87 Deep Learning for Microparticle Characterisation in Medicines Manufacturing

Nikilkumar Patel | Dr. Javier Cardona

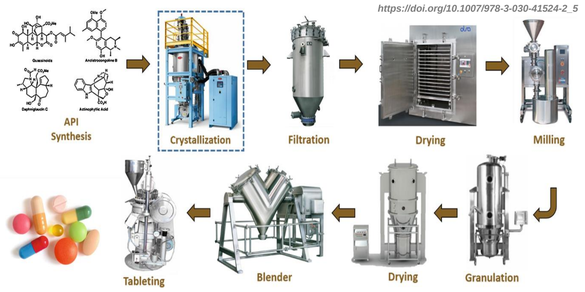
EE997 - MSc. Machine Learning & Deep Learning, Department of Electronic & Electrical Engineering

University of Strathclyde Engineering

CMAC FUTURE MANUFACTURING RESEARCH HUB

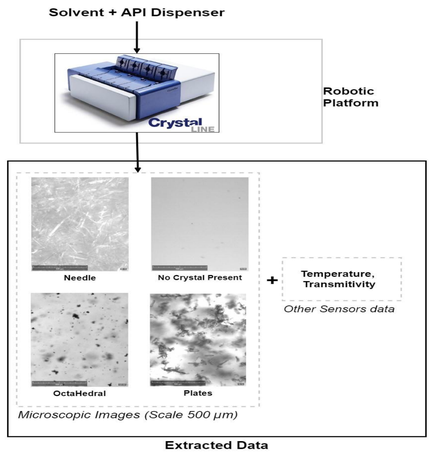
1. Introduction

Crystallisation in Pharma



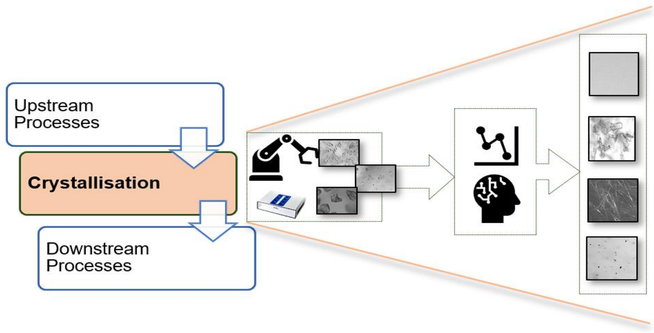
- To isolate the compound as a solid with the polymorphic at a high yield and with minimal impurities
- Robust crystallization = Better Quality product

Crystal Outcome Screening



2. Current Solution and Challenges

Crystal Shape Classification using Supervised Deep Learning



Supervised Models for Computer Vision Tasks

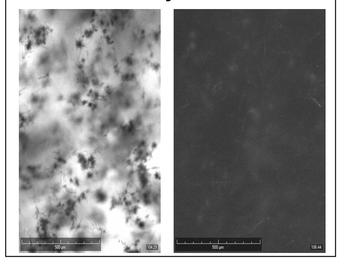
ImageNet Classifiers
- ResNet (18,34,50)
- EfficientNet and more..

Object Segmentations
- R-CNN
- Mask R-CNN and more..

Challenges in Supervised Learning

Annotations Cost
- Manual: Crowd Source vs Experts
- Automated Platforms : Amazon Rekognition, GCP
- Estimated Market Value **in 2022**: **US$0.8B** vs Forecasted Market Value **by 2027 US$3.6B**

Task Complexity
- Is this a Plate or Needle? What to label if it is visibly too dark?



Common Sense
- Deviation of the truth from Human perception
- Learnings can be rigid and does not always work accurately

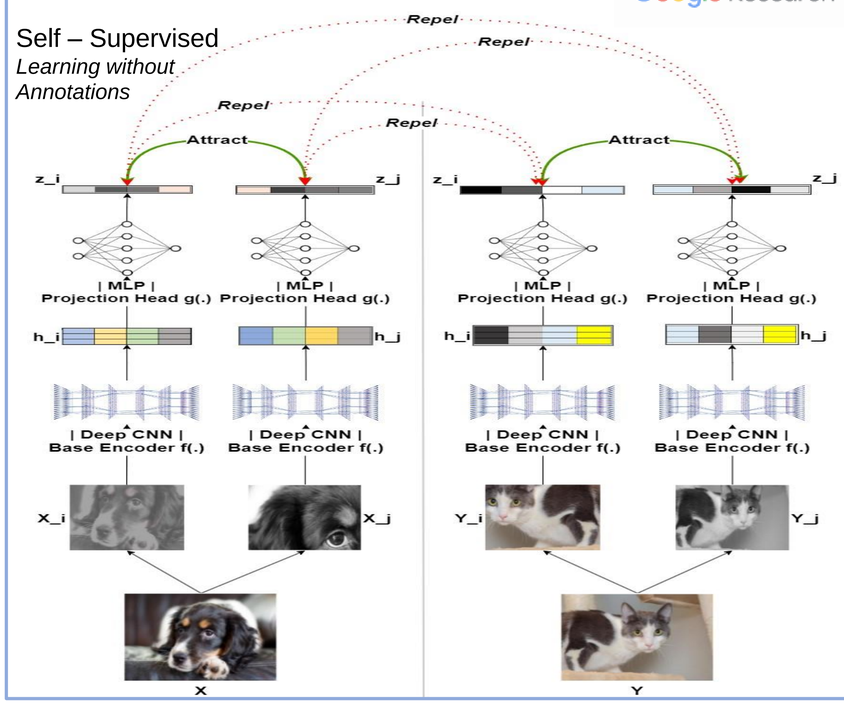
Source: www.freecodecamp.org

3. Proposed Methodology

SimCLR: Simple Framework for Contrastive Learning of Visual Representations

Google Research

Self – Supervised
Learning without Annotations



Data Augmentation

Contrastive Loss

NT- Xent Pairwise Loss

$$\mathbf{L} = \frac{\sum_{k=1}^{N}[\, l(2k-1, 2k) + l(2k, 2k-1)]}{2N}$$

$$\textbf{Pair loss: } l(i,j) = -\log \frac{e^{S_{i,j}/T}}{\sum_{k=1}^{2N} e^{S_{i,k}/T}}$$

$$\textbf{Cosine Similarity: } S_{i,j} = sim(X_i, X_j)$$

$$T : \textbf{Temperature}$$

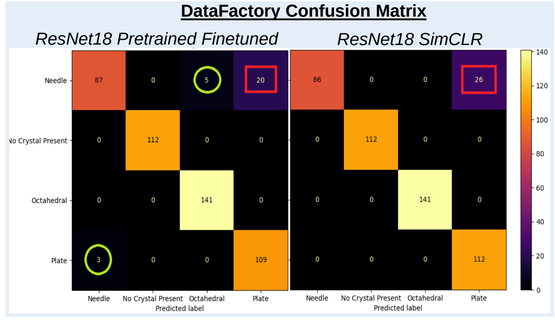
Network Architecture

Simple structure, No Memory Bank

4. Results and Discussion

Dataset	Datapoints	Learning Rule	Model	Accuracy	F1
CIFAR10	60000	Supervised	ResNet18 ImageNet Pretrained & Finetuned	38.14	37.83
CIFAR10	60000	Supervised	ResNet18 Fully Trained	72.84	72.61
CIFAR10	***60000***	***Self-Supervised***	***ResNet18 encoder SimCLR***	***59.48***	***58.25***
DataFactory	4344	Supervised	ResNet18 ImageNet Pretrained & Finetuned	94.13	93.99
DataFactory	4344	Supervised	ResNet18 Fully Trained	92.45	92.42
DataFactory	***4344***	***Self-Supervised***	***ResNet18 encoder SimCLR***	***94.55***	***94.47***

DataFactory Confusion Matrix

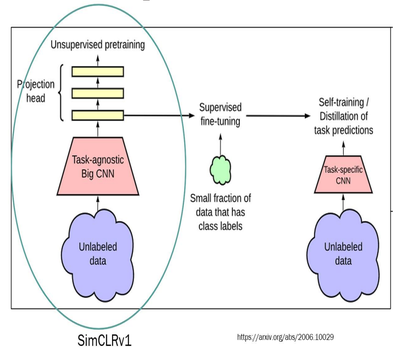


- SimCLR outperforms all supervised models for our experiment data; can match supervised models if trained for longer on CIFAR10 data with better suitable optimizer
- Having required no labels at all for training, SimCLR seems to address the challenges faced by supervised approach, with decent accuracy and more distinctive classification
- For N augments per image, training data increases N times; longer training time and more computation power

5. Future Direction

SimCLR v2 Semi Supervised

Google Research



- Modified architecture and training with small fraction of annotations.
- Training with **10% annotations** SimCLRv2 surpasses SOTA both supervised and unsupervised models.