GIO: Gradient Information Optimization for Training Dataset Selection

Dante Everaert

Chris Potts

Data Selection Problem

- Very common problem, especially with ever larger quantities of available data
- When training a model, it can be advantageous to train on a subset of available data

Data Selection Problem

- Very common problem, especially with ever larger quantities of available data
- When training a model, it can be advantageous to train on a subset of available data
  - Data is variable in quality (e.g. crowdworkers, scraped, etc)

Data Selection Problem

- Very common problem, especially with ever larger quantities of available data
- When training a model, it can be advantageous to train on a subset of available data
  - Data is variable in quality (e.g. crowdworkers, scraped, etc)
  - We have a budget

Data Selection Problem

- Very common problem, especially with ever larger quantities of available data
- When training a model, it can be advantageous to train on a subset of available data
  - Data is variable in quality (e.g. crowdworkers, scraped, etc)
  - We have a budget
  - Data needs to be aligned with something (e.g. within a domain)

Data Selection Problem: A Generic Setup

Generic Setup:

- Call the distribution of all available data $\boldsymbol{G}$
- Call the distribution of existing selected data $\boldsymbol{D}$ (can be empty)
- Call the distribution of the ideal selected data $\boldsymbol{X}$

Data Selection Problem: A Generic Setup

Generic Setup:

- Call the distribution of all available data $\boldsymbol{G}$
- Call the distribution of existing selected data $\boldsymbol{D}$ (can be empty)
- Call the distribution of the ideal selected data $\boldsymbol{X}$

Note: No assumption on labels, domain, task, etc. Just generic

Data Selection Problem: A Generic Solution

- Now that we have a generic $\boldsymbol{G}$, $\boldsymbol{D}$ and a target $\boldsymbol{X}$, *any* data selection problem reduces to:

Data Selection Problem: A Generic Solution

- Now that we have a generic $\boldsymbol{G}$, $\boldsymbol{D}$ and a target $\boldsymbol{X}$, *any* data selection problem reduces to:

Identify a subset $\boldsymbol{V}$ of $\boldsymbol{G}$ such that the set $\boldsymbol{D} \cup \boldsymbol{V}$ contains the most information about ideal state $\boldsymbol{X}$

Data Selection Problem: A Generic Solution

- Now that we have a generic $\boldsymbol{G}$, $\boldsymbol{D}$ and a target $\boldsymbol{X}$, *any* data selection problem reduces to:

<u>Identify a subset $\boldsymbol{V}$ of $\boldsymbol{G}$ such that the set $\boldsymbol{D} \cup \boldsymbol{V}$ contains the most information about ideal state $\boldsymbol{X}$</u>

Formally:

$$\text{Choose data } V \subseteq G \text{ such that } \int_{\Omega} p_X(\mathbf{x}) \log \frac{p_X(\mathbf{x})}{p_{D \cup V}(\mathbf{x})} d\mathbf{x} \text{ is minimized}$$

How to minimize the KL divergence?

- Naïve approach: iteratively build the selected set (***D***) by adding the point from ***G*** which most minimizes the KL divergence at each step

$$D \leftarrow D + \underset{\mathbf{v}_i \in G}{\operatorname{argmin}} \int_{\Omega} p_X(\mathbf{x}) \log \frac{p_X(\mathbf{x})}{p_{D \cup \{\mathbf{v}_i\}}(\mathbf{x})} d\mathbf{x}$$

How to minimize the KL divergence?

- Naïve approach: iteratively build the selected set ($\boldsymbol{D}$) by adding the point from $\boldsymbol{G}$ which most minimizes the KL divergence at each step

$$D \leftarrow D + \underset{\mathbf{v}_i \in G}{\operatorname{argmin}} \int_\Omega p_X(\mathbf{x}) \log \frac{p_X(\mathbf{x})}{p_{D \cup \{\mathbf{v}_i\}}(\mathbf{x})} d\mathbf{x}$$

- Intractable – need to recompute the distributions and integral for every point in $\boldsymbol{G}$ at every step

Solution: GIO (Gradient Information Optimization)

1. Rewrite $p_{D\cup\{\mathbf{v}_i\}}(\mathbf{x}) = g(\mathbf{x}, \mathbf{v}_i)$ to eliminate $\boldsymbol{D}$ as it is not changing

   The integral to optimize becomes: $\underset{\mathbf{v}_i \in G}{\operatorname{argmin}} \int_\Omega p_X(\mathbf{x}) \log \frac{p_X(\mathbf{x})}{g(\mathbf{x}, \mathbf{v}_i)} d\mathbf{x}$

Solution: GIO (Gradient Information Optimization)

1. Rewrite $p_{D\cup\{\mathbf{v}_i\}}(\mathbf{x}) = g(\mathbf{x}, \mathbf{v}_i)$ to eliminate $\boldsymbol{D}$ as it is not changing

   The integral to optimize becomes: $\underset{\mathbf{v}_i \in G}{\operatorname{argmin}} \int_\Omega p_X(\mathbf{x}) \log \frac{p_X(\mathbf{x})}{g(\mathbf{x}, \mathbf{v}_i)} d\mathbf{x}$

2. Eliminate p_X and x: Since p_X is unchanging and the integral implicitly removes x, it defines a functional $F\left[g(\mathbf{v})\right]$

Solution: GIO (Gradient Information Optimization)

1. Rewrite $p_{D\cup\{\mathbf{v}_i\}}(\mathbf{x}) = g(\mathbf{x}, \mathbf{v}_i)$ to eliminate $\boldsymbol{D}$ as it is not changing

   The integral to optimize becomes: $\underset{\mathbf{v}_i \in G}{\operatorname{argmin}} \int_\Omega p_X(\mathbf{x}) \log \frac{p_X(\mathbf{x})}{g(\mathbf{x}, \mathbf{v}_i)} d\mathbf{x}$

2. Eliminate p_X and x: Since p_X is unchanging and the integral implicitly removes x, it defines a functional $F\left[g(\mathbf{v})\right]$

2a. Relax the constraint $\mathrm{v}_i \in G$ to the space of all possible v and solve for v_{opt}

Solution: GIO (Gradient Information Optimization)

1. Rewrite $p_{D\cup\{\mathbf{v}_i\}}(\mathbf{x}) = g(\mathbf{x}, \mathbf{v}_i)$ to eliminate ***D*** as it is not changing

   The integral to optimize becomes: $\underset{\mathbf{v}_i \in G}{\operatorname{argmin}} \int_\Omega p_X(\mathbf{x}) \log \frac{p_X(\mathbf{x})}{g(\mathbf{x}, \mathbf{v}_i)} d\mathbf{x}$

2. Eliminate p_X and x: Since p_X is unchanging and the integral implicitly removes x, it defines a functional $F\left[g(\mathbf{v})\right]$

2a. Relax the constraint $\mathrm{v}_i \in G$ to the space of all possible v and solve for v_{opt}

Altogether: $\mathbf{v}_{k+1} \leftarrow \mathbf{v}_k - \gamma \cdot \frac{\partial}{\partial \mathbf{v}_k} \left(\int_\Omega p(\mathbf{x}) \log \frac{p(\mathbf{x})}{g(\mathbf{x}, \mathbf{v}_k)} d\mathbf{x} \right)$

Solution: GIO (Gradient Information Optimization)

3. Once we have v_{opt}, we can just pick the nearest point to v_{opt} in $\boldsymbol{G}$ and that will be the optimal v_i to add to the selected data

Solution: GIO (Gradient Information Optimization)

3. Once we have v_{opt}, we can just pick the nearest point to v_{opt} in $\boldsymbol{G}$ and that will be the optimal v_i to add to the selected data

4. Repeat until KL divergence stops decreasing – information is maximized between selected data distribution and target distribution

Solution: GIO (Gradient Information Optimization)

3. Once we have v_{opt}, we can just pick the nearest point to v_{opt} in $\boldsymbol{G}$ and that will be the optimal v_i to add to the selected data

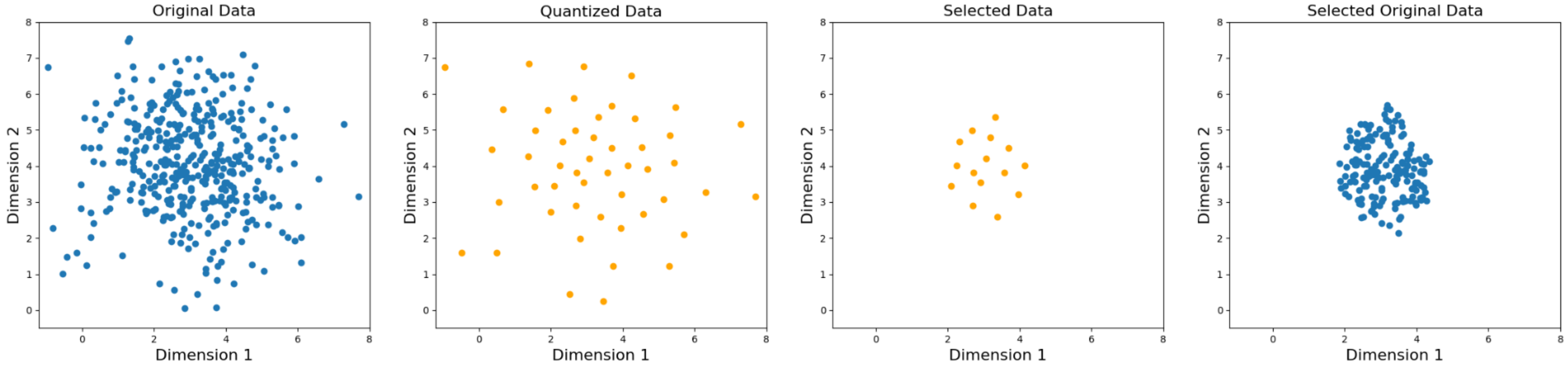
4. Repeat until KL divergence stops decreasing – information is maximized between selected data distribution and target distribution

Note: still no assumption about labels, domain, task etc!

GIO at Scale

Quantization-Explosion:

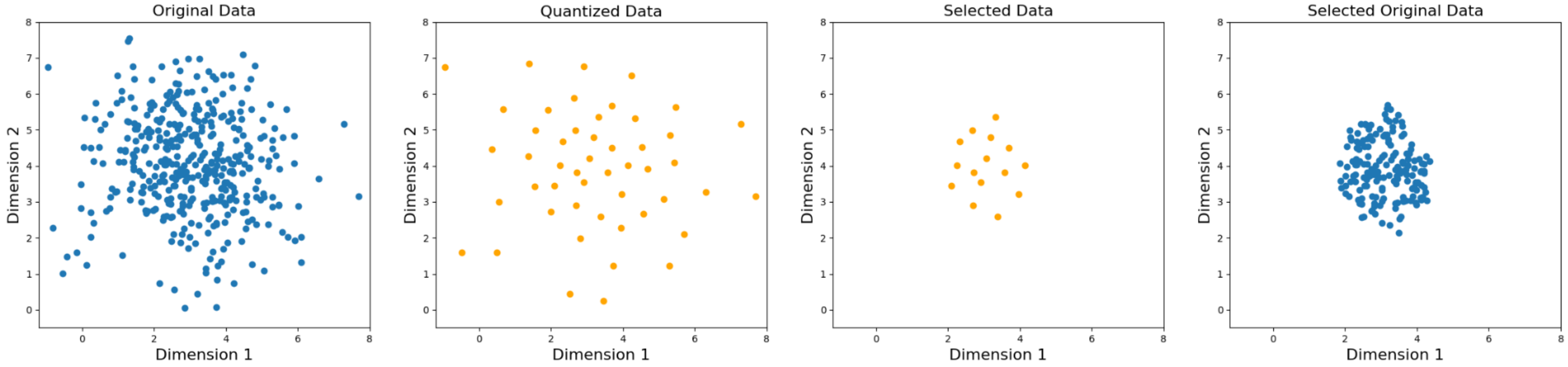
Instead of every point, first quantize the data with K-Means, perform the algorithm, then explode to the original data based on cluster membership



GIO at Scale

Quantization-Explosion:

Instead of every point, first quantize the data with K-Means, perform the algorithm, then explode to the original data based on cluster membership



KL Divergence Estimator[1]: $$\frac{1}{m}\sum_{k=1}^{m}\frac{1}{n}\left[\sum_{i=1}^{n} d\cdot\log\nu_k(i) - d\cdot\log\rho_l(i)\right] + \frac{1}{m}\sum_{k=1}^{m}\log\frac{l\cdot m}{k(n-1)}$$

[1]From Wang et. Al., modified to be an average due to the 0 gradient problem. See Appendix for proof and details

Experimental Results

Experiment 1: Beating WMT-14 from “Attention is all you Need”

Key result: A model trained on GIO-selected data matches and in some cases outperforms a model trained on full data, using only **50%** of the data – and beats all comparative methods in 10/12 evaluations

Experimental Results

Experiment 1: Beating WMT-14 from "Attention is all you Need"

Key result: A model trained on GIO-selected data matches and in some cases outperforms a model trained on full data, using only **50%** of the data – and beats all comparative methods in 10/12 evaluations

Experiment 2: Selecting High-quality Data (Spelling Correction Task)

Key result: GIO selects 73% high quality data from a mixed set, compared to <60% for comparative methods

Experimental Results

Experiment 1: Beating WMT-14 from “Attention is all you Need”

Key result: A model trained on GIO-selected data matches and in some cases outperforms a model trained on full data, using only **50%** of the data – and beats all comparative methods in 10/12 evaluations

Experiment 2: Selecting High-quality Data (Spelling Correction Task)

Key result: GIO selects 73% high quality data from a mixed set, compared to <60% for comparative methods

Experiment 3: Reducing Training Set Size to 25% (Image – FashionMNIST)

Key result: GIO-selected data leads to only a 2.3% performance loss, compared to 3% with random selection

Conclusion

GIO is a robust domain- and task- agnostic method and applies to any data with continuous representation out of the box with few assumptions

Conclusion

GIO is a robust domain- and task- agnostic method and applies to any data with continuous representation out of the box with few assumptions

A model trained on **GIO**-selected data can match or outperform models trained on the full set and outperforms all comparative methods on various tasks

Conclusion

GIO is a robust domain- and task- agnostic method and applies to any data with continuous representation out of the box with few assumptions

A model trained on **GIO**-selected data can match or outperform models trained on the full set and outperforms all comparative methods on various tasks

GIO can be used to select high quality data, aligned data to certain domains/intent, reduce the train set size to fit a budget, and more

More Information

Paper: https://arxiv.org/pdf/2306.11670.pdf

Github: https://github.com/daeveraert/gradient-information-optimization/tree/main

Pip Install: “pip install grad-info-opt”

My Contact Information: Dante Everaert, dante.everaert@gmail.com