Actionable Interpretability via Causal Hypergraphs:
Unravelling Batch Size Effects in Deep Learning

Zhongtian Sun[1] Anoushka Harit[2] Pietro Liò[2]

[1]University of Kent, UK [2]University of Cambridge, UK

Motivation

Why does batch size affect generalisation?

Although batch size is widely known to affect convergence and generalisation, especially in graph/text models, existing explanations remain heuristic and non-interventionist.

We address:

- Lack of causal understanding of batch-size effects.
- No modelling of higher-order (joint) training interactions.
- No bridge between interpretability and training-time control.

Our solution: HGCNet, a causal hypergraph framework that models batch dynamics structurally, enabling do-calculus–based training insights.

HGCNet: Causal Hypergraph Framework

We treat batch size B as a root intervention acting via:

$$B \to N \to S \to C \to G$$

Nodes:

- N: Gradient noise
- S: Sharpness (Hessian)
- C: Complexity (e.g. norm, margin)
- G: Generalisation (Test Acc.)

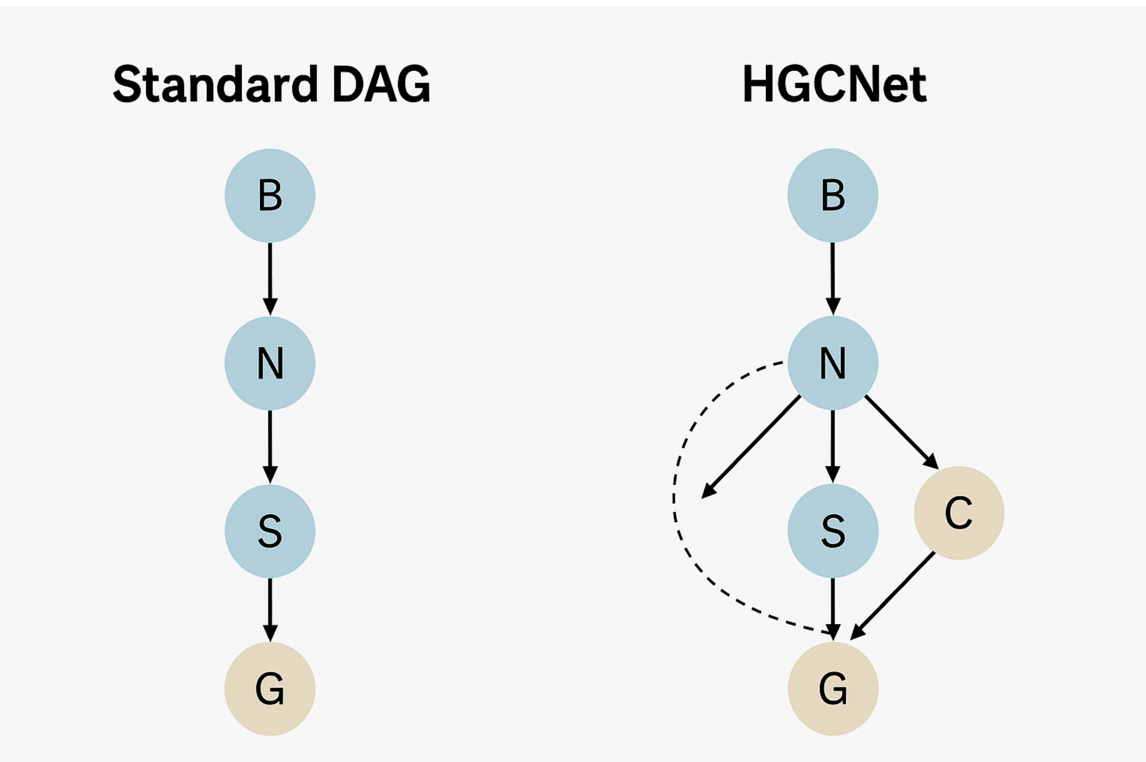

Figure 1. Causal Hypergraph Structure in HGCNet

Hyperedge: {N, S} → C enables joint mediation.
Implication: Enables do-calculus estimation and training-time policy derivation via ATE curves.

HGCNet Algorithm and Estimation

Input: Dataset D, batch sizes $B \in \{16, 32, \ldots, 512\}$

Steps:

1. Estimate gradient variance $N(B)$
2. Compute Hessian-based sharpness S
3. Estimate complexity $C = f(N, S)$
4. Measure generalisation G
5. Fit structural equations, apply:
$$P(G \mid do(B = b)) = \sum_{N,S,C} P(G|C)\, P(C|N,S)\, P(S|N)\, P(N|B = b)$$
6. Derive ATE curves and counterfactual predictions.

HGCNet Construction Details

We model the training process using a directed hypergraph: nodes are stochastic variables (e.g., gradient noise, sharpness), and hyperedges capture joint causal interactions. Key design:

- Edges like {N, S} → C encode higher-order effects (e.g., joint influence of noise and sharpness on complexity).
- Conditional independence relations enable efficient ATE and do-calculus inference.
- Training-time policy can be derived by counterfactuals: $do(B = b')$ reveals expected test performance.

This approach allows edge removal as a structured ablation tool, identifying dominant mediators.

Theoretical Insights

We treat batch size B as a *policy variable*, inducing downstream effects via gradient noise and sharpness:

$$B \to N \to S \to C \to G$$

We leverage:

- *Causal mediation analysis* to identify how training-time variables influence generalisation.
- *Do-calculus* for counterfactual estimation of policies like batch size intervention.
- ATE (Average Treatment Effect) curves to formalise generalisation tradeoffs.

Empirical Setup & Main Results

Domains & Datasets:

- **Graphs:** Cora, CiteSeer
- **Text:** PubMed, Amazon Reviews

Models: GCN, GAT, PI-GNN, BERT, RoBERTa, **HGCNet**

Measured Quantities:

- **Gradient Noise** (variance of updates)
- **Hessian Sharpness** (spectral norm)
- **Model Complexity** (norm, margin)
- **Generalisation Accuracy / Precision**

Main Results (ATE Estimate):

Dataset	B=16	B=512	Gain
Cora	83.9%	80.5%	**+3.4%**
CiteSeer	79.1%	76.0%	**+3.1%**
PubMed	88.2%	85.1%	**+3.1%**
Amazon	92.4%	89.0%	**+3.4%**

Smaller batches causally improve generalisation.

Causal Ablation (Edge Removal):

Ablation	Drop in Generalisation (G)
Remove Noise Node (N)	−3.4%
Remove Sharpness Node (S)	−2.8%
Remove {N,S} → C Edge	−2.0%
SAM-only Control	−1.5%

Gradient noise is the dominant causal mediator.

Broader Impact

Scientific Impact:

- Introduces the first causally grounded model for analysing batch size effects in training dynamics.
- Provides a foundation for causal benchmarking of generalisation–efficiency tradeoffs.

Practical Relevance:

- Enables interpretable, theory-driven batch size selection policies.
- Reduces reliance on heuristic tuning, especially in resource-constrained or safety-critical domains.

Future Applications:

- Can inform policy decisions in curriculum learning, fairness-aware training, and robust model deployment.
- Lays groundwork for causal training-time interventions in domains such as reinforcement learning and automated ML.

References

Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M. and Tang, P.T.P., 2016. On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836.
Pearl, J., 2009. Causality. Cambridge university press.
Peters, J., Janzing, D. and Schölkopf, B., 2017. Elements of causal inference: foundations and learning algorithms (p. 288). The MIT Press.

Quick Access

Paper PDF

LinkedIn