Sum-of-Parts: Self-Attributing Neural Networks with End-to-End Learning of Feature Groups

Weiqiu You, Helen Qu, Marco Gatti, Bhuvnesh Jain, Eric Wong
University of Pennsylvania

ICML International Conference On Machine Learning

Per-feature SANNs fail fundamentally

Prediction of a self-attributing neural network f decomposes to predictions hi for groups G_i, weighted by coefficients θ_i.

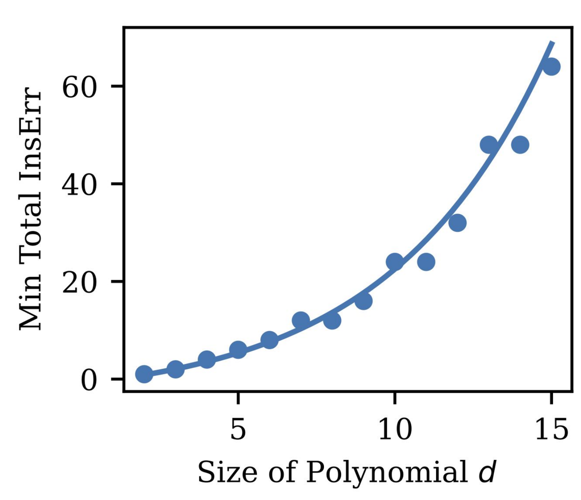
$$f(x) = \sum_{i=1}^{m} \theta(x)_i h(x_{G_i})$$

Theorem: Per-feature SANNs (i.e. |Gi| = 1) inherently cannot avoid exponential insertion and deletion errors. ☹️

Exponential Insertion Error ☹️

$$\text{InsErr}(G, \alpha, S) = \left| f^*(x_S) - f^*(0_d) - \sum_{G_i \subseteq S} \alpha_i \right|$$

$$\text{where } (x_S)_j = \begin{cases} x_j & \text{if } j \in S \\ 0 & \text{otherwise} \end{cases}$$



The total insertion error over all possible insertions is $\sum_{S \subseteq [d]} \text{InsErr}(G, \alpha, S)$.

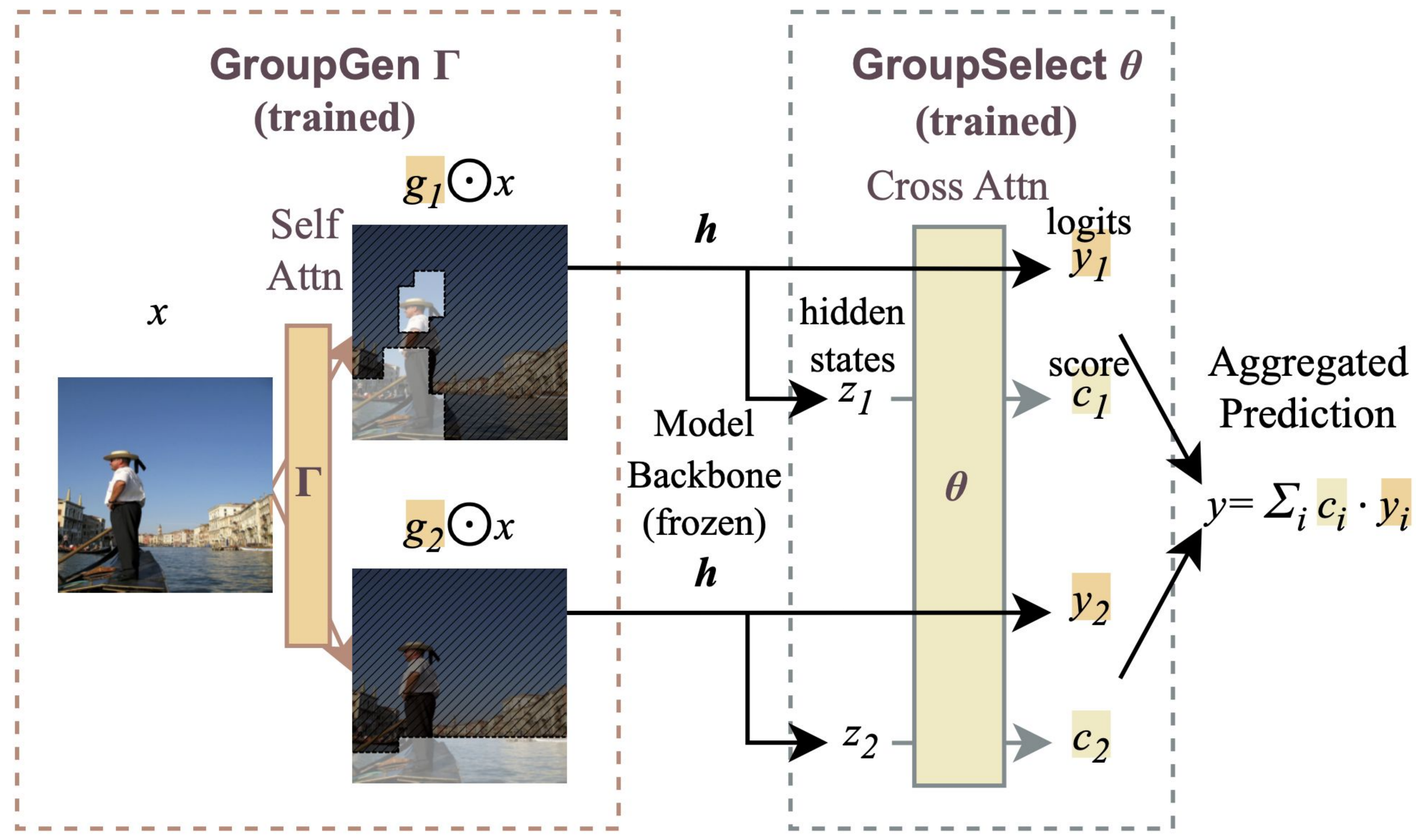
A similar problem arises in deletion. Most existing SANNs are per-feature, thus theoretically limited.

Group-based SANNs overcome the limits

To break the performance-accuracy trade-off, we must move from per-feature SANN to per-group SANN.

Theorem: Group-based SANN (i.e. |Gi| > 1) can achieve zero ins/del error for m-term polynomials with only m groups. 😍

We propose Sum-of-Parts, a *group-based* SANN that leverages the theory to get SOTA performance



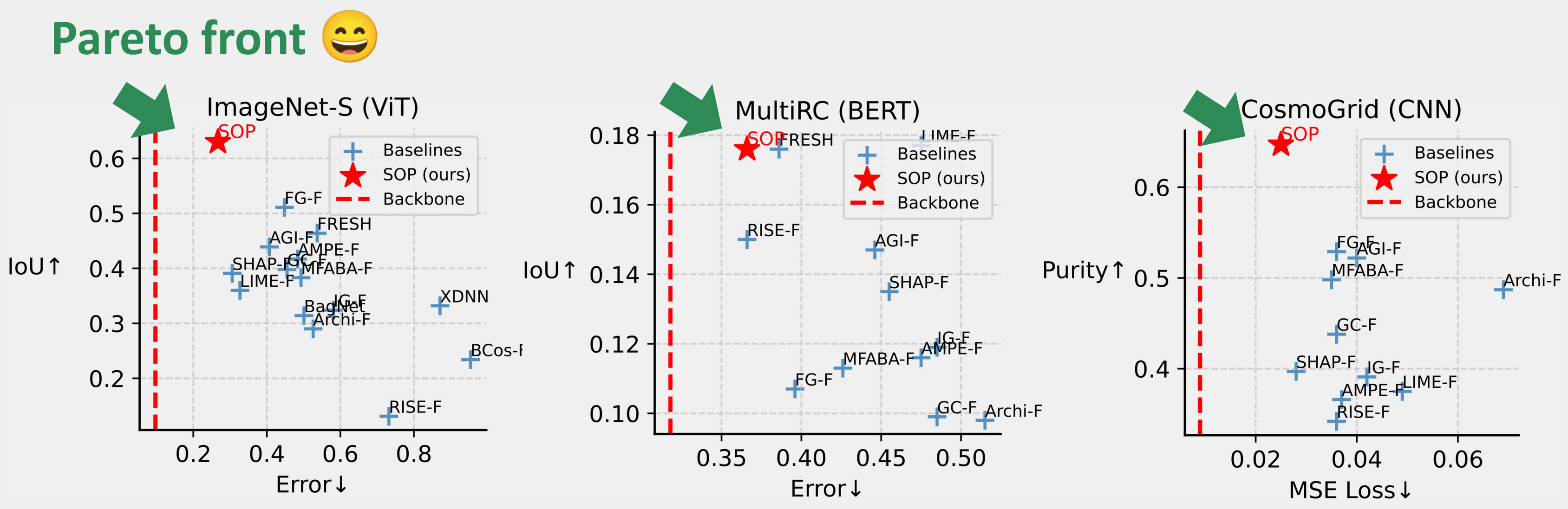
1. **Group Generator** creates feature groups via self attention.
2. **Model Backbone** makes a prediction with each group.
3. **Group Selector** scores each group with cross attention.

$$f(x) = \sum_{i=1}^{m} \underbrace{\theta(\Gamma(x), x)_i}_{\substack{\text{group selector} \\ \textbf{(trained)}}} \cdot \underbrace{h(g_i \odot x)}_{\substack{\text{backbone predictor} \\ \textit{(frozen)}}}, \quad \text{where } \underbrace{g_i \in \Gamma(x)}_{\substack{\text{group generator} \\ \textbf{(trained)}}}.$$

The whole pipeline is trained end-to-end to learn the suitable groups from data.

Sum-of-Parts achieves lowest error + highest purity!

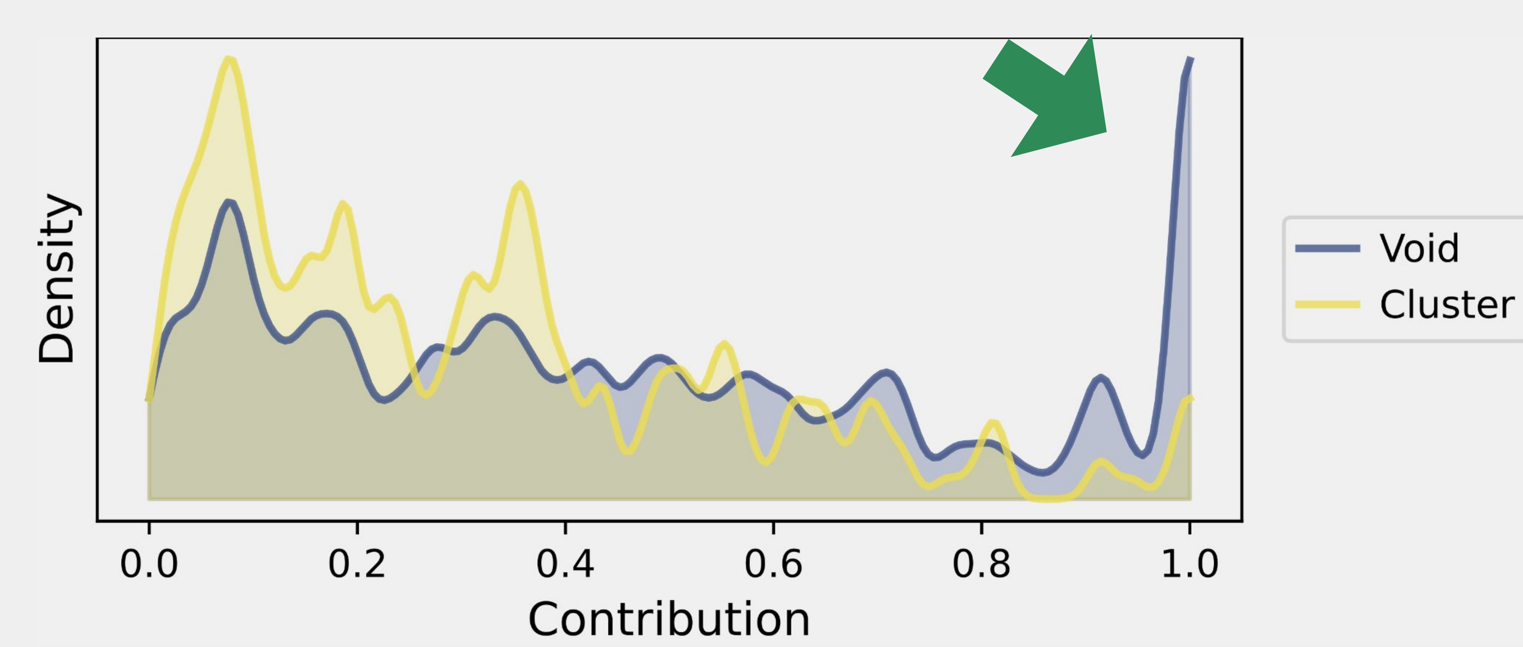
Pareto front 😄



SOP's groups capture objects better

SOP discovers new patterns in cosmology

1. Many voids contribute 100% 😯



2. cosmological structures weigh higher when predicting Ωm 😯