	
Model	**PCC** (↑)	**RMSE MAP** (↓)	**PCC** (↑)	**RMSE MAP** (↓)
W2V2$_{\text{Base}}$	0.89 (0.02)	0.42 (0.03)	0.80 (0.01)	0.54 (0.01)
with Dropout	0.89 (0.01)	0.43 (0.01)	**0.83** (0.03)	**0.51** (0.04)
with NVIB	**0.90** (0.01)	**0.41** (0.02)	**0.83** (0.02)	**0.51** (0.03)

Table 2: Language identification for CommonLanguage (ID), FLEURS (OOD) and VoxPopuli (OOD) speech datasets. Average test F1 scores (0–1) are reported with standard deviation across 5 seeds.

	CommonLanguage (ID)	FLEURS (OOD)	VoxPopuli (OOD)
Model	**F1** (↑)	**F1** (↑)	**F1** (↑)
W2V2$_{\text{Large}}$	**0.82** (0.01)	0.90 (0.02)	**0.86** (0.02)
with Dropout	0.81 (0.01)	0.90 (0.01)	0.82 (0.02)
with NVIB	**0.82** (0.01)	**0.91** (0.02)	0.85 (0.02)

Initialisation of NVIB layers. The initialisation ensures empirical equivalence with each pretrained model after adjusting $(\tau_\sigma^2, \tau_\alpha)$, allowing the attention weights to ignore the prior component in NVIB layers. While Fehr and Henderson (2024) empirically initialise the prior component, we simplify this by setting $\boldsymbol{\mu}^p = \mathbf{0}$, $(\boldsymbol{\sigma}^p)^2 = \mathbf{1}$, and $\alpha_0^p = 1$. During fine-tuning, $\boldsymbol{\mu}^p$ remains learnable, while $(\boldsymbol{\sigma}^p)^2$ and α_0^p are fixed. However, stacking NVIB across layers in deeper models reduces equivalence precision. In such cases, NVIB is omitted from the later layers of the model. The parameters τ_σ and τ_α influence equivalence during training and evaluation.τ_σ controls initial Gaussian noise during fine-tuning but is unnecessary for evaluation equivalence, as mean embeddings are used. τ_α reweights Dirichlet pseudo-counts to ensure input embeddings outweigh the prior in attention.

Fine-tuning hyperparameters. Following Henderson and Fehr (2023), we set the number of samples per component to one but omit their conditional prior, which prevents posterior collapse during training from scratch. The KL divergence is weighted by λ_G and λ_D, respectively. We optimize these hyperparameters through a log-scaled grid search $[10^{\,0}, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}]$, tying λ_G and λ_D and selecting the best values based on validation performance.

3.1 Speech Out-of-Distribution Evaluation

Language identification and automated assessment of speech are crucial tasks in the development of audio transmission systems, but are challenging due to many factors related to: the acoustic environment; variation in recording hardware and software; speaker characteristics; and evaluation conditions (Gierlich and Kettler, 2006; Chinen, 2021; Cooper et al., 2022). The prediction of perceived speech quality is formulated as a regression task to estimate the scores of human listeners (ITU-T, 1996), whereas language identification is a classification task given an audio sample. Given the diverse array of factors that can impact speech, generalisation is essential in these tasks.

Speech quality assessment. We fine-tune and evaluate on the NISQA (Mittag et al., 2021) dataset, which contains English speech recordings from live calls with network impairments and simulated distortions. We perform OOD testing on the TencentWithReverberation (Tencent) Chinese speech corpus (Yi et al., 2022), which introduces new conditions such as: simulated and real reverberation; and different labelling conditions. Following ITU-T (2020), we evaluate our models using the Pearson's correlation coefficient (PCC) and root-mean-square error after mapping with a first-order polynomial function (RMSE MAP).

We fine-tuned the pretrained Wav2vec2-base model (Baevski et al., 2020), a 12-layer Transformer encoder, using mean-squared-error (MSE) loss. Fine-tuning was conducted with the Adam optimizer (Kingma and Ba, 2014), a constant learning rate of $1e^{-5}$, a batch size of 16, and for 5 epochs. NVIB was applied to layers 0–10, with projections initialized using $\tau_\sigma = 0.1$ and $\tau_\alpha = 10$. The best-performing model used $\lambda_G = \lambda_D = 1e^{-2}$. Table 1 shows that NVIB regularisation achieves the highest correlation on the in-distribution (ID) data. On the OOD dataset, NVIB regularisation achieves comparable generalisation improvements while exhibiting a lower standard deviation.

Speech language identification. We fine-tune our models on the CommonLanguage (Ravanelli et al., 2021) dataset which consists of 22K training audios from 45 languages. We evaluate on two OOD datasets with overlapping languages: FLEURS (Conneau et al., 2023) with 27 languages; and VoxPopuli (Wang et al., 2021) with 11 lanuages. The FLEURS dataset is read speech, which is closer to CommonLanguage. Whereas, the VoxPopuli dataset is more challenging as it contains spontaneous speech from the European Parliament.

We fine-tuned the pretrained Wav2vec2-large model (Baevski et al., 2020), a 24-layer Transformer encoder, using cross-entropy loss for the language identification classification task. Fine-tuning was performed with the AdamW optimizer (Loshchilov and Hutter, 2019), a learning rate of $3e^{-5}$, a scheduler with linear warm-up and decay, a batch size of 4, and for 10 epochs with mixed precision (16-bit) and gradient norm clipping of 1. NVIB was applied to layers 0–16, with projections initialized using $\tau_\sigma = 0$ and $\tau_\alpha = 10$. The best-performing model used $\lambda_G = \lambda_D = 1e^{-7}$. Table 2 reports the F1 classification scores, showing that NVIB matches ID performance and outperforms the dropout-regularised baseline on the OOD datasets.

3.2 Text Out-of-Distribution Classification

We consider the CivilComments (CC) (Borkan et al., 2019) task which is part of the WILDS (Koh et al., 2021) curated set of datasets that represent real-life distribution shifts. CC classifies the presence of toxicity in online comments which is an important task of monitoring internet content. The task is a binary classification task of determining if a comment is toxic, and contains a subpopulation shift between 8 demographic identities classes, meaning that the training and test domains overlap, but their relative proportions differ. We measure generalisation by the accuracy of the lowest performing subpopulation *worst-group* (WG).

We fine-tuned the pretrained TinyBERT model (Turc et al., 2019), a two-layer Transformer encoder, using cross-entropy loss. Fine-tuning was performed with the AdamW optimizer (Loshchilov and Hutter, 2019), a constant learning rate of $5e^{-5}$, a batch size of 1024, and for 50 epochs with mixed precision (16-bit) and gradient norm clipping of 0.1. NVIB was applied to all layers, with projections initialized using $\tau_\sigma = 0.1$ and $\tau_\alpha = 1$, and a linear KL annealing warmup was used during fine-tuning. The best-performing model used $\lambda_G = \lambda_D = 1e^{-1}$.

Table 3: Text classification on CC train (ID) and test (OOD). Average accuracy (%) is reported across 5 seeds with standard deviation and the *best* OOD model.

	CC Train (ID)	CC Test (OOD)
Model	**WG** (↑)	**WG** (↑)
BERT$_{\text{Tiny}}$	78.12 (14.33) *99.00*	49.14 (5.56) *61.03*
with Dropout	**91.05** (1.49) *91.16*	**60.10** (3.11) *63.97*
with NVIB	80.12 (10.69) *76.30*	55.01 (6.15) *61.03*

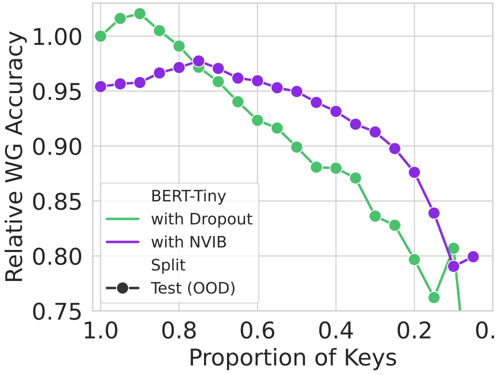

Figure 2: Worst-group (WG) test (OOD) accuracy as a function of attention key sparsity for the best OOD models, relative to dropout without sparsity.

Table 3 shows the generalisation improvement of this task through regularisation. On average, NVIB regularisation improves OOD generalisation over the unregularised baseline, though it remains less effective than dropout. However, introducing sparsity in the attention keys based on their attention magnitude, as shown in Figure 2, improves OOD accuracy and sustains it across a wide range of sparsity levels. NVIB naturally induces key sparsity by reducing the weight of embeddings relative to the prior component during attention calculations. To remove keys, we mask embeddings with the lowest average attention magnitudes. Further inspection of the attention patterns in Figure 3 reveals a clear focus on toxic words as spurious keys are dropped and attention shifts to the prior token. The alignment with toxic content becomes more pronounced as sparsity increases. Additional examples can be

Table 4: Graph link prediction on FB15k-237. Test set ranking metrics (0–1) are reported, based on the best model selected from validation set performance.

	FB15k-237			
Model	**MRR (↑)**	**H@1 (↑)**	**H@3 (↑)**	**H@10 (↑)**
BLP-BERT$_{\text{Tiny}}$	0.164	0.100	0.175	0.288
with Dropout	0.162	0.097	0.172	0.288
with NVIB	**0.167**	**0.103**	**0.180**	**0.294**

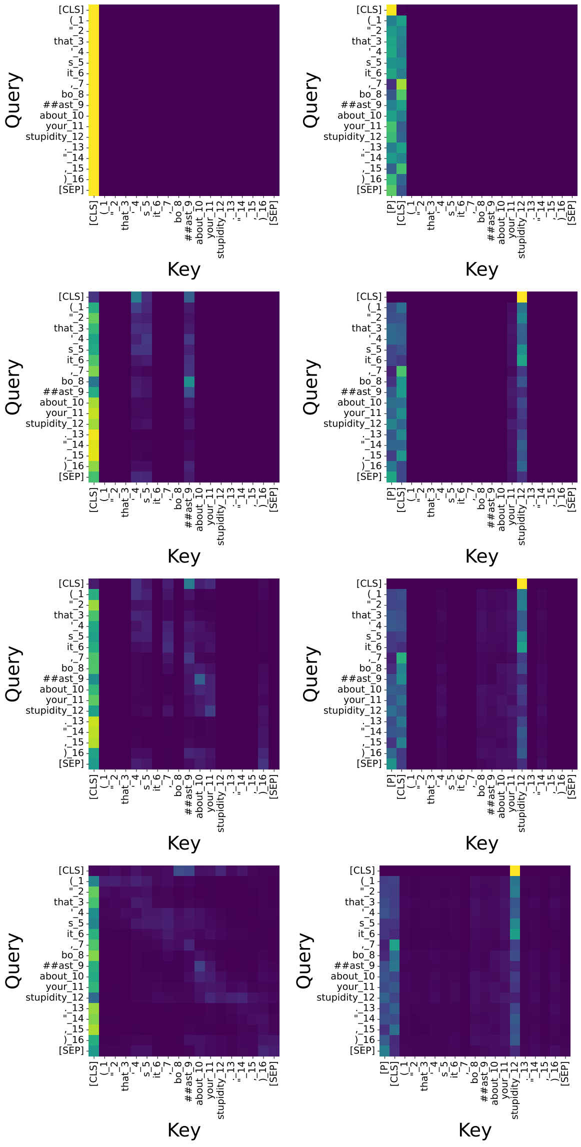

Figure 3: Attention plot for the best models on Civil-Comments showing a single head of the last layer. Left: with dropout, Right: with NVIB. Top-Bottom: Proportion of keys retained [0.1, 0.25, 0.5, 1.0]. Sentence: ("that's it, boast about your stupidity.").

found in Appendix Figures 6, 7 and 8.

3.3 Graph Link Prediction

Link prediction is a graph-based problem that involves predicting whether a link exists between two nodes in a graph. This is widely used for recommendation and prediction in social networks, citation links and biological interactions (Kumar et al., 2020; Xia et al., 2021). We build upon the BERT for Link Prediction (BLP) model (Daza et al., 2020) which operates on a set of triples (h, r, t), where h and t represent the head and tail node, while r represents the relation between those two nodes.

We evaluate on the FB15k-237 dataset (Daza et al., 2020). This dataset follows an inductive setting, where new entities and triples are dynamically incorporated into the graph during evaluation. We evaluate the models by querying them with $(h, r, ?\,)$ and $(?\,, r, t)$ triples, and assess their performance using two metrics: Mean Reciprocal Rank (MRR), which measures the model's ability to rank the correct triple, and H@k, which calculates the proportion of correct triples ranked within the top-k results. We fine-tuned the pretrained TinyBERT model (Turc et al., 2019), a two-layer Transformer encoder, using a distance-based TransE loss function. Fine-tuning was performed with the RAdam optimizer (Liu et al., 2020), a cosine learning rate scheduler with a value of $8e^{-5}$, a batch size of 256, and for 40 epochs with mixed precision (16-bit) and gradient norm clipping of 1. NVIB was applied to both layers, with projections initialised using $\tau_\sigma = 0.1$ and $\tau_\alpha = 1$. The best-performing model used $\lambda_G = \lambda_D = 1e^{-3}$.

Table 4 presents the test set results, which highlights the advantage of the NVIB-regularised model over typical regularisation methods like dropout. This advantage may stem from the presence of new entities in the head or tail positions, which require a higher level of generalisation.

3.4 Image Few-Shot Classification

Few-shot classification aims to train models capable of classifying images with limited labelled examples per category. Meta-learning (Vinyals et al., 2016) achieves this by meta-training on several *episodes*, enabling generalisation to new tasks

Table 5: Image classification on CIFAR-FS (ID). Test episodes accuracy (%) with standard deviation.

	CIFAR-FS (ID)	
Model	**Acc** (↑)	**Std** (↓)
DeiT$_{\text{Small}}$	93.57	5.71
with Dropout	93.55	5.61
with NVIB	**93.88**	**5.58**

with previously unseen classes. To generalise effectively, the classifier must transfer knowledge from the training distribution to unseen testing distributions while avoiding spurious correlations and shortcuts (Zheng et al., 2024; Zhang et al., 2024).

The following experiments were conducted within a meta-learning-based few-shot classification framework (Hu et al., 2022), using the pre-trained DeiT-Small (Touvron et al., 2021), a 12-layer Transformer encoder, with cross-entropy loss. Fine-tuning was performed using the AdamW optimizer (Loshchilov and Hutter, 2019), a constant weight decay of 0.05, and a linear warm-up with cosine decay learning rate scheduler $1e^{-4}$. The model was trained for 50 epochs with mixed precision (16-bit) and a batch size of 1. For classification, we used the prototypical network (ProtoNet) (Snell et al., 2017), which dynamically creates class centroids for each episode and performs nearest centroid classification (Hu et al., 2022). In this experiment we initialised the prior $\mu^p = \mathbf{0}$ and did not allow it to be learnable.

Few-shot in-distribution. We evaluate the ID performance using the CIFAR-FS (Bertinetto et al., 2019) dataset. Following Hu et al. (2022), we conduct experiments in a 5-way, 5-shot setting. Each episode consists of a "support set" with 5 classes and 5 samples per class for training, and a "query set" with 5 classes and 15 examples per class for testing. The experiment includes 2000 episodes for meta-training and 2000 episodes for testing. NVIB was applied to layers 0–5, with projections initialized using $\tau_\sigma = 0$ and $\tau_\alpha = 0$. The best-performing model used $\lambda_G = \lambda_D = 1e^{-2}$.

Table 5 reports the average classification accuracy and standard deviation over all test episodes for CIFAR-FS in few-shot classification. Compared to the baseline and Dropout, we observe that NVIB regularisation improves accuracy with lower variance across all test episodes.

Few-shot out-of-distribution. To evaluate the OOD few-shot classification performance, we use the Meta-Dataset (Triantafillou et al., 2019). This benchmark is a diverse set of 10 image datasets, including, ImageNet-1k, MSCOCO (COCO), Traffic Signs (Sign), Describable Textures (DTD), FGVCx Fungi (Fungi), Omniglot, VGG Flower (Flower), CUB-200-2011 (CUB), FGVCAircraft (Acraft) and QuickDraw (QDraw). We meta-train the models on ImageNet-1k and then meta-test them on the remaining datasets.

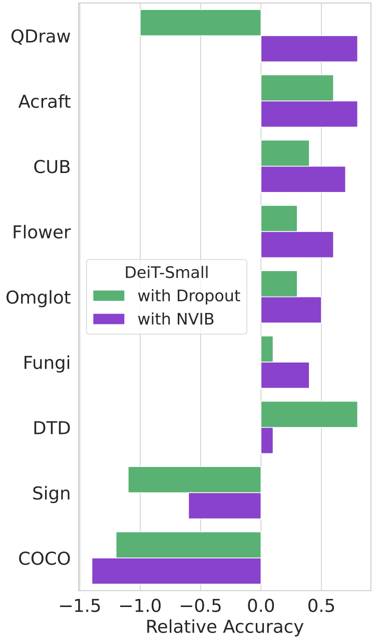

Figure 4: Percentage point improvement in test accuracy relative to the unregularised baseline on the Meta-Dataset benchmark (OOD).

We follow the methodology outlined in Hu et al. (2022), where the number of ways sampled ranges from 5 to 50, with a maximum support size of 500 and a maximum query size of 10. For our Transformer encoder, we apply NVIB to layers 0–5, with projections initialised using $\tau_\sigma = 0$ and $\tau_\alpha = -3$. The best-performing model used NVIB regularization parameters of $\lambda_G = \lambda_D = 1e^{-3}$. Figure 4 shows that the NVIB-regularised model achieves the highest performance on 6 out of 9 OOD datasets and outperforms the dropout-regularised model in 7 out of 9 cases.

3.5 Image Privacy Classification

Image privacy classification is a crucial task in safeguarding sensitive visual content, requiring models to accurately and robustly identify privacy-sensitive information. An effective classifier must generalise across data variations to minimize the risk of private information leakage. We consider the PrivacyAlert (Zhao et al., 2022) dataset, which contains images labelled as either private (25%) or public.

Table 6: Image privacy classification on PrivacyAlert. Average test F1 scores (%) are reported with standard deviation across 5 seeds.

	PrivacyAlert (ID)	
Model	**F1** (↑)	**Std** (↓)
DeiT$_{\text{Tiny}}$	78.41	0.10
with Dropout	76.26	2.47
with NVIB	**79.40**	**0.72**

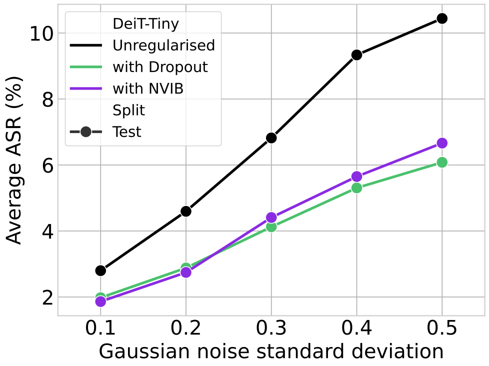

Figure 5: Average attack success rate (ASR) on privacy classifiers, reported across 5 models with varying Gaussian noise standard deviations. A lower ASR indicates greater model robustness.

The private category is largely composed of images from the Nudity/Sexual class, with limited representation from the Medical and Personal Information categories.

We fine-tuned the pretrained DeiT-Tiny (Touvron et al., 2021), a 12-layer Transformer encoder, with cross-entropy loss. Fine-tuning was conducted with the AdamW optimizer (Loshchilov and Hutter, 2019), a constant learning rate of $5e^{-6}$, a batch size of 32, and for 80 epochs with mixed precision (16-bit). NVIB was applied to all layers, with projections initialized using $\tau_\sigma = 0.5$ and $\tau_\alpha = 8$. The best-performing model used $\lambda_G = \lambda_D = 1e^{-3}$. We evaluate classification performance using F1 scores and assess robustness by perturbing images with zero-mean Gaussian noise at varying standard deviations. Robustness is measured using the attack success rate (ASR), which is the proportion of correctly classified images that are misclassified after perturbation. Table 6 shows that NVIB outperforms the dropout-regularised baseline. Additionally, Figure 5 indicates that NVIB achieves robustness comparable to dropout while offering improvements over an unregularised vision baseline.

4 Discussion

Our results suggest that including NVIB regularisation improves the model's ability to distinguish signal from noise. This is supported by performance gains observed in tasks such as speech quality prediction (Table 1) and few-shot and privacy image classification (Tables 5 & 6). We attribute this to NVIB's Bayesian nature, which effectively models statistical uncertainty. During fine-tuning, NVIB introduces noise into the latent representations, which enhances its ability to generalise across noisy feature spaces such as background disturbances and capture variations present in both audio and images.

NVIB regularisation shifts the model's attention from relying on superficial, spurious features to deeper features which generalise better out-of-distribution. This is evident in consistent improvements across tasks that require generalisation to unseen entities, such as graph linking (Table 4) and visual meta-learning (Figure 4). We believe this is due to the additional prior tokens, which disentangle and reweight attention away from spurious tokens (attention maps in Figures 3, 6, 7 & 8). Additionally, this effect is observed in sustained performance with increased sparsity and robustness (Figure 2 & 5).

5 Conclusion

In this work, we contribute to fine-tuning with Nonparametric Variational Information Bottleneck regularisation by demonstrating improved generalisation across multiple modalities and models. We extend NVIB to pretrained models by proposing a novel learnable prior mean embedding per layer for greater adaptability, clipping Dirichlet pseudo-counts for training stability, and simplifying the NVIB denoising attention function at evaluation time.

Future work. In future work, we aim to scale our experiments to include models with larger parameter sizes and explore training from scratch, where regularisation may be even more beneficial. While our current focus prioritized simplicity and uniformity, we are encouraged to evaluate additional baselines and tasks across each modality. Furthermore, we see significant promise in applying NVIB to language modelling, particularly with large language models (LLMs).

6 Limitations

Our experiments offer a broad exploration of NVIB regularisation across various models, tasks, and modalities, but they do not delve deeply into any one area. While NVIB shows effectiveness, the experiments were conducted on moderate-sized models, and future work should focus on scaling to larger models for a more thorough understanding. Additionally, the emphasis on simplicity and uniformity in the experimental design leaves room for exploring additional baselines and tasks across different domains.

The performance gains are relatively modest. In some cases, NVIB outperforms methods like dropout and consistently surpasses models without regularisation. A key finding of this work is that NVIB's regularisation behaviour resembles that of dropout when fine-tuned. However, the introduction of key sparsity opens up opportunities for future efficiency gains and enhanced interpretability. While NVIB adds complexity, we see this as an important step in understanding embedding distributions and their interactions through attention.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*, Toulon, France. OpenReview.net.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*.

Melika Behjati, Fabio James Fehr, and James Henderson. 2023. Learning to abstract with nonparametric variational information bottleneck. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. 2019. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Srinadh Bhojanapalli, Ayan Chakrabarti, Andreas Veit, Michal Lukasik, Himanshu Jain, Frederick Liu, Yin-Wen Chang, and Sanjiv Kumar. 2021. Leveraging redundancy in attention with reuse transformers. *ArXiv*, abs/2110.06821.

Yuchen Bian, Jiaji Huang, Xingyu Cai, Jiahong Yuan, and Kenneth Ward Church. 2021. On attention redundancy: A comprehensive study. In *North American Chapter of the Association for Computational Linguistics*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *ArXiv*, abs/1904.10509.

Michael Chinen. 2021. Marginal effects of language and individual raters on speech quality models. *IEEE Access*, 9:127320–127334.

Weiqin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, and David Suter. 2022. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13012–13022.

Sanghyuk Chun. 2024. Improved probabilistic image-text representations. In *The Twelfth International Conference on Learning Representations*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2022. Generalization ability of mos prediction networks. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8442–8446.

Daniel Daza, Michael Cochez, and Paul T. Groth. 2020. Inductive entity representations from text via link prediction. *Proceedings of the Web Conference 2021*.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine

Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. 2023. Scaling vision transformers to 22 billion parameters. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7480–7512. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

William Fedus, Barret Zoph, and Noam M. Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39.

Fabio James Fehr and James Henderson. 2024. Nonparametric variational regularisation of pretrained transformers. In *First Conference on Language Modeling*.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

H.W. Gierlich and F. Kettler. 2006. Advanced speech quality testing of modern telecommunication equipment: An overview. *Signal Processing*, 86(6):1327–1340. Applied Speech and Audio Processing.

Zongyan Han, Zhenyong Fu, and Jian Yang. 2020. Learning the redundancy-free features for generalized zero-shot object recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12862–12871.

James Henderson. 2020. The unstoppable rise of computational linguistics in deep learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6294–6306, Online. Association for Computational Linguistics.

James Henderson and Fabio James Fehr. 2023. A VAE for transformers with nonparametric variational information bottleneck. In *The Eleventh International Conference on Learning Representations*.

Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M. Hospedales. 2022. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *CVPR*.

ITU-T. 1996. Methods for subjective determination of transmission quality. Recommendation P.800, International Telecommunication Union, Geneva, Switzerland.

ITU-T. 2020. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. Recommendation P.1401, International Telecommunication Union, Geneva, Switzerland.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. Wilds: A benchmark of in-the-wild distribution shifts. *Preprint*, arXiv:2012.07421.

Ajay Kumar, Shashank Sheshar Singh, Kuldeep Singh, and Bhaskar Biswas. 2020. Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 553:124289.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Colorado Reed, Jun Zhang, Dongsheng Li, Kurt Keutzer, and Han Zhao. 2021a. Invariant information bottleneck for domain generalization. In *AAAI Conference on Artificial Intelligence*.

Pengyong Li, Jun Wang, Yixuan Qiao, Hao Chen, Yihuan Yu, Xiaojun Yao, Peng Gao, Guowang Xie, and Sen Song. 2021b. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Briefings in bioinformatics*.

Jiachen Lian, Chunlei Zhang, and Dong Yu. 2022. Robust disentangled variational speech representation learning for zero-shot voice conversion. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6572–6576.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Rabeeh Karimi mahabadi, Yonatan Belinkov, and James Henderson. 2021. Variational information bottleneck for effective low-resource fine-tuning. In *International Conference on Learning Representations*.

Arya D. McCarthy, Xian Li, Jiatao Gu, and Ning Dong. 2020. Addressing posterior collapse with mutual information for improved variational neural machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *Interspeech 2021*.

Alexandru Nelus and Rainer Martin. 2021. Privacy-preserving audio classification using variational information feature extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2864–2877.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Conference on Empirical Methods in Natural Language Processing*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying WEI, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, volume 33, pages 12559–12571. Curran Associates, Inc.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *North American Chapter of the Association for Computational Linguistics*.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*.

Qingyun Sun, Jianxin Li, Hao Peng, Jia Wu, Xingcheng Fu, Cheng Ji, and Philip S. Yu. 2022. Graph structure learning with variational information bottleneck. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4165–4174. 36th AAAI Conference on Artificial Intelligence, AAAI 2022 ; Conference date: 22-02-2022 Through 01-03-2022.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *ArXiv*, physics/0004057.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. 2019. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv: Computation and Language*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Jiawei Wu, Xiaoya Li, Xiang Ao, Yuxian Meng, Fei Wu, and Jiwei Li. 2020a. Improving robustness and generality of nlp models using disentangled representations. *ArXiv*, abs/2009.09587.

Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. 2020b. Graph information bottleneck. In *Advances in Neural Information Processing Systems*, volume 33, pages 20437–20448. Curran Associates, Inc.

Feng Xia, Ke Sun, Shuo Yu, Abdul Aziz, Liangtian Wan, Shirui Pan, and Huan Liu. 2021. Graph learning: A survey. *IEEE Transactions on Artificial Intelligence*, 2:109–127.

Gaoxiong Yi, Wei Xiao, Yiming Xiao, Babak Naderi, Sebastian Möller, Wafaa Wardah, Gabriel Mittag, Ross Culter, Zhuohuang Zhang, Donald S. Williamson, Fei Chen, Fuzheng Yang, and Shidong Shang. 2022. ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications. In *Proc. Interspeech 2022*, pages 3308–3312.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113.

Cenyuan Zhang, Xiang Zhou, Yixin Wan, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. 2022. Improving the adversarial robustness of NLP models by information bottleneck. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3588–3598, Dublin, Ireland. Association for Computational Linguistics.

Min Zhang, Haoxuan Li, Fei Wu, and Kun Kuang. 2024. Metacoco: A new few-shot classification benchmark with spurious correlation. In *The Twelfth International Conference on Learning Representations*.

Chenye Zhao, Jasmine Mangat, Sujay Koujalgi, Anna Squicciarini, and Cornelia Caragea. 2022. Privacyalert: A dataset for image privacy prediction. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1352–1361.

Guangtao Zheng, Wenqian Ye, and Aidong Zhang. 2024. Benchmarking spurious bias in few-shot image classifiers. In *European Conference on Computer Vision*.

A Introduction to NVIB

Henderson and Fehr (2023) define Nonparametric Variational Information Bottleneck (NVIB) by generalising the standard attention layer to a Bayesian model where embeddings are distributions over the latent space. A key insight of this approach is that the latent space of attention-based representations can be viewed as nonparametric mixture distributions. In this interpretation, the vectors accessed via attention define a mixture of impulse distributions. Since a Transformer embedding is a set of vectors that dynamically scale with the complexity of the input, the corresponding latent space of these mixture distributions is inherently nonparametric in nature. In this formulation, the attention function is interpreted as Bayesian "query denoising" using the latent distribution as the prior. The authors define *denoising attention* as a generalisation of the attention function to query denoising.

A.1 Denoising attention

Denoising attention is a generalisation of attention which interprets the latent space of Transformers as a nonparametric mixture distribution. Henderson and Fehr (2023) provide a constructive proof of exact equivalence to the standard attention function. When standard attention accesses the latent space of Transformers, which is a set of embedding vectors $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$ via weight matrices $\boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{d\times d}$ to keys and values, respectively, and projects the accessing input vector $\boldsymbol{u'} \in \mathbb{R}^{1\times d}$ via the weight matrix $\boldsymbol{W}^Q \in \mathbb{R}^{d\times d}$ to a query. By letting $\boldsymbol{u} = (\boldsymbol{u}'\boldsymbol{W}^Q(\boldsymbol{W}^K)^\top) \in \mathbb{R}^{1\times d}$, the standard scaled dot product attention function can be rewritten as $(\mathrm{Attn}(\boldsymbol{u}, \boldsymbol{Z})\ \boldsymbol{W}^V)$, with $\mathrm{Attn}(\boldsymbol{u}, \boldsymbol{Z})$ defined in terms of a sum over the vectors $\boldsymbol{z}_i$ in $\boldsymbol{Z}$, or equivalently defined in terms of an integral over a distribution which is only non-zero at the $\boldsymbol{z}_i$:

$$\begin{aligned}\mathrm{Attn}(\boldsymbol{u}, \boldsymbol{Z}) &= \mathrm{softmax}\left(\frac{1}{\sqrt{d}}\boldsymbol{u}\boldsymbol{Z}^\top\right)\boldsymbol{Z} \quad (6)\\ &= \mathrm{DAttn}(\boldsymbol{u}; F_{\boldsymbol{Z}})\end{aligned}$$

$$\mathrm{DAttn}(\boldsymbol{u}; F) = \int_{\boldsymbol{v}} \frac{f(\boldsymbol{v}) g(\boldsymbol{u}; \boldsymbol{v}, \sqrt{d}\boldsymbol{I})}{\int_{\boldsymbol{v}} f(\boldsymbol{v}) g(\boldsymbol{u}; \boldsymbol{v}, \sqrt{d}\boldsymbol{I})\, d\boldsymbol{v}} \boldsymbol{v} d\boldsymbol{v} \tag{7}$$

$$F_{\boldsymbol{Z}} = \sum_{i=1}^{n} \frac{\exp(\frac{1}{2\sqrt{d}}||\boldsymbol{z}_i||^2)}{\sum_{i=1}^{n} \exp(\frac{1}{2\sqrt{d}}||\boldsymbol{z}_i||^2)} \delta_{\boldsymbol{z}_i} \tag{8}$$

where $\delta_{\boldsymbol{z}_i}$ is an impulse distribution at $\boldsymbol{z}_i$, $f(\cdot)$ is the probability density function for distribution F, and $g(\boldsymbol{u};\ \boldsymbol{v}, \sqrt{d}\boldsymbol{I})$ is the multivariate Gaussian function with diagonal variance of $\sqrt{d}$. This alternative definition $\mathrm{DAttn}(\boldsymbol{u};\ F_{\boldsymbol{Z}})$ is *denoising attention*. It subsumes standard attention in that any attention-based embedding $\boldsymbol{Z}$ has an equivalent mixture of impulse distributions, namely $F_{\boldsymbol{Z}}$, where denoising attention $\mathrm{DAttn}(\boldsymbol{u};\ F_{\boldsymbol{Z}})$ gives us exactly the same result as attention $\mathrm{Attn}(\boldsymbol{u}, \boldsymbol{Z})$, for all queries $\boldsymbol{u}$. This is an elegant result, which in practice allows us to define a nonparametric distribution over the latent embeddings of Transformers. Appendix B covers the exact equations for denoising attention and pseudocode.

A.2 Distributions over mixture distributions

Given this generalisation of attention-based representations to nonparametric mixture distributions, Bayesian nonparametrics can be used to define distributions over the latent space. Henderson and Fehr (2023) propose to use Dirichlet Processes (DPs) to define distributions over mixture distributions, so an NVIB layer first embeds its input vectors into a DP representation by mapping them to the parameters $(\boldsymbol{\mu}^q, \boldsymbol{\sigma}^q, \boldsymbol{\alpha}^q)$ of a DP. A DP is defined by a base distribution G_0^q for generating the vectors for the component impulse distributions, and a pseudo-count α_0^q for generating their mixture weights.

$$\alpha_0^q = \sum_i \alpha_i^q \tag{9}$$

$$G_0^q = \sum_i \frac{\alpha_i^q}{\alpha_0^q} \mathcal{N}(\boldsymbol{\mu}_i^q, \boldsymbol{I}(\boldsymbol{\sigma}_i^q)^2) \tag{10}$$

Following this definition, G_0^q is itself a mixture distribution, consisting of one Gaussian component from the prior plus one Gaussian component for each vector input to the NVIB layer. These DPs represent the posterior $q(F|x)$). The prior $p(F)$ is a DP specified by the parameters $(\boldsymbol{\mu}^p, \boldsymbol{\sigma}^p, \alpha^p)$ of its pseudo-count α^p and its uni-modal base distribution $G_0^p = \mathcal{N}(\boldsymbol{\mu}^p, \boldsymbol{\sigma}^p)$. In this work, we allow the prior $\boldsymbol{\mu}^p$ to be learned, which allows the prior to be centred in the latent embedding space. However, to maintain noise during regularisation, we set the prior variance $(\boldsymbol{\sigma}^p)^2 = \boldsymbol{1}$ and the prior's pseudo-count $\alpha_0^p = 1$.

A.3 NVIB regularisation

During training, NVIB regularises the information passing through the NVIB layer by sampling latent representations from its DP embedding. This process introduces noise and removes redundant information, enhancing model generalisation. The level of noise is learned by the DP parameters $(\boldsymbol{\mu}^q, \boldsymbol{\sigma}^q, \boldsymbol{\alpha}^q)$ within the NVIB layer. To maintain noise during training, a Kullback-Leibler (KL) divergence loss term is included between the embedding distribution and the DP prior. Since the prior DP is input independent, the KL term enforces an information bottleneck by minimising the information retained in the DP embedding. During evaluation, the NVIB layer uses the mean latent representation, which is the base distribution G_0^q of the DP embedding.

The evidence lower bound (ELBO) is a widely used objective in variational Bayesian methods, serving as a tractable approximation to the log-likelihood of the observation x, where x represents the input. The ELBO is formulated as follows:

$$\begin{aligned} \log(p(x)) \geq \ & \mathbb{E}_{q(F|x)} \log(p(x|F)) \\ & - \mathbb{KL}(q(F|x)||p(F)) \end{aligned} \tag{11}$$

where the reconstruction loss is defined as:

$$\mathcal{L}_R = -\mathbb{E}_{q(F|x)} \log(p(x|F)) \tag{12}$$

The ELBO's decomposition consists of two key terms: the reconstruction loss $\mathcal{L}_R$, computed using samples F drawn from the approximate posterior $q(F|x)$, and the KL divergence between this posterior and the prior $p(F)$. In this work, we replace the reconstruction loss with a task specific loss $\mathcal{L}_T$. Henderson and Fehr (2023) further divided the KL term into $\mathcal{L}_G$, corresponding to Gaussian distributions, and $\mathcal{L}_D$, corresponding to Dirichlet distributions.

$$\mathcal{L}_D + \mathcal{L}_G \approx \mathrm{D}_{\mathbb{KL}}(q(F|x) \,||\, p(F)) \tag{13}$$

This gives us the following loss terms for the KL divergence, where Γ is the gamma function and ψ is the digamma function:

$$\begin{aligned}\mathcal{L}_D = \log\Gamma(\alpha_0^q) - \log\Gamma(\alpha_0^{p'}) \\ + (\alpha_0^q - \alpha_0^{p'})\,(-\psi(\alpha_0^q) + \psi(\alpha_0^q)) \\ + \left(\log\Gamma(\alpha_0^{p'}) - \log\Gamma(\alpha_0^q)\right)\end{aligned} \tag{14}$$

$$\begin{aligned}\mathcal{L}_G = \frac{1}{2}\kappa_0 \sum_{i=1}^{n+1} \frac{\alpha_i^q}{\alpha_0^q} \sum_{h=1}^{d} \Bigg(\frac{(\mu_{ih}^q - \mu_h^p)^2}{(\sigma_h^p)^2} - 1 \\ + \frac{(\sigma_{ih}^q)^2}{(\sigma_h^p)^2} - \log\frac{(\sigma_{ih}^q)^2}{(\sigma_h^p)^2} \Bigg)\end{aligned} \tag{15}$$

Since we only draw a single sample per component, thus $\kappa_0 = n{+}1$. However, in practice we scale both $\mathcal{L}_G$ and $\mathcal{L}_D$ by the number of components $(n+1)$ such that the loss is invariant to sequence length. We introduce two hyperparameters to control the relative weight of the above three parts of the loss, which defines our VIB loss $\mathcal{L}$.

$$\mathcal{L} = \mathcal{L}_T \;+\; \lambda_D \mathcal{L}_D \;+\; \lambda_G \mathcal{L}_G \tag{16}$$

A.4 Including NVIB into pretrained models

Fehr and Henderson (2024) define an identity initialisation for NVIB such that the latent embeddings have negligible uncertainty and denoising attention is effectively equivalent to standard attention. This allows pretrained attention-based models to be reinterpreted as Nonparametric Variational models. By only changing the initialisation, away from the identity and towards an empirically estimated prior, an effective post-training regularisation is added. The authors found that this information-theoretic regularisation lead to improvements in OOD text generalisation in summarisation and translation without fine-tuning.

B Simplifying denoising attention

In this section, we provide the implementation details for *denoising multihead attention*. We define the set of Transformer latent embedding vectors as $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$ and set of pre-projected queries as $\boldsymbol{U'} \in \mathbb{R}^{m\times d}$. We assume the latent projection matrices are square such that $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{d\times d}$ and biases $\boldsymbol{b}^Q, \boldsymbol{b}^K, \boldsymbol{b}^V \in \mathbb{R}^d$ are used to linearly project to the queries, keys and values, respectively. We define the standard attention weights before the softmax as follows:

$$\boldsymbol{A} = \frac{1}{\sqrt{d}} \underbrace{(\boldsymbol{U'}\boldsymbol{W}^Q + \boldsymbol{b}^Q)}_{\boldsymbol{Q}} \underbrace{(\boldsymbol{Z}\boldsymbol{W}^K + \boldsymbol{b}^K)^\top}_{\boldsymbol{K}^\top} \tag{17}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Typically, for multihead attention the projected query $\boldsymbol{Q}$ and keys $\boldsymbol{K}$ are split into heads. In this definition, we split the linear projections by a divisible number of heads h such that $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{h\times d\times \frac{d}{h}}$ and biases $\boldsymbol{b}^Q, \boldsymbol{b}^K, \boldsymbol{b}^V \in \mathbb{R}^{h\times \frac{d}{h}}$, so that $\boldsymbol{Q} \in \mathbb{R}^{h\times m\times \frac{d}{h}}$ and $\boldsymbol{K} \in \mathbb{R}^{h\times n\times \frac{d}{h}}$. We can then specify multihead attention by defining a matrix of attention scores $\boldsymbol{A} \in \mathbb{R}^{h\times m\times n}$, for each head i:

$$\boldsymbol{A}_i = \frac{1}{\sqrt{d/h}}((\boldsymbol{Q}_i(\boldsymbol{W}_i^K)^\top \boldsymbol{Z}^\top + \boldsymbol{Q}_i(\boldsymbol{b}_i^K)^\top) \tag{18}$$

where the bias term $\boldsymbol{Q}_i(\boldsymbol{b}_i^K)^\top \in \mathbb{R}^m$ is added across all n keys, and thus is normalised out in the softmax below. The scaling term also considers the heads and is division by $\sqrt{d/h}$. For denoising attention, each head's query is projected into the space of the original set of vectors $\boldsymbol{Z}$, namely $\boldsymbol{U}_i{=}\boldsymbol{Q}_i(\boldsymbol{W}_i^K)^\top$, and so is still in $\mathbb{R}^{m\times d}$. Thus, each head can be viewed as doing denoising attention in the same way as single-head attention, with the only difference being that the variance of the theoretical query noise is now $\sqrt{d/h}\boldsymbol{I}$.

Training. Given these sampled weights and vectors, the training-time denoising attention function is the same as the standard attention function with two changes: (1) the keys come from the sampled vectors $\boldsymbol{Z} \in \mathbb{R}^{(n+1)\times d}$, which include a vector sampled from the prior component; and (2) each key has an attention bias $\boldsymbol{b} \in \mathbb{R}^{(n+1)}$ which is determined by its weight $\boldsymbol{\pi} \in \mathbb{R}^{(n+1)}$. Summing over heads i, the training-time denoising attention function DAttn(.) is defined as follows:

$$\begin{aligned}\text{DAttn}(.) = \sum_i \text{Softmax}(\boldsymbol{A}_i \\ \underbrace{+ \log(\boldsymbol{\pi}) - \frac{1}{2\sqrt{d/h}}\|\boldsymbol{Z}\|^2)}_{\boldsymbol{b}} \\ \times \underbrace{(\boldsymbol{Z}\boldsymbol{W}_i^V + \boldsymbol{b}_i^V)}_{\boldsymbol{V}_i}\end{aligned} \tag{19}$$

The biases $\boldsymbol{b}$ are defined by adding the log of the sampled weights $\log(\boldsymbol{\pi}) \in \mathbb{R}^{(n+1)}$ from the NVIB layer and subtracting the scaled squared-L2-norms of the sampled vectors $\frac{1}{2\sqrt{d/h}}\|\boldsymbol{Z}\|^2 \in \mathbb{R}^{(n+1)}$. For multihead attention we only need to reuse the same biases $\boldsymbol{b}$ for each head, just like we reuse the same vectors $\boldsymbol{Z}$ for each head.

Evaluation. During the evaluation, as for training, the NVIB layer outputs the isotropic Gaussian parameters $\boldsymbol{\mu} \in \mathbb{R}^{(n+1)\times d}, \boldsymbol{\sigma} \in \mathbb{R}^{(n+1)\times d}$ and Dirichlet parameters $\boldsymbol{\alpha} \in \mathbb{R}^{(n+1)}$. For evaluation the base distribution is used. The parameters are taken directly without sampling such that we use the expectation of the distribution. We can write the denoising attention scores $\boldsymbol{A} \in \mathbb{R}^{h\times m\times(n+1)}$, for each head i, as follows:

$$\boldsymbol{A}_i = \boldsymbol{Q}_i(\boldsymbol{W}_i^K)^\top(\frac{\boldsymbol{\mu}}{\sqrt{d/h}})^\top + \frac{1}{\sqrt{d/h}}\boldsymbol{Q}_i(\boldsymbol{b}_i^K)^\top \tag{20}$$

where the bias term $\boldsymbol{Q}_i(\boldsymbol{b}_i^K)^\top \in \mathbb{R}^m$ is added across all n keys, and thus is normalised out in the softmax below. For this attention score matrix $\boldsymbol{A}$, multihead evaluation denoising attention adds the same key biases $\boldsymbol{c} \in \mathbb{R}^{h\times(n+1)}$ across all m queries and h heads. For ease of notation we define $\alpha_0 = \sum_{j=1}^{d} \alpha_j$. Thus, the test-time denoising attention function DAttn(.) is defined as follows:

$$\begin{aligned}\text{DAttn}(.) = \sum_i \text{Softmax}(\boldsymbol{A}_i \\ \underbrace{+ \log(\frac{\boldsymbol{\alpha}}{\alpha_0}) - \frac{1}{2}\|\frac{\boldsymbol{\mu}}{\sqrt{d}}\|^2)}_{\boldsymbol{b}} \\ \times \underbrace{(\boldsymbol{\mu}\boldsymbol{W}_i^V + \boldsymbol{b}_i^V)}_{\boldsymbol{V}_i}\end{aligned} \tag{21}$$

This simplifies previous implementations of Henderson and Fehr (2023) and Fehr and Henderson (2024) by removing the additional variance term in the bias $\boldsymbol{b}$ and the interpolation between the query and value vectors. This makes the training and test time denoising attention functions more similar and reduces computation requirements.

C Attention maps

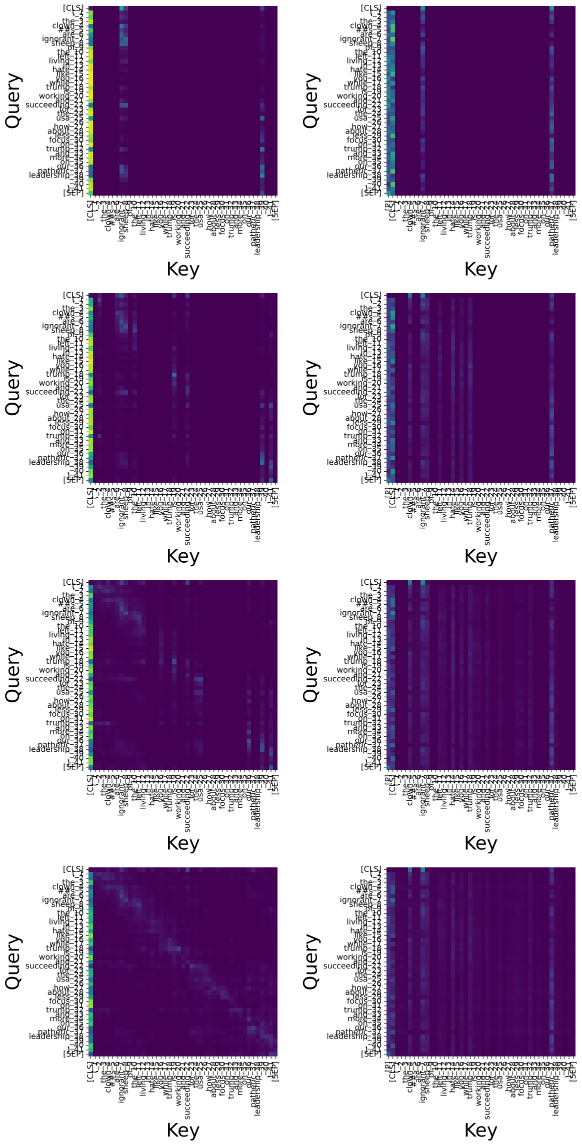

Figure 6: Attention plot for the best models on Civil-Comments. The plots show a single head of the last layer. Left: with dropout, Right: with NVIB. Top-Bottom: Proportion of keys retained [0.1, 0.25, 0.5, 1.0]. Sentence: ('the clowns are ignorant sheep of the left living in hate like you while trump is working and succeeding for the usa. how about less focus on trump and more on our pathetic leadership'.) NVIB highlights 'ignorant' and 'pathetic'.

Pseudocode: Attention and Denoising Attention during training (single-head). Left: Standard Attention. Right: Denoising Attention.

```
class Attention():
    def __init__(self, d):
        # Projections to Q, K, V [d,d]
        self.q = linear(d, d)
        self.k = linear(d, d)
        self.v = linear(d, d)

    def forward(self, u, z):
        # queries       u: [B, M, d]
        # keys / values z: [B, N, d]
        d = keys.shape(2)

        # Project to Q, K, V
        q = self.q(u)
        k = self.k(z) / sqrt(d)
        v = self.v(z)

        # Attention scores [B, M, N]
        attn = q @ k.transpose()

        # Attention probabilities [B, M, N]
        attn = softmax(attn)

        # Value projection [B, M, d]
        out = attn @ v

        return out
```

```
class DenoisingAttention():
    def __init__(self, d):
        # Projections to Q, K, V [d,d]
        self.q = linear(d, d)
        self.k = linear(d, d)
        self.v = linear(d, d)

    def forward(self, u, z, pi):
        # queries       u: [B, M, d]
        # keys / values z: [B, N+1, d]
        d = keys.shape(2)

        # Project to Q, K, V
        q = self.q(u)
        k = self.k(z) / sqrt(d)
        v = self.v(z)

        # NVIB bias [B, 1, N+1]
        b = log(pi)
            - 1/(2*sqrt(d))*l2norm(z)**2

        # Attention scores [B, M, N+1]
        attn = q @ k.transpose() + b

        # Attention probabilities [B, M, N+1]
        attn = softmax(attn)

        # Value projection [B, M, d]
        out = attn @ v

        return out
```

Pseudocode: Denoising Attention during evaluation (single-head). Left: Previous implementation including extra bias term and query value interpolation. Right: Current simplified implementation.

```
class DenoisingAttention():
    def __init__(self, d):
        # Projections to Q, K, V [d,d]
        self.q = linear(d, d)
        self.k = linear(d, d)
        self.v = linear(d, d)

    def forward(self, u, mu, sigma2, alpha):
        # queries         u: [B, M, d]
        # keys / values mu: [B, N+1, d]
        d = keys.shape(2)

        # Project to Q, K, V
        q = self.q(u)
        k = self.k(mu / (sqrt(d)+sigma2))
        # v is used in interpolation

        # NVIB bias [B, 1, N+1]
        b = log(alpha / sum(alpha))
            - 1/(2*(sqrt(d)+sigma2))*l2norm(mu)**2
            - sum(log(sqrt(sqrt(d)+sigma2)))

        # Attention scores [B, M, N+1]
        attn = q @ k.transpose() + b

        # Attention probabilities [B, M, N+1]
        attn = softmax(attn)

        # Query projection to key-space [B, M, d]
        u_k = self.k(q)

        # Value interpolation [B, M, d]
        out = (attn @ (sigma2/(sqrt(d)+sigma2)))*u_k
             + attn @ ((sqrt(d)/(sqrt(d)+sigma2)))*mu
        out = self.v(out)


        return out
```

```
class DenoisingAttention():
    def __init__(self, d):
        # Projections to Q, K, V [d,d]
        self.q = linear(d, d)
        self.k = linear(d, d)
        self.v = linear(d, d)

    def forward(self, u, mu, alpha):
        # queries       u: [B, M, d]
        # keys/values mu: [B, N+1, d]
        d = keys.shape(2)

        # Project to Q, K, V
        q = self.q(u)
        k = self.k(mu) / sqrt(d)
        v = self.v(mu)

        # NVIB bias [B, 1, N+1]
        b = log(alpha/sum(alpha))
            - 1/(2*sqrt(d))*l2norm(mu)**2

        # Attention scores [B, M, N+1]
        attn = q @ k.transpose() + b

        # Attention probabilities [B, M, N+1]
        attn = softmax(attn)

        # Value projection [B, M, d]
        out = attn @ v

        return out
```

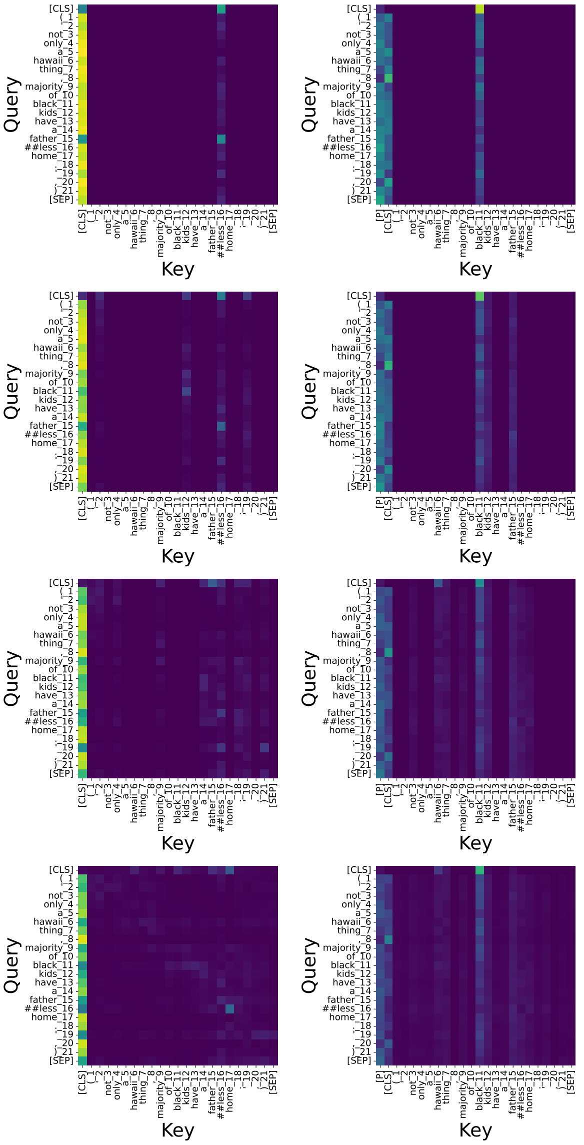

Figure 7: Attention plot for the best models on Civil-Comments. The plots show a single head of the last layer. Left: with dropout, Right: with NVIB. Top-Bottom: Proportion of keys retained [0.1, 0.25, 0.5, 1.0]. Sentence: ('not onlu a hawaii thing, majority of black kids have a fatherless home.-') NVIB highlights 'black', 'kids' and 'father'.

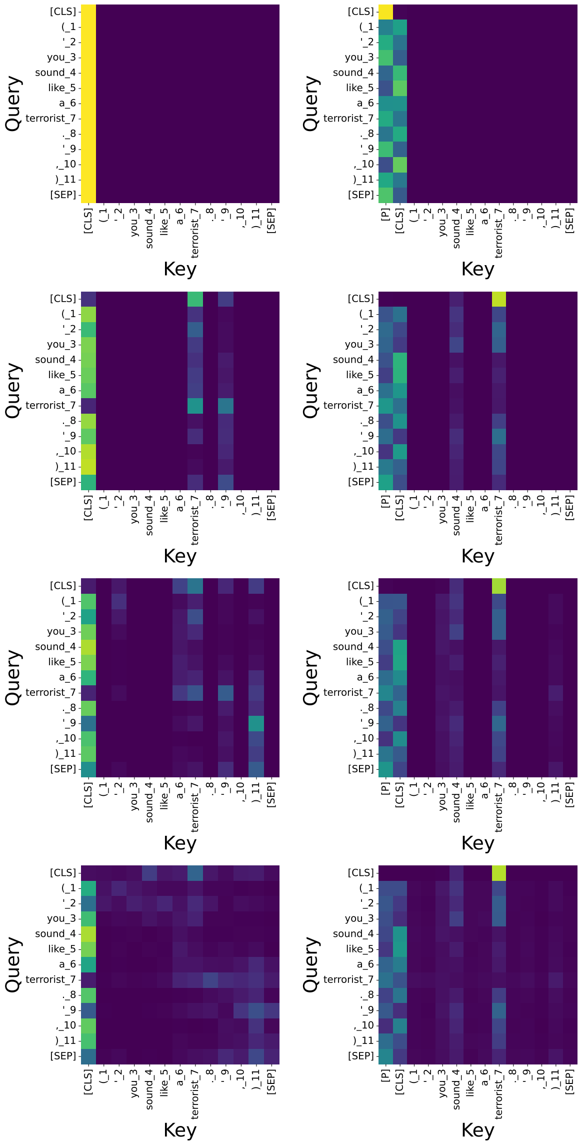

Figure 8: Attention plot for the best OOD models on CC with increasing sparsity. The plots show a single head of the last layer. Left: with dropout, Right: with NVIB. Top-Bottom: Proportion of keys retained [0.1, 0.25, 0.5, 1.0]. Sentence: ('you sound like a terrorist.'). NVIB highlights 'sound' and 'terrorist'.