Novelty Assessment Report

Paper: GRADIEND: Feature Learning within Neural Networks Exemplified through Biases
PDF URL: https://openreview.net/pdf?id=1vBNAnAgCD
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

AI systems frequently exhibit and amplify social biases, leading to harmful consequences in critical areas. This study introduces a novel encoder-decoder approach that leverages model gradients to learn a feature neuron encoding societal bias information such as gender, race, and religion. We show that our method can not only identify which weights of a model need to be changed to modify a feature, but even demonstrate that this can be used to rewrite models to debias them while maintaining other capabilities. We demonstrate the effectiveness of our approach across various model architectures and highlight its potential for broader applications.

Core Task Landscape

This paper addresses: **Learning and Modifying Societal Biases in Neural Network Models Using Gradient-Based Feature Encoding**

A total of **15 papers** were analyzed and organized into a taxonomy with **10 categories**.

Taxonomy Overview

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

- **Gradient-Based Bias Detection and Feature Attribution**
- **Gradient-Based Bias Mitigation Through Representation Modification**
- **Adversarial and Activation-Based Bias Mitigation**
- **Data-Level Bias Mitigation Through Augmentation and Intervention**
- **Fairness-Constrained Optimization and Unified Learning Frameworks**
- **Bias Testing and Evaluation Frameworks**
- **Multimodal and Domain-Invariant Bias Considerations**

Complete Taxonomy Tree

- Learning and Modifying Societal Biases in Neural Network Models Using Gradient-Based Feature Encoding Survey Taxonomy
- Gradient-Based Bias Detection and Feature Attribution
  - Gradient Attention and Saliency-Based Bias Detection (3 papers)
  - [4] Gradient attention balance network: Mitigating face recognition racial bias via gradient attention (Linzhi Huang, 2023) View paper
  - [12] Neural Networks Bias Mitigation Through Fuzzy Logic and Saliency Maps (Sahar Shah, 2025) View paper
  - [14] Detecting bias in image classification using model explanations (Schrasing Tong, 2020) View paper
  - Gradient-Based Feature Importance and Bias Attribution (2 papers)
  - [9] Bias also matters: Bias attribution for deep neural network explanation (Shengjie Wang, 2019) View paper
  - [13] PMPO: A Self-Optimizing Framework for Creating High-Fidelity Measurement Tools for Social Bias in Large Language Models (Z Wang, 2023) View paper
- Gradient-Based Bias Mitigation Through Representation Modification
  - Gradient-Based Projection and Feature Removal ★ (2 papers)
  - [0] GRADIEND: Feature Learning within Neural Networks Exemplified through Biases (Anon et al., 2026) View paper
  - [6] Shielded representations: Protecting sensitive attributes through iterative gradient-based projection (Iskander, 2023) View paper
  - Gradient Penalization in Latent Space (1 papers)
  - [5] From Hope to Safety: Unlearning Biases of Deep Models via Gradient Penalization in Latent Space (Maximilian Dreyer, 2023) View paper
- Adversarial and Activation-Based Bias Mitigation (1 papers)
  - [3] Gradient Based Activations for Accurate Bias-Free Learning (Kurmi, 2022) View paper
- Data-Level Bias Mitigation Through Augmentation and Intervention
  - Causal Intervention and Corpus Augmentation (1 papers)
  - [2] Gender bias in neural natural language processing (Kaiji Lu, 2020) View paper
  - Counterfactual Interpolation Augmentation (1 papers)
  - [10] Counterfactual Interpolation Augmentation (CIA): A Unified Approach to Enhance Fairness and Explainability of DNN. (Yao Qiang, 2022) View paper
- Fairness-Constrained Optimization and Unified Learning Frameworks (1 papers)
  - [11] An Analytical Framework for Bias Mitigation in Credit Scoring Systems through Fairness-Constrained Neural Optimization (Nagaraj, 2025) View paper

- Bias Testing and Evaluation Frameworks (2 papers)
  - [1] Explainable evaluation framework for facial expression recognition in web-based learning environments (Amira Mouakher, 2025) View paper
  - [8] Black-box fairness testing with shadow models (Weipeng Jiang, 2023) View paper
- Multimodal and Domain-Invariant Bias Considerations (2 papers)
  - [7] Road Ahead for Multiâ□□modal Intelligent Sensing in the Deep Learning Era (Ahmed. Zoha, 2024) View paper
  - [15] Towards trustworthy data analytics: Algorithmic tools for interpretability and fairness (Cong, 2022) View paper

Narrative

Core task: learning and modifying societal biases in neural network models using gradient-based feature encoding. The field has organized itself around several complementary strategies. One major branch focuses on gradient-based bias detection and feature attribution, where researchers trace how sensitive model predictions are to protected attributes. A second branch emphasizes gradient-based bias mitigation through representation modification, employing techniques such as projection and feature removal to reshape internal embeddings. Parallel to these are adversarial and activation-based mitigation methods, data-level interventions through augmentation, fairness-constrained optimization frameworks that integrate bias penalties directly into training objectives, and dedicated bias testing and evaluation suites. Finally, multimodal and domain-invariant considerations address bias across vision-language systems and cross-domain generalization. Together, these branches reflect a progression from diagnosing bias to actively correcting it at multiple levels of the learning pipeline.

Within the representation-modification branch, a particularly active line of work explores how to surgically remove or shield biased features from learned embeddings. GRADIEND[0] exemplifies this approach by using gradient information to identify and encode features that carry societal bias, then modifying representations to reduce reliance on those features. Closely related, Shielded Representations[6] also targets feature-level interventions to protect against unwanted correlations, while Unlearning Biases Gradient[5] investigates gradient-driven unlearning strategies to erase biased associations post-training. These methods share a common emphasis on leveraging backpropagation signals to pinpoint and neutralize bias, yet they differ in whether they act during training, fine-tuning, or as a post-hoc correction. Open questions remain about the trade-offs between utility preservation and bias reduction, and how these gradient-based techniques scale across diverse datasets and model architectures.

Related Works in Same Category

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

1. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection

Authors: Iskander, Shadi, Radinsky, Kira, Belinkov, et al. (6 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Natural language processing models tend to learn and encode social biases present in the data. One popular approach for addressing such biases is to eliminate encoded information from the model's representations. However, current methods are restricted to removing only linearly encoded information.In this work, we propose Iterative Gradient-Based Projection (IGBP), a novel method for removing non-linear encoded concepts from neural representations. Our method consists of iteratively training neu...

Relationship Analysis

Both papers belong to the Gradient-Based Projection and Feature Removal category, using iterative gradient-based optimization to remove bias-encoded features from neural representations. They overlap in their core approach of leveraging gradients to identify and project away sensitive attribute information (gender, race) from model representations. The key difference is that the original paper (GRADIEND) learns an encoder-decoder architecture to modify model weights directly for rewriting models, while the candidate paper (IGBP) uses a projective loss with trained probe classifiers to project representations onto hypersurfaces at inference time without modifying the underlying model parameters.

Contributions Analysis

Overall novelty summary. The paper proposes a novel encoder-decoder architecture (GRADIEND) that uses model gradients to learn feature neuron encodings of societal biases, then modifies model weights to debias transformers while preserving other capabilities. According to the taxonomy, this work resides in the 'Gradient-Based Projection and Feature Removal' leaf under 'Gradient-Based Bias Mitigation Through Representation Modification'. This leaf contains only two papers total (including the original), indicating a relatively sparse research direction within the broader field of gradient-based bias mitigation, which itself is one of several major branches addressing societal bias in neural networks.

The taxonomy reveals that the paper sits within a representation-modification branch, distinct from parallel approaches such as adversarial methods, data-level augmentation, fairness-constrained optimization, and bias detection frameworks. Neighboring leaves include 'Gradient Penalization in Latent Space' (which penalizes sensitivity rather than removing features) and 'Gradient Attention and Saliency-Based Bias Detection' (focused on detection rather than mitigation). The scope notes clarify that this leaf specifically targets iterative projection or removal of bias-encoded features using gradient-based optimization, excluding adversarial training or data-level interventions that appear in separate taxonomy branches.

Among 29 candidates examined across three contributions, none were found to clearly refute the paper's claims. Contribution A (encoder-decoder architecture) examined 10 candidates with 0 refutable; Contribution B (weight modification method) examined 10 candidates with 0 refutable; Contribution C (orthogonal class pairs) examined 9 candidates with 0 refutable. The sibling paper in the same taxonomy leaf (Shielded Representations) addresses feature-level interventions but appears to differ in approach. Given the limited search scope (top-K semantic search plus citation expansion, not exhaustive), these statistics suggest the specific combination of encoder-decoder architecture and gradient-based weight modification for debiasing may represent a novel technical approach within this sparse research direction.

Based on the limited literature search covering 29 candidates, the work appears to occupy a relatively unexplored position within gradient-based representation modification for bias mitigation. The sparse taxonomy leaf (only 2 papers) and absence of clearly refutable prior work among examined candidates suggest potential novelty, though the analysis cannot rule out relevant work outside the top-K semantic matches or in adjacent research communities not captured by the taxonomy structure.

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

Contribution 1: GRADIEND encoder-decoder architecture for feature learning

Description: The authors introduce GRADIEND, a novel encoder-decoder architecture that learns a single scalar feature neuron from model gradients. The encoder compresses gradients into a feature representation, while the decoder learns which model weights need modification to change that feature.

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. Attention LinkNet-152: a novel encoder-decoder based deep learning network for automated spine segmentation

URL: View paper

Brief Assessment

Attention LinkNet[39] focuses on spine segmentation from CT images using an encoder-decoder architecture for image processing, not on learning features from model gradients in language models as GRADIEND does.

---

2. On the uses of large language models to design end-to-end learning semantic communication

URL: View paper

Brief Assessment

LLM Semantic Communication[43] focuses on semantic communication systems using LLMs for text transmission over channels, not on learning features from model gradients or bias detection/modification in neural networks.

---

3. Architecting contextual gradient synthesis for knowledge representation in large language models

URL: View paper

Brief Assessment

Contextual Gradient Synthesis[37] mentions an encoder-decoder structure with gradient components, but the provided context is too fragmentary to establish whether this represents prior work that refutes the novelty of GRADIEND's specific approach to learning scalar feature neurons from model gradients.

---

4. Dm-codec: Distilling multimodal representations for speech tokenization

URL: View paper

Brief Assessment

DM-Codec[36] focuses on speech tokenization using encoder-decoder architectures for audio processing, not on learning features from model gradients in language models as GRADIEND does.

---

5. Fractal gradient reconstitution in large language models: A framework for internal representation coherence through recursive tensor reassembly

URL: View paper

Brief Assessment

Fractal Gradient Reconstitution[38] discusses gradient information flow through encoder-decoder bridges in bidirectional contexts and numerical optimization within LLM training, but does not present an encoder-decoder architecture that learns scalar feature neurons from model gradients for the purpose of identifying which weights need modification to change specific features like bias.

---

6. PGC-Net: A Novel Encoder-Decoder Network With Path Gradient Flow Control for Cell Counting

URL: View paper

Brief Assessment

PGC-Net[42] focuses on cell counting using encoder-decoder networks with gradient flow control for density map regression, not on learning features from model gradients in language models as GRADIEND does.

---

7. Mobile edge intelligence for large language models: A contemporary survey

URL: View paper

Brief Assessment

Mobile Edge Intelligence[35] focuses on deploying large language models at the network edge for mobile applications, not on encoder-decoder architectures for learning features from model gradients in language models.

---

8. Enhanced encoderâ□□decoder architecture for accurate monocular depth estimation

URL: View paper

Brief Assessment

Enhanced Encoder Decoder[40] focuses on monocular depth estimation using an encoder-decoder architecture for computer vision tasks, not on learning features from model gradients in language models as GRADIEND does.

---

9. Contextual gradient recomposition for sequential coherence preservation in large language model token generation

URL: View paper

Brief Assessment

Contextual Gradient Recomposition[41] focuses on gradient reweighting during inference for token generation coherence, not on learning feature neurons from gradients via encoder-decoder architectures as in the original paper.

---

10. Transformers Get Stable: An End-to-End Signal Propagation Theory for Language Models

URL: View paper

Brief Assessment

Transformers Get Stable[44] focuses on signal propagation theory for training stability in deep transformers, not on encoder-decoder architectures for learning features from model gradients.

Contribution 2: Method for identifying and modifying model weights to debias transformers

Description: The authors demonstrate that their approach can identify specific model weights associated with societal biases (gender, race, religion) and modify these weights to reduce bias while preserving language modeling performance, achieving state-of-the-art results for gender debiasing when combined with INLP.

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. Post-hoc Spurious Correlation Neutralization with Single-Weight Fictitious Class Unlearning

URL: View paper

Brief Assessment

Spurious Correlation Neutralization[25] focuses on post-hoc single-weight editing for spurious correlation mitigation in image classifiers, not on identifying and modifying transformer weights for debiasing language models. The technical approaches and application domains differ fundamentally.

2. Chatgpt based data augmentation for improved parameter-efficient debiasing of llms

URL: View paper

Brief Assessment

ChatGPT Data Augmentation[20] focuses on generating synthetic training data for debiasing using ChatGPT and adapter tuning, rather than identifying and modifying specific model weights associated with biases. The approaches are methodologically distinct.

3. Identifying and adapting transformer-components responsible for gender bias in an English language model

URL: View paper

Brief Assessment

Transformer Gender Bias[23] focuses on identifying attention heads responsible for gender bias and fine-tuning them, rather than directly modifying weights across all model parameters to debias while maintaining capabilities as described in the original contribution.

4. Debiasing attention mechanism in transformer without demographics

URL: View paper

Brief Assessment

Debiasing Transformer Demographics[18] addresses debiasing transformers without demographic information. The original paper's approach uses gradient-based feature learning to identify and modify weights, while the candidate's focus and methodology are unclear from the provided excerpt.

5. Id-xcb: Data-independent debiasing for fair and accurate transformer-based cyberbullying detection

URL: View paper

Brief Assessment

ID-XCB[21] focuses on cyberbullying detection and uses adversarial training with fairness constraints, not on identifying and modifying specific transformer weights associated with societal biases as described in the original contribution.

6. Editing models with task arithmetic

URL: View paper

Brief Assessment

Task Arithmetic[17] focuses on editing models through task vectors for multi-task learning and task transfer, not specifically on identifying and modifying weights for debiasing transformers while preserving language modeling performance.

7. Fredformer: Frequency debiased transformer for time series forecasting

URL: View paper

Brief Assessment

Fredformer[16] addresses frequency bias in time series forecasting transformers, not societal bias debiasing. The paper focuses on balancing frequency feature learning rather than identifying and modifying weights to reduce gender, race, or religious biases while preserving language modeling capabilities.

8. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models

URL: View paper

Brief Assessment

Parameter-Efficient Debiasing Analysis[24] focuses on parameter-efficient tuning methods (prefix tuning, prompt tuning, adapter tuning) combined with counterfactual data augmentation, not on identifying and modifying specific weights associated with biases as described in the original contribution.

9. Debiasing CLIP: Interpreting and Correcting Bias in Attention Heads

URL: View paper

Brief Assessment

Debiasing CLIP[19] focuses on identifying and modifying attention heads in vision transformers (CLIP models) for debiasing, not general transformer weight modification for language models. The technical approach differs fundamentally from the original paper's gradient-based encoder-decoder method for language model debiasing.

10. Curriculum Debiasing: Toward Robust Parameter-Efficient Fine-Tuning Against Dataset Biases

URL: View paper

Brief Assessment

Curriculum Debiasing[22] focuses on curriculum learning strategies for PEFT methods to address spurious correlations, not on identifying and modifying specific transformer weights to debias models as described in the original contribution.

Contribution 3: Gradient-based feature learning using orthogonal class pairs

Description: The authors propose using gradient differences between factual and orthogonal (counterfactual) token prediction tasks to learn targeted features with desired interpretations. This contrasts with unsupervised methods like Sparse AutoEncoders that discover features without guaranteeing specific semantic meanings.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Tigtec: Token importance guided text counterfactuals

URL: View paper

Brief Assessment

TIGTEC[29] uses gradient-based token importance for counterfactual text generation, not for learning targeted feature neurons with desired interpretations. The methods serve different purposes: TIGTEC targets tokens for text editing, while the original paper learns feature representations within neural networks.

2. Gradients of counterfactuals

URL: View paper

Brief Assessment

Gradients of Counterfactuals[30] focuses on using gradients of scaled input images (counterfactuals) to understand feature importance in already-trained networks for interpretability. The original paper proposes learning feature neurons during training using gradient differences between factual and orthogonal token predictions, which is a fundamentally different approach for targeted feature learning rather than post-hoc interpretation.

---

3. Enhancing textual counterfactual explanation intelligibility through Counterfactual Feature Importance

URL: View paper

Brief Assessment

Counterfactual Feature Importance[27] focuses on explaining counterfactual text modifications using gradient-based importance attribution, not on learning targeted feature representations within neural networks through orthogonal token prediction tasks.

---

4. From Prediction to Action: Counterfactual Explanations and Ensemble Learning for Explainable Maternal Health Risk Modelling

URL: View paper

Brief Assessment

Maternal Health Risk[33] focuses on maternal health risk prediction using ensemble learning and counterfactual explanations (DICE). It does not address gradient-based feature learning or orthogonal class pairs for interpretable representations in neural networks.

---

5. Tabcf: Counterfactual explanations for tabular data using a transformer-based vae

URL: View paper

Brief Assessment

TabCF[28] focuses on counterfactual explanations for tabular data using a transformer-based VAE, not on gradient-based feature learning from orthogonal token prediction tasks for interpretable representations.

---

6. Interpretable instance disease prediction based on causal feature selection and effect analysis

URL: View paper

Brief Assessment

Causal Feature Selection[31] focuses on counterfactual-based feature selection for disease prediction in healthcare, not on learning interpretable features through gradient differences between factual and orthogonal token prediction tasks in language models.

---

7. GradCFA: A Hybrid Gradient-Based Counterfactual and Feature Attribution Explanation Algorithm for Local Interpretation of Neural Networks

URL: View paper

Brief Assessment

GradCFA[32] focuses on counterfactual explanations and feature attribution for model interpretability in classification tasks, not on learning feature representations within neural networks using gradient differences between factual and counterfactual token predictions.

---

8. Causal Discovery and Inference through Next-Token Prediction

URL: View paper

Brief Assessment

Causal Discovery Prediction[34] focuses on causal reasoning and counterfactual inference in transformers trained for next-token prediction, not on gradient-based feature learning for interpretable representations or bias mitigation.

---

9. Auditing Black-Box AI Systems Using Counterfactual Explanations

URL: View paper

Brief Assessment

Auditing Black-Box AI[26] focuses on using counterfactual explanations and gradient-based methods for auditing and explaining AI model decisions, not for learning targeted feature representations within neural networks as described in the original contribution.

Appendix: Text Similarity Detection

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

References

- [0] GRADIEND: Feature Learning within Neural Networks Exemplified through Biases View paper
- [1] Explainable evaluation framework for facial expression recognition in web-based learning environments View paper
- [2] Gender bias in neural natural language processing View paper
- [3] Gradient Based Activations for Accurate Bias-Free Learning View paper
- [4] Gradient attention balance network: Mitigating face recognition racial bias via gradient attention View paper
- [5] From Hope to Safety: Unlearning Biases of Deep Models via Gradient Penalization in Latent Space View paper
- [6] Shielded representations: Protecting sensitive attributes through iterative gradient-based projection View paper
- [7] Road Ahead for Multiâ□□modal Intelligent Sensing in the Deep Learning Era View paper
- [8] Black-box fairness testing with shadow models View paper
- [9] Bias also matters: Bias attribution for deep neural network explanation View paper
- [10] Counterfactual Interpolation Augmentation (CIA): A Unified Approach to Enhance Fairness and Explainability of DNN. View paper
- [11] An Analytical Framework for Bias Mitigation in Credit Scoring Systems through Fairness-Constrained Neural Optimization View paper
- [12] Neural Networks Bias Mitigation Through Fuzzy Logic and Saliency Maps View paper
- [13] PMPO: A Self-Optimizing Framework for Creating High-Fidelity Measurement Tools for Social Bias in Large Language Models View paper
- [14] Detecting bias in image classification using model explanations View paper
- [15] Towards trustworthy data analytics: Algorithmic tools for interpretability and fairness View paper
- [16] Fredformer: Frequency debiased transformer for time series forecasting View paper

- [17] Editing models with task arithmetic View paper
- [18] Debiasing attention mechanism in transformer without demographics View paper
- [19] Debiasing CLIP: Interpreting and Correcting Bias in Attention Heads View paper
- [20] Chatgpt based data augmentation for improved parameter-efficient debiasing of llms View paper
- [21] Id-xcb: Data-independent debiasing for fair and accurate transformer-based cyberbullying detection View paper
- [22] Curriculum Debiasing: Toward Robust Parameter-Efficient Fine-Tuning Against Dataset Biases View paper
- [23] Identifying and adapting transformer-components responsible for gender bias in an English language model View paper
- [24] An empirical analysis of parameter-efficient methods for debiasing pre-trained language models View paper
- [25] Post-hoc Spurious Correlation Neutralization with Single-Weight Fictitious Class Unlearning View paper
- [26] Auditing Black-Box AI Systems Using Counterfactual Explanations View paper
- [27] Enhancing textual counterfactual explanation intelligibility through Counterfactual Feature Importance View paper
- [28] Tabcf: Counterfactual explanations for tabular data using a transformer-based vae View paper
- [29] Tigtec: Token importance guided text counterfactuals View paper
- [30] Gradients of counterfactuals View paper
- [31] Interpretable instance disease prediction based on causal feature selection and effect analysis View paper
- [32] GradCFA: A Hybrid Gradient-Based Counterfactual and Feature Attribution Explanation Algorithm for Local Interpretation of Neural Networks View paper
- [33] From Prediction to Action: Counterfactual Explanations and Ensemble Learning for Explainable Maternal Health Risk Modelling View paper
- [34] Causal Discovery and Inference through Next-Token Prediction View paper
- [35] Mobile edge intelligence for large language models: A contemporary survey View paper
- [36] Dm-codec: Distilling multimodal representations for speech tokenization View paper
- [37] Architecting contextual gradient synthesis for knowledge representation in large language models View paper
- [38] Fractal gradient reconstitution in large language models: A framework for internal representation coherence through recursive tensor reassembly View paper
- [39] Attention LinkNet-152: a novel encoder-decoder based deep learning network for automated spine segmentation View paper
- [40] Enhanced encoderâ□□decoder architecture for accurate monocular depth estimation View paper
- [41] Contextual gradient recomposition for sequential coherence preservation in large language model token generation View paper
- [42] PGC-Net: A Novel Encoder-Decoder Network With Path Gradient Flow Control for Cell Counting View paper
- [43] On the uses of large language models to design end-to-end learning semantic communication View paper
- [44] Transformers Get Stable: An End-to-End Signal Propagation Theory for Language Models View paper