Semi-Supervised Learning for Materials Informatics

Kat Nykiel

Purdue University

November 20, 2023

Bridging the Gap

Materials Informatics

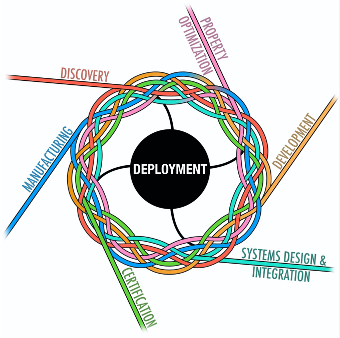

Figure 1: Materials Genome Initiative: (mgi.gov)

Semi-Supervised Learning

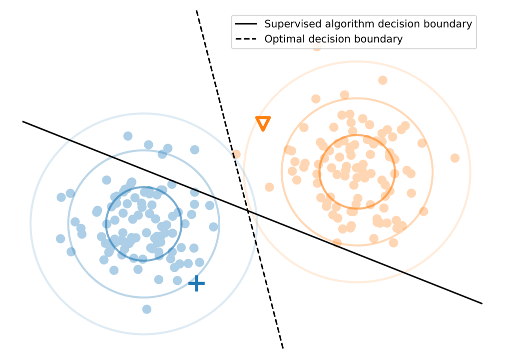

Figure 2: SSL Example (van Engelen, 2020)

Collaboration between these two fields is limited; we need to bridge the gap

State-of-the-Art

Materials Informatics

- *MatBench* dataset provides benchmark for model performance
- GNN models show high predictive accuracy, on par with simulations

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{X}, [M, u]) \qquad (1)$$

- $v_i \in \mathcal{V}$ is atom information for i
- $e_{ij} \in \mathcal{E}$ is the bond information

Semi-Supervised Learning

- *FlexMatch* uses a *flexible confidence threshold* to identify high-confidence pseudo-labels

$$\sigma_t(c) = \sum_{n=1}^{N} 1(\max(p_{m,t}(y|u_n)) > \tau) \cdot 1(\operatorname{argmax}(p_{m,t}(y|u_n)) = c) \qquad (2)$$

Restriction: FlexMatch uses 32x32 images as input

How can we apply this model to materials informatics?

Representing Crystal Structures as Images

Idea

Use graph representations of materials to generate images

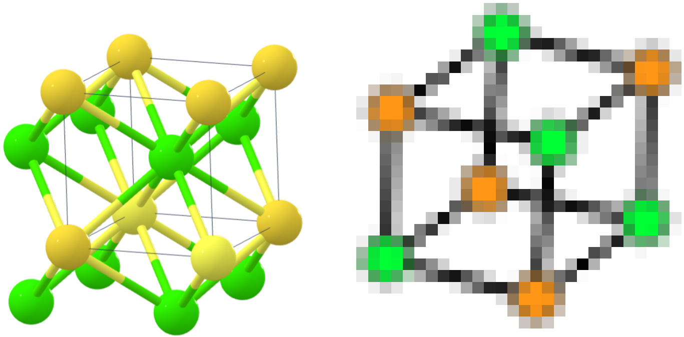

Figure 3: NaCl crystal structure and its 32x32 embedding

Data Augmentation

- Random horizontal flipping
- Random cropping with padding
- Normalization of pixel values

Hyperparameters

- 1 epoch
- 5000 iterations
- 10,000 data points
- 500 labels
- 80/10/10 $\mathcal{T}/\mathcal{E}/\mathcal{V}$

Benchmarking FlexMatch on CIFAR-10 Dataset

CIFAR-10 Dataset

Table 1: Measured *FixMatch* and *FlexMatch* performance

Method	Accuracy
FixMatch	0.936
FlexMatch	0.867

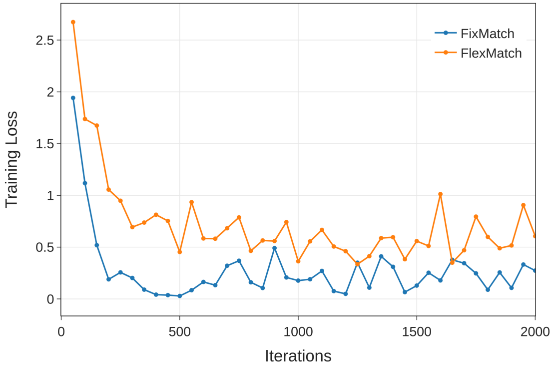

Figure 4: Training loss on CIFAR-10 dataset

FixMatch slightly outperforms FlexMatch on CIFAR-10 dataset*

Benchmarking FlexMatch on MatBench Dataset

MatBench Dataset

Table 2: Measured *FixMatch* and *FlexMatch* performance

Method	Accuracy
FixMatch	0.582
FlexMatch	0.586

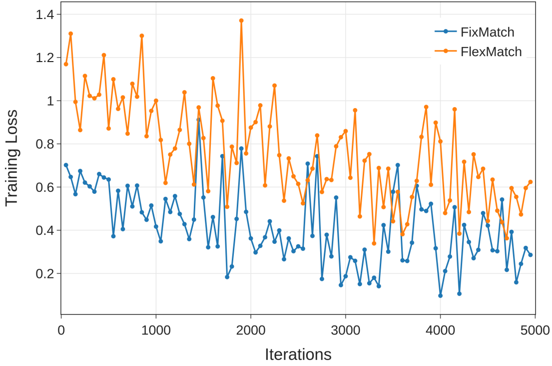

Figure 5: Training loss on MatBench dataset

Both FixMatch and FlexMatch struggle to train on this data; more work is needed

Conclusions and Future Work

Conclusions

- FixMatch outperforms FlexMatch on CIFAR-10 dataset, contradictory to paper results.
- More work is needed to overcome the image translation barrier between materials informatics and semi-supervised algorithms.

Future Work

- Train models for more epochs
- Refine dataset to better capture chemistry
- Publish!