What is common btw Negative Transfer, Uncertainty and Food Recognition?

Petia Radeva

Collaboration with:
Eduardo Aguilar, Bhalaji Nagarajan

University of Barcelona &
Computer Vision Center

petia.ivanova@ub.edu

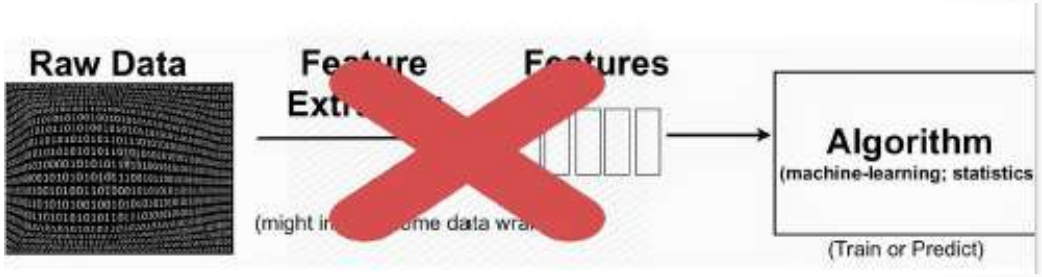
Contents

- Negative Transfer
- Uncertainty modeling
- GANs
- Epistemic uncertainty-based Data augmentation
- Conclusions

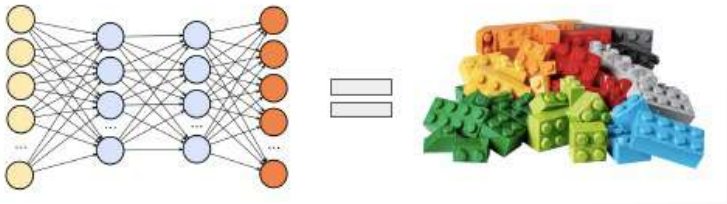
What makes DL so popular?

Advantages:

- 1. Self-learned high-level features representations



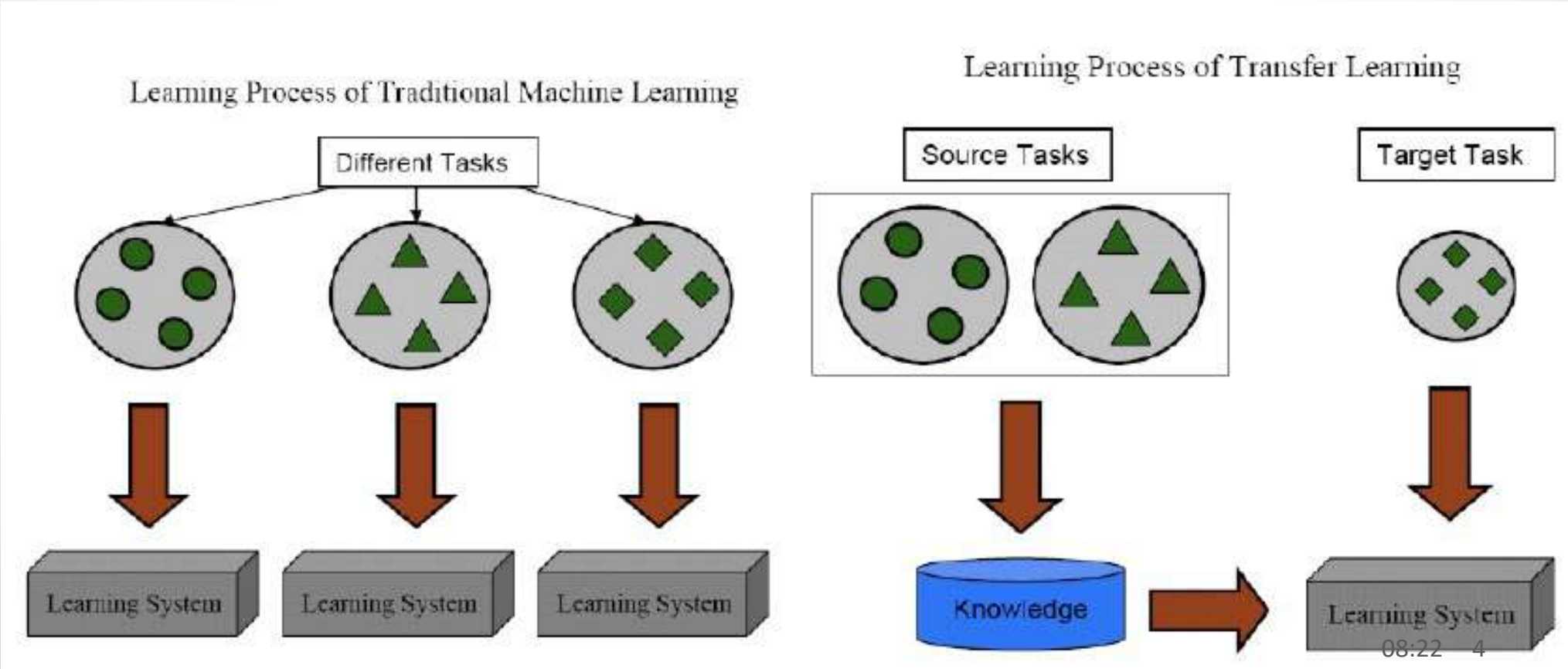
- 2. Modularity



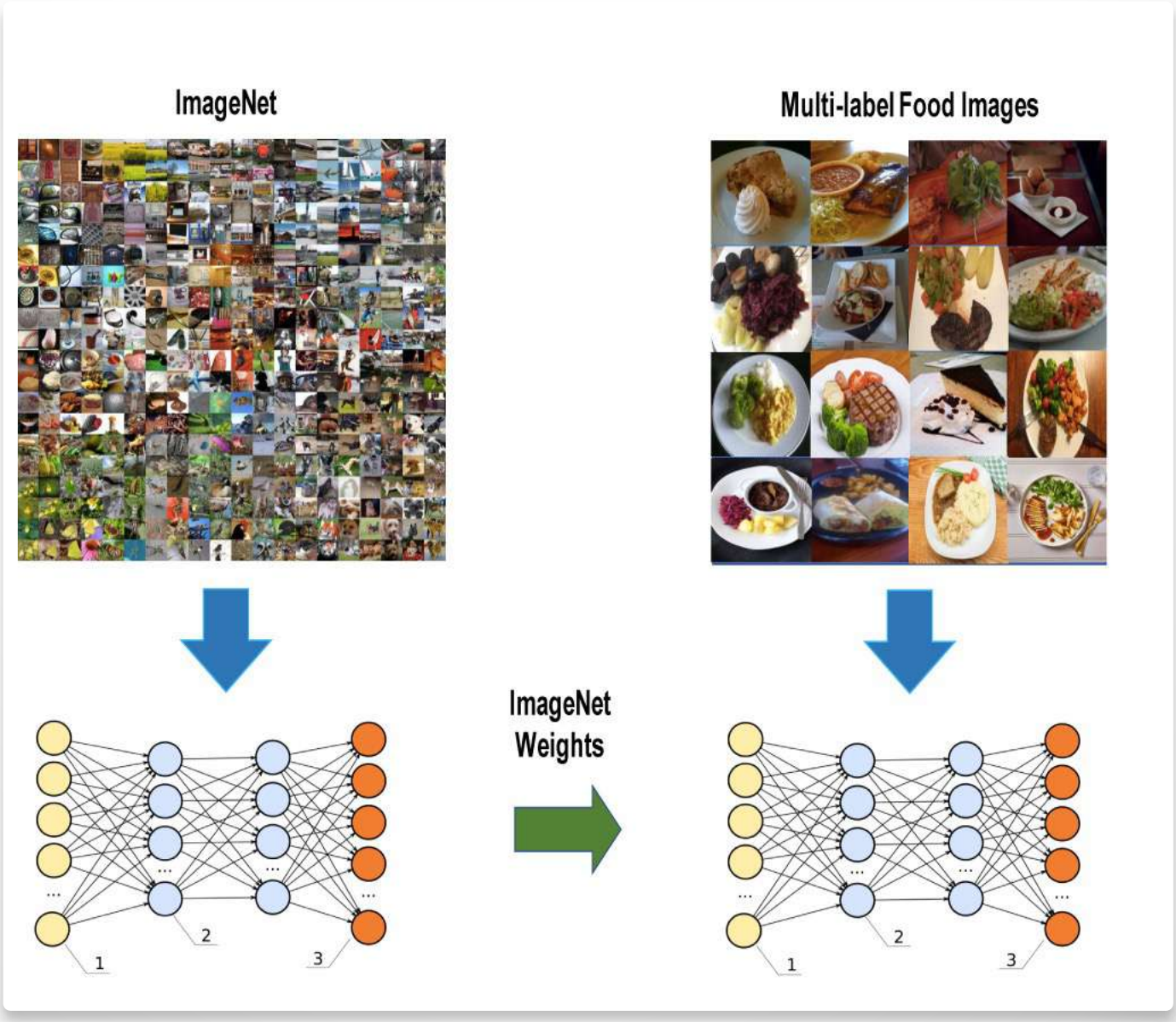
- 3. Transfer Learning

Transfer Learning

- Building every model from scratch is time-consuming and expensive
- There are many existing knowledge. Can we reuse them?



Food Recognition (FR)



Why is the food recognition a challenge?

Negative transfer

- Negative transfer: when source domain data and task contribute to reduced performance of learning in the target domain

- Causes:
  - Domains are too dissimilar [RMKD05]
  - Tasks are not well-related [BH03], etc.

- Similarity measures
  - Cosine similarity
  - Kullback-Leibler divergence
  - Jensen-Shannon divergence
  - Maximum Mean Discrepancy [BGR+02], etc.

Single- vs Multi-Label FR

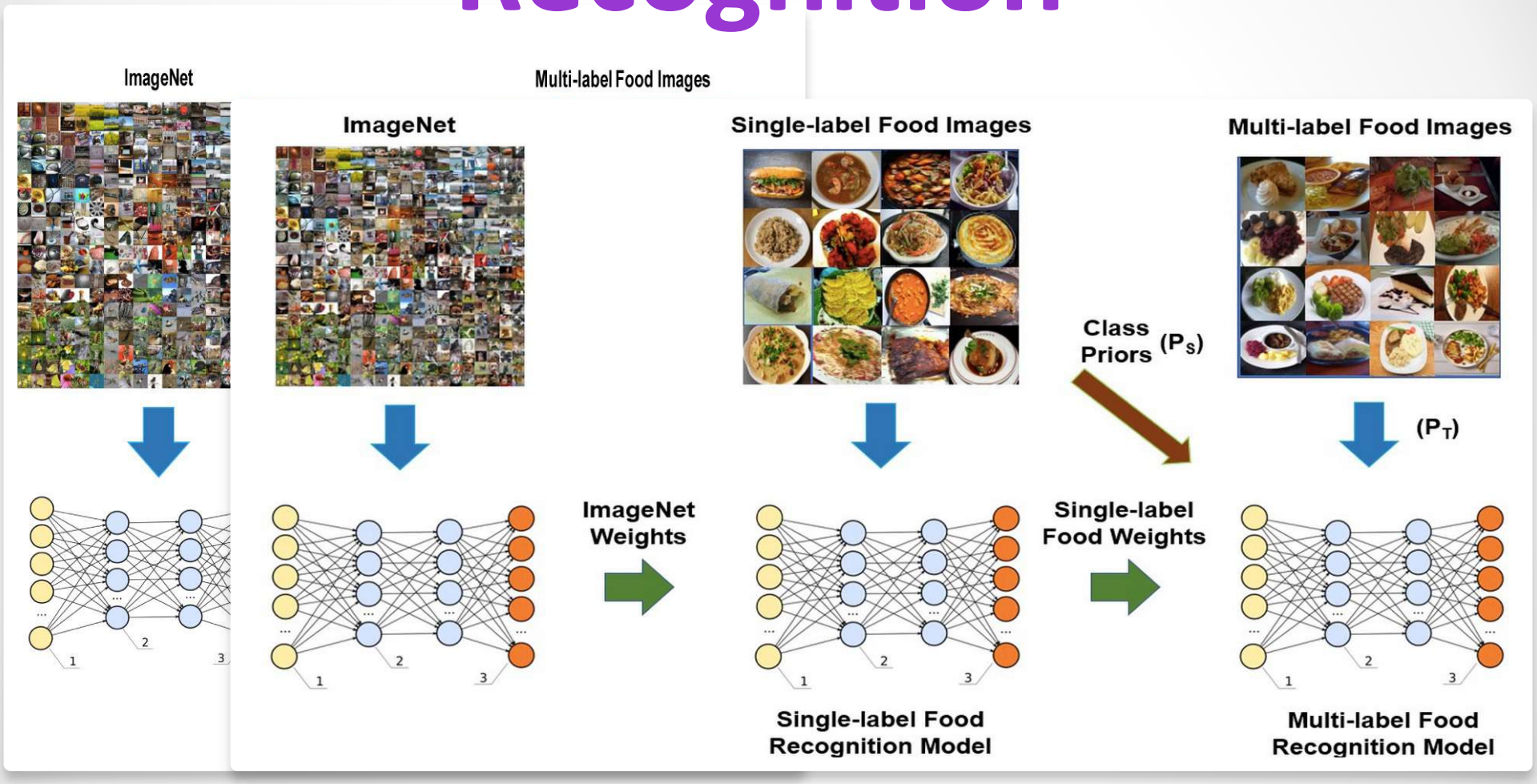
- Food-101 Food-201 our Combo-plates

Single-Label Recognition always has better performance than Multi-label Rec.

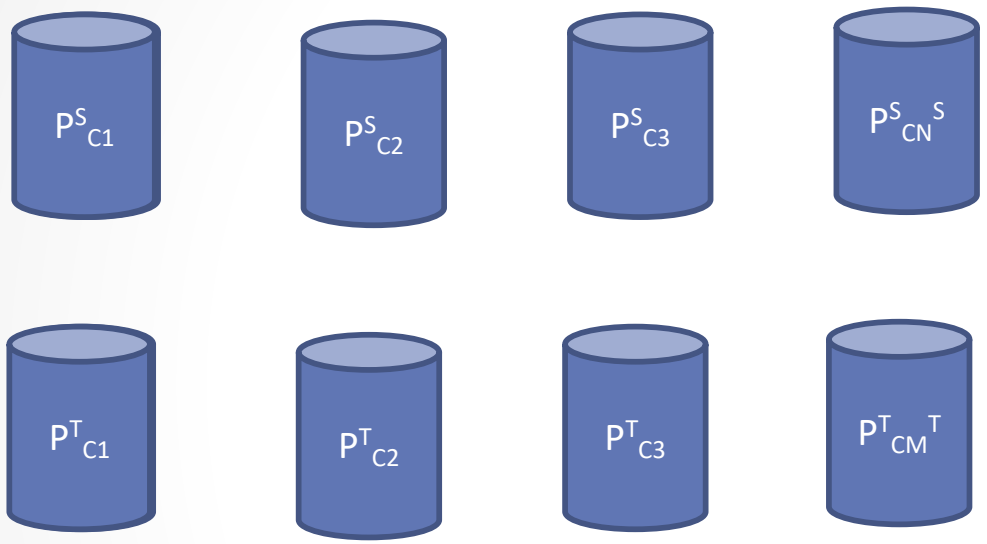
Most public datasets for FR are single-label.

- However, we used to eat multiple food (so real FR should be a multi-label FR).

Single-to-Multi-Label Recognition



Class Priors for Improved Transferability



P^S is the prior of the source domain classes

P^T is the prior of the target domain classes

- The ratio of priors, P^T/P^S, can be used to enhance the transferability of the domain knowledge.

- Bhalaji Nagarajan, Eduardo Aguilar, Petia Radeva, S2ML-TL Framework for Multi-label Food Recognition. ICPR Workshops (5) 2020: 629-646

S2ML-TL Framework

- Prior Computation: $P(T_i) = \frac{1}{N_t} \sum_{n=1}^{N_t} y_i^n$

$$P(S_i) = \frac{1}{N_s} \sum_{j=1}^{N_s} y_j,\ j \in \{all\ source\ classes\ containing\ i\}$$

- During TL, the training of the target domain starts with initialization of weights from the source domain.

$$r_i^b = \beta \left[\alpha \frac{P_T}{P_S} + (1 - \alpha) P_T \right]$$

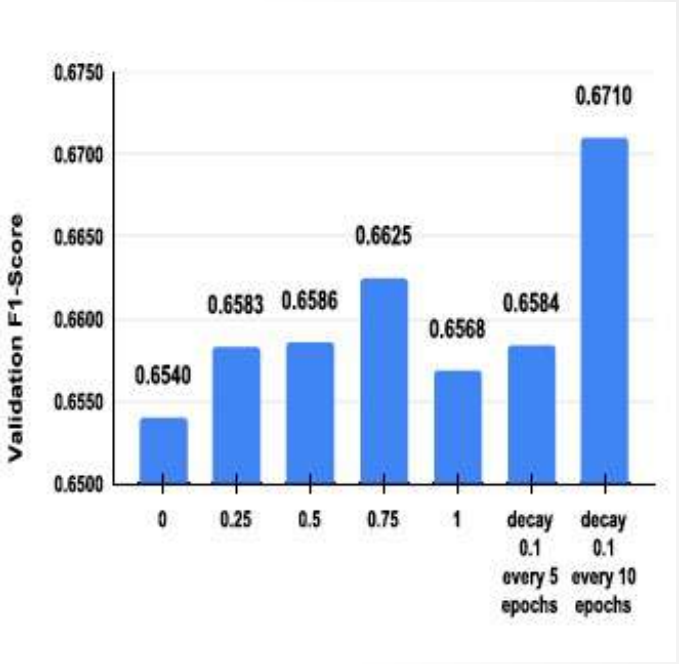
- Prior-induced ML loss function:

$$l_p(x, y) = \frac{1}{C} \sum_{c=1}^{C} [\ y_i.log(p(y_i)) + (1 - y_i).log(1 - p(y_i))\ ] * r_i^b$$

Validation

Validation

- Datasets: Food101, Foos201, Combo-plates
- Evaluation metrics: Precision, Recall and F1-score
- Implementation details: InsepctionresnetV2, Resnet50, etc.
- Hyper-parameter selection:



Performance of models with different α values

- Bhalaji Nagarajan, Eduardo Aguilar, Petia Radeva, S2ML-TL Framework for Multi-label Food Recognition. ICPR Workshops (5) 2020: 629-646

Validation

Model	Validation data			Test data		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Standard TL	0.7209	0.5865	0.6468	0.7152	0.5840	0.6430
ERM	0.6991	0.5667	0.6260	0.6900	0.5700	0.6200
KL	0.7030	0.6212	0.6596	0.6984	0.6173	0.6553
Priors	0.7045	0.6229	**0.6612**	0.7000	0.6200	**0.6600**

Model performance of InceptionResnetV2 on Combo-plates

Model	Validation data			Test data		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Standard TL	0.7204	0.4215	0.5319	0.7322	0.4636	0.5678
ERM	0.7518	0.4546	0.5666	0.7493	0.4800	0.5852
KL	0.7918	0.4317	0.5587	0.7740	0.4370	0.5586
Priors	0.7767	0.5877	**0.6691**	0.7313	0.5400	**0.6213**

Model performance of Resnet50 on Food201

Ablation study

Model	$\alpha * \frac{P_T}{P_S}$	P_T	β	Validation data			Test data		
				Prec.	Recall	F1 Sc.	Prec.	Recall	F1 Sc.
$\alpha = 1$ (w/o β)	x	-	-	0.7537	0.5685	0.6481	0.7394	0.5666	0.6415
$\alpha = 1$ (w β)	x	-	x	0.7237	0.6012	0.6568	0.7200	0.6000	0.6500
Decayed α (w/o β)	x	-	-	0.7281	0.5958	0.6553	0.7112	0.5932	0.6469
Decayed α (w β)	x	-	x	0.7118	0.6094	0.6567	0.6963	0.6086	0.6495
Target priors (w/o β)	-	x	-	0.6972	0.6238	0.6585	0.6811	0.6140	0.6458
Target priors (w β)	-	x	x	0.7092	0.6068	0.6540	0.7050	0.5994	0.6479
Proposal (w/o β)	x	x	-	0.6996	0.6034	0.6480	0.7045	0.6069	0.6521
Proposal (w β)	x	x	x	0.7114	0.6349	**0.6710**	0.7011	0.6127	**0.6539**

- Bhalaji Nagarajan, Eduardo Aguilar, Petia Radeva, S2ML-TL Framework for Multi-label Food Recognition. ICPR Workshops (5) 2020: 629-646

Conclusions (I)

- Using single-label recognition increases the learning ability of the multi-label recognition task.
- The recognition performance could be further boosted by using class priors.
- Classes that are showing a negative trend are classes:
  - difficult to classify (such as ketchup)
  - have a larger variation inside the same class (such as potatoes).
- The performance of such classes used to have
  - Fine-grained nature of the dataset
  - High classification uncertainty.

Contents

- Negative Transfer
- Uncertainty modeling
- GANs
- Epistemic uncertainty-based Data augmentation
- Conclusions


Let’s talk about uncertainty

In DL many unanswered questions...

- Why ~~doesn't~~ does my model work?
- What does my model know?
- Why does my model predict this and not that?

- **Our models are black boxes and not interpretable...**
- Physicians and others need to understand why a model predicts an output.

Uncertainty in ML

- For Computer scientists, computers and algorithms are **deterministic**.

"Still, it can be surprising that machine learning makes heavy use of probability theory."

- The reason that the **answers are unknown** is because of uncertainty.
- The solution is to **systematically evaluate different solutions** until a good or good-enough set of features and/or algorithm is discovered for a specific prediction problem.

- https://machinelearningmastery.com/uncertainty-in-machine-learning/

Noise in observations

- Noise refers to **variability or randomness** in the observation.
- The real world, and in turn, real data, is **messy or imperfect**.
  - As practitioners, we must remain sceptical of the data and **develop systems to expect and even harness this uncertainty**.

https://machinelearningmastery.com/uncertainty-in-machine-learning/

Incomplete Coverage of the Domain

- In statistics, a random sample refers to a collection of observations chosen from the domain without systematic bias.
  - **However, there will always be some bias.**
- A suitable level of **variance and bias** in the sample is **required** such that the **sample is representative** of the task or project for which the data or model will be used.
  - Often, we have <u>little control</u> over the sampling process.

Incomplete Coverage of the Domain

- In all cases, we will never have all of the observations.
  - If we did, a predictive model would not be required.
- This is why we **split a dataset into train and test** sets or use resampling methods like k-fold cross-validation.
  - We do this to handle the uncertainty in the representativeness of our dataset and estimate the performance of a modelling procedure on data not used in that procedure.

https://machinelearningmastery.com/uncertainty-in-machine-learning/

Imperfect Model of the Problem

- This is often summarized as “all models are wrong,”
- Aphorism by George Box:

  “All models are wrong but some are useful”

- This does not **apply** just to the model, the artefact, but the **whole procedure** used to prepare it, including the choice and preparation of data, choice of training hyperparameters, and the interpretation of model predictions.

https://machinelearningmastery.com/uncertainty-in-machine-learning/

Imperfect Model of the Problem

- Another type of error is an **error of omission**.

"In many cases, it is more practical to use a simple but uncertain rule rather than a complex but certain one, even if the true rule is deterministic and our modelling system has the fidelity to accommodate a complex rule."

- Given we know that the models will make errors, we handle this uncertainty by seeking a model that is good enough.

https://machinelearningmastery.com/uncertainty-in-machine-learning/

Model uncertainty

1. Given a model trained with several pictures of fruits, a user asks the model to decide what is the object using a photo of a chocolate cake.

Who is the “responsible” for this?

$$S(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

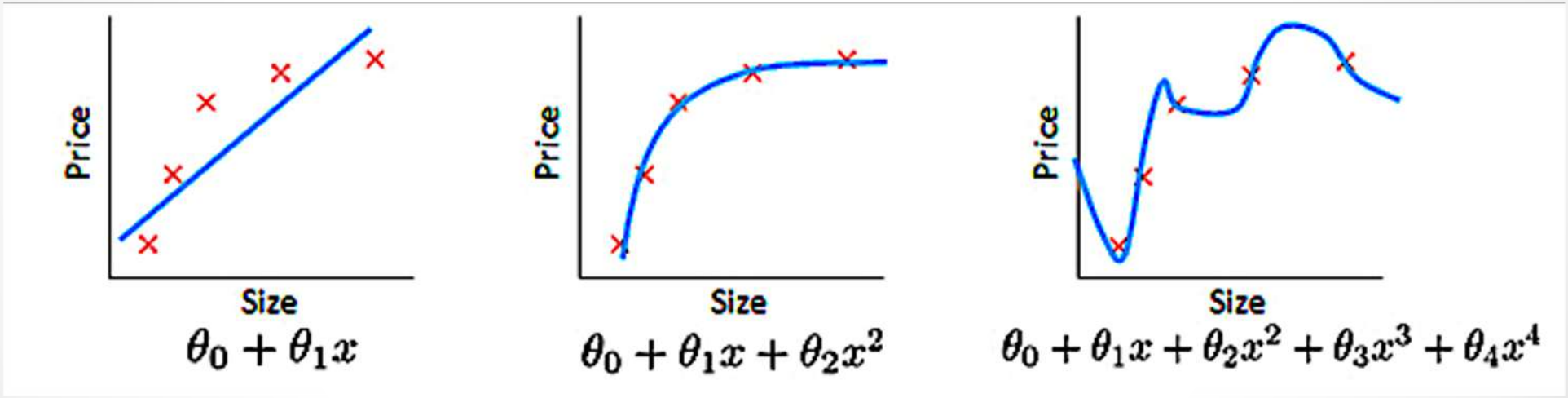
Model uncertainty

2. We have different types of images to classify fruits, where one of the category comes with a lot of clutter/noise/occlusions.

- Adapted from Gal (2016)

Model uncertainty

3. What is the best model parameters that best explain a given dataset? What model structure should we use?



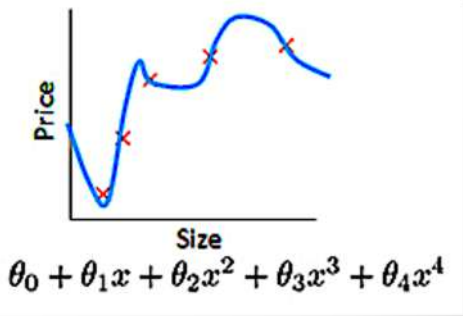
Gal (2016)

Types of uncertainty in Bayesian modeling

Aleatoric – captures the noise inherent in the observations

- heteroscedastic – data-dependent
- homoscedastic – constant for different data points,
  - but can be task-dependent.

- **Epistemic** – model uncertainty
  - Can be explained away given enough data
  - Uncertainty about the model parameters
  - Uncertainty about the model structure



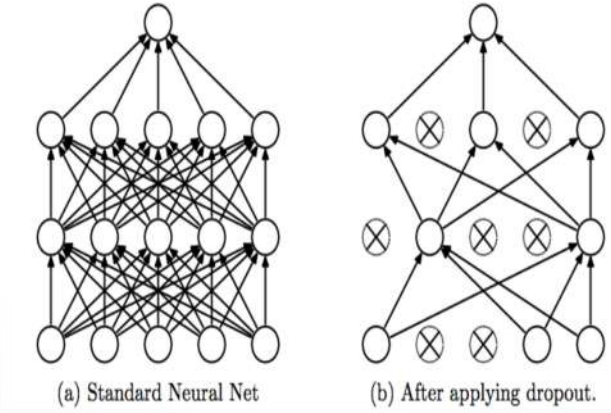
Trustworthy and uncertainty aware methods as a solution to the problem of unreliable models

- Bayes-by-Backprop
- Stochastic Weight Averaging Gaussian (SWAG)
- Deep Ensemble Models
- Monte-Carlo Dropout
- Stochastic Variational Inference Networks
- Multiplicative Normalizing Flow Networks
- Evidential Deep Learning Models -Dirichlet Distribution
- Temperature-Scaling

How to estimate the Epistemic Uncertainty?

Gal and Ghahramani showed that dropout at inference time gives an uncertainty estimator:

1. Infer y|x multiple times, each time sample a different set of nodes to drop out.
2. Average the predictions to get the final prediction E(y|x).
3. Calculate the sample variance of the predictions.



How to estimate the Epistemic Uncertainty?

The Epistemic Uncertainty (EU) can be expressed as follows:

where

$$EU(x_t) = -\sum_{c=1}^{C} \overline{p(y_c = \hat{y}_c|x_t)} \ln(\overline{p(y_c = \hat{y}_c|x_t)}),$$

K Monte Carlo dropout simulations

$$\overline{p(y_c = \hat{y}_c|x)} = \frac{1}{K} \sum_{k=1}^{K} p(y_c^k = \hat{y_c^k}|x).$$

Class uncertainty

What to do with the difficult classes?

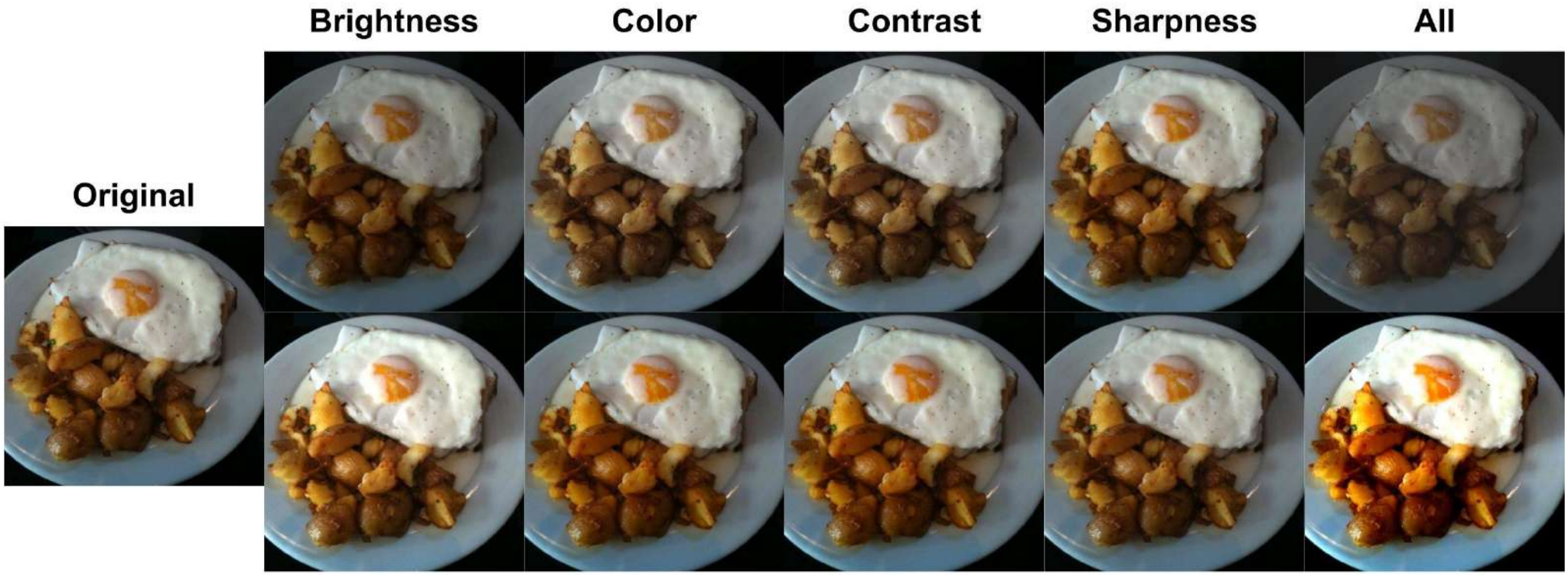
Are all classes well represented/easily discriminable?

Adapted from Gal (2016)

How to augment difficult classes?

- - data augmentation

Use Uncertainty for Data Augmentation



Different augmentation results

How to augment difficult clases?

- classic data augmentation, or
- creating synthetic images. How?

Contents

- Negative Transfer
- Uncertainty modeling
- GANs
- Epistemic uncertainty-based Data augmentation
- Conclusions

The Biggest Breakthrough In The History Of AI

- Celebrated computer scientist Yann LeCun observed:

“GANs and the variations that are now being proposed is the most interesting idea in the last 10 years in ML, in my opinion.”

VP and Chief AI Scientist, Facebook
Silver Professor of Computer Science, Data Science, Neural Science, and Electrical and Computer Engineering, New York University.
ACM Turing Award Laureate, (sounds like I'm bragging, but a condition of accepting the award is to write this next to you name)
Member, National Academy of Engineering

So use GANs to create synthetic samples from other samples.

But:

Can there be data that require more attention than others?

Samples of the Smallest and Largest EU within the same class of dish


Caesar Salad


Ravioli


Steak

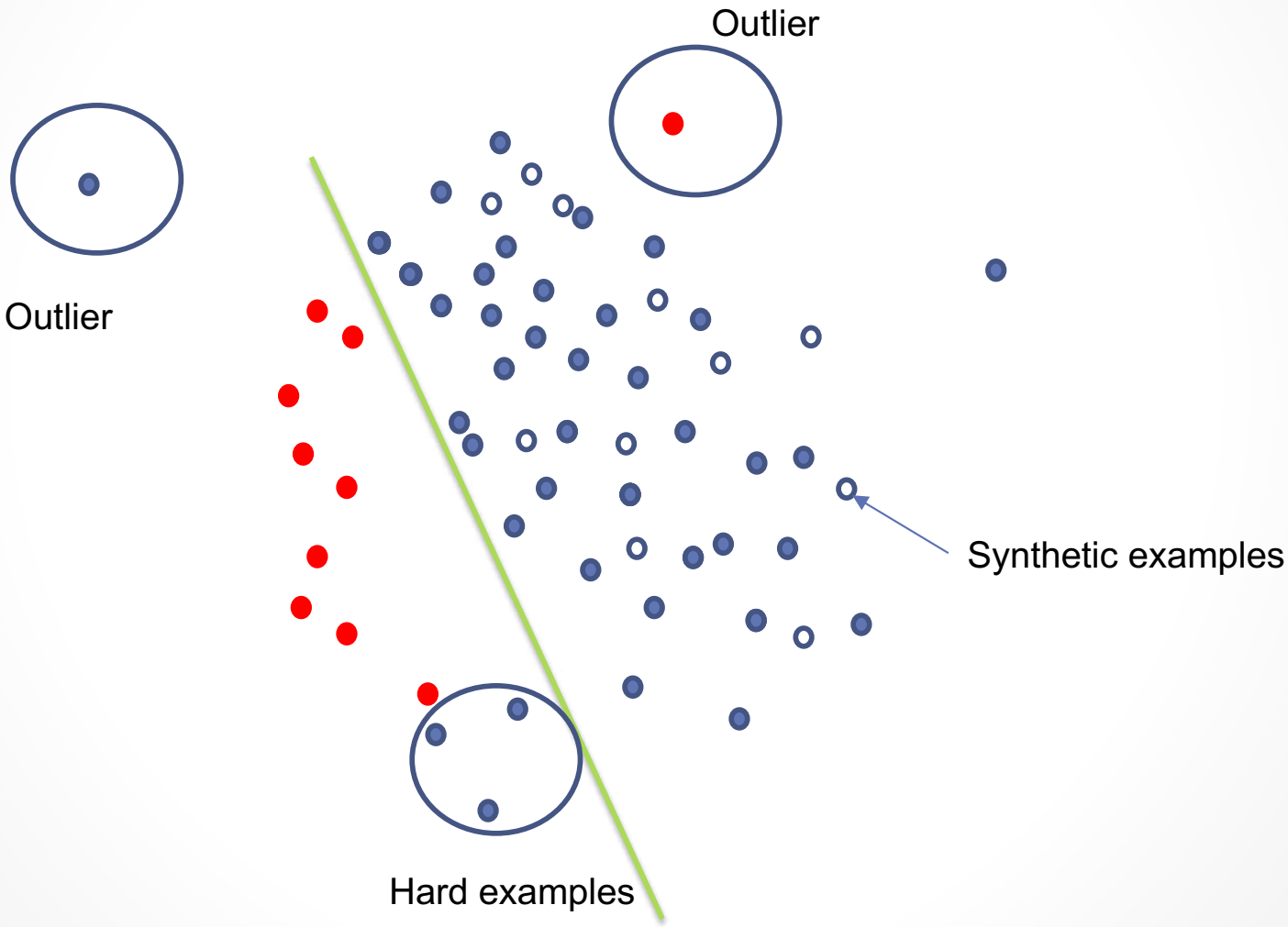

Tacos

Images by E. Aguilar, and P. Radeva. Uncertainty-aware Integration of Local and Flat Classifiers for Food Recognition, Pattern Recognition Letters, 2020

Contents

- Uncertainty modeling
- GANs
- Data augmentation based on epistemic uncertainty for food analysis
- Conclusions

Can there be data that require more attention than others?



Can there be data that require more attention than others?

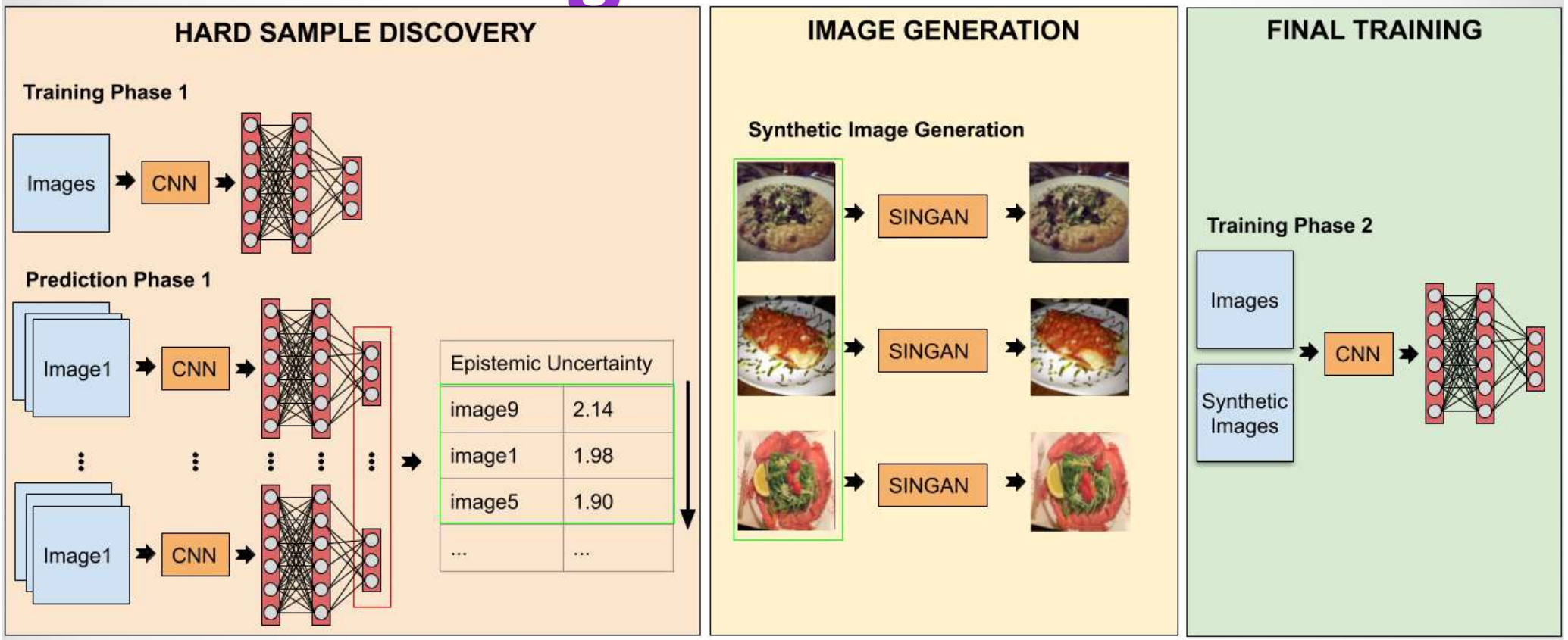
- Within a dataset, there could be more and less complex classes.
- Some images could be more beneficial for the classification than others.

More emphasis should be placed on these data.

- **Hypothesis**: Estimating the uncertainty can help us decide the most appropriate classes to perform additional data augmentation methods.

- How to create additional synthetic data?

Use Uncertainty for Data Augmentation

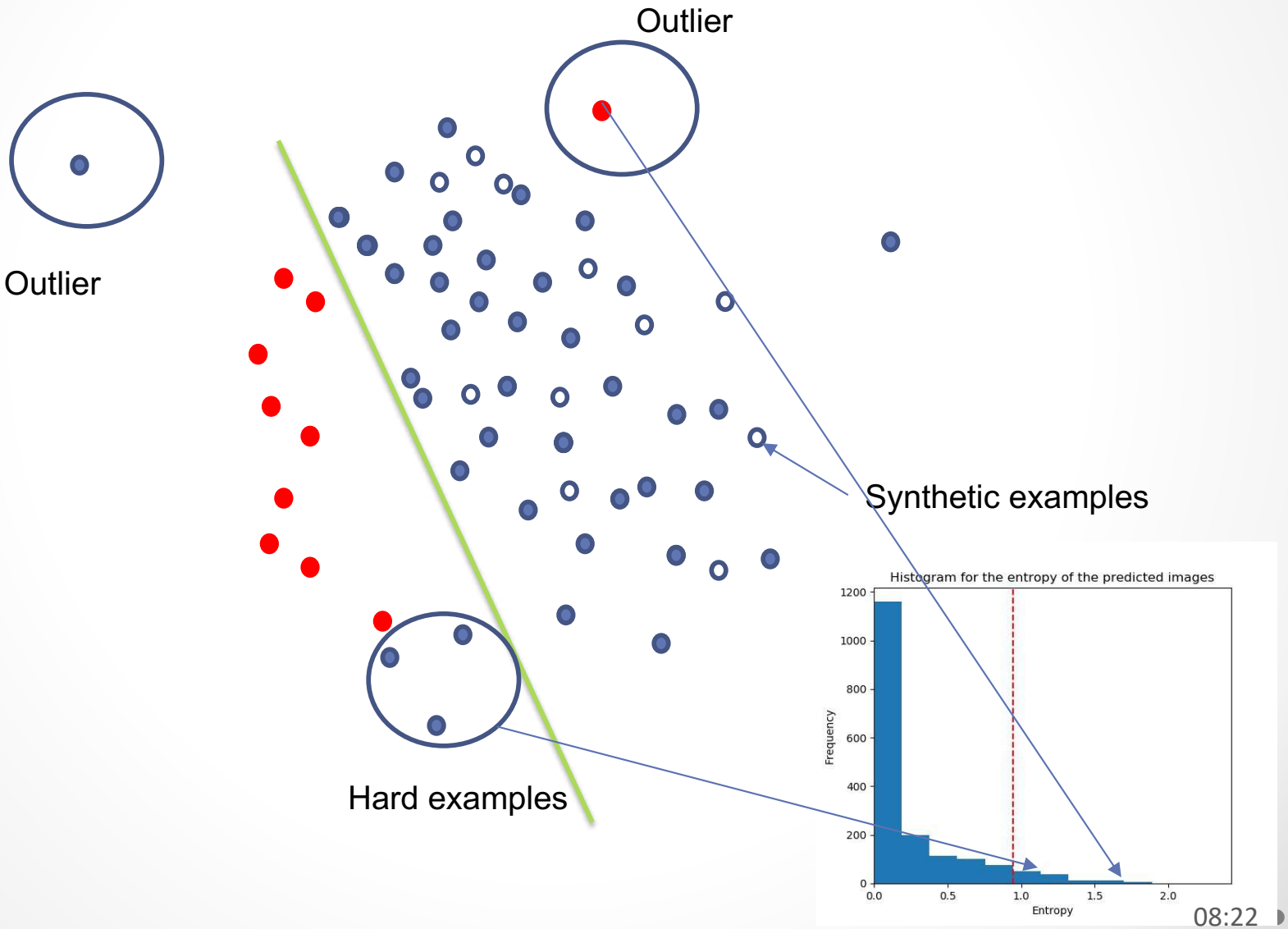


Use the data augmentation applied class-conditionally to improve the results in terms of accuracy and also to reduce the overall epistemic uncertainty.

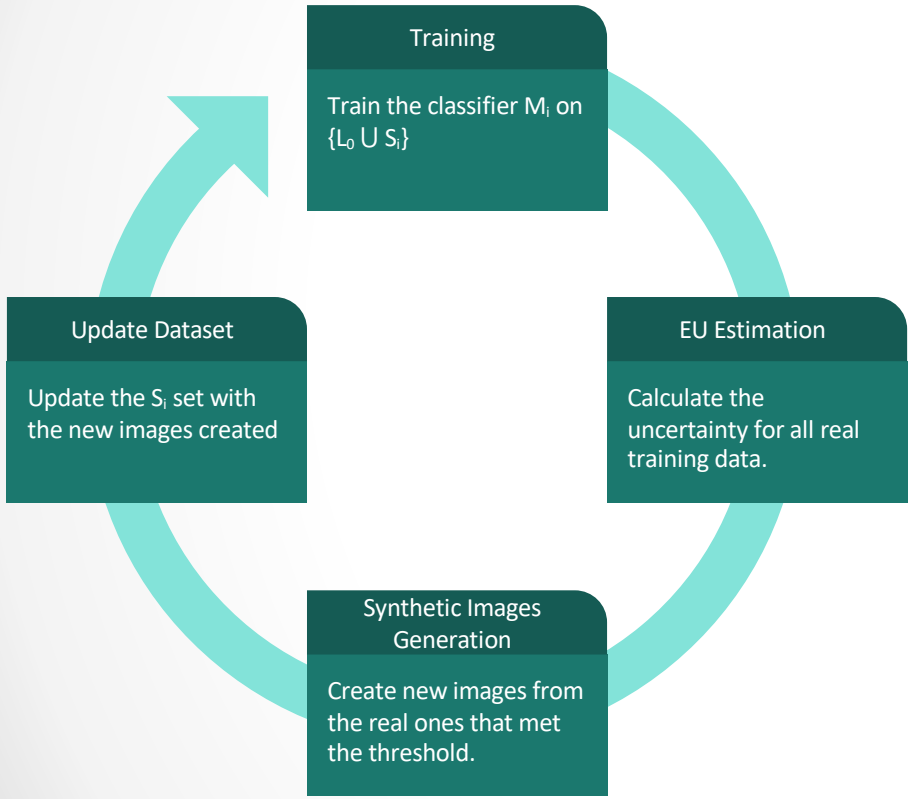
During the prediction phase, the same image is fed to the CNN several times to calculate the epistemic uncertainty given by the model for that image

E. Aguilar, and P. Radeva. "Class-conditional Data Augmentation Applied to Image Classification." International Conference on Computer Analysis of Images and Patterns (CAIP),2019.

Can there be data that require more attention than others?



UDA Procedure



Algorithm 1: UDA procedure

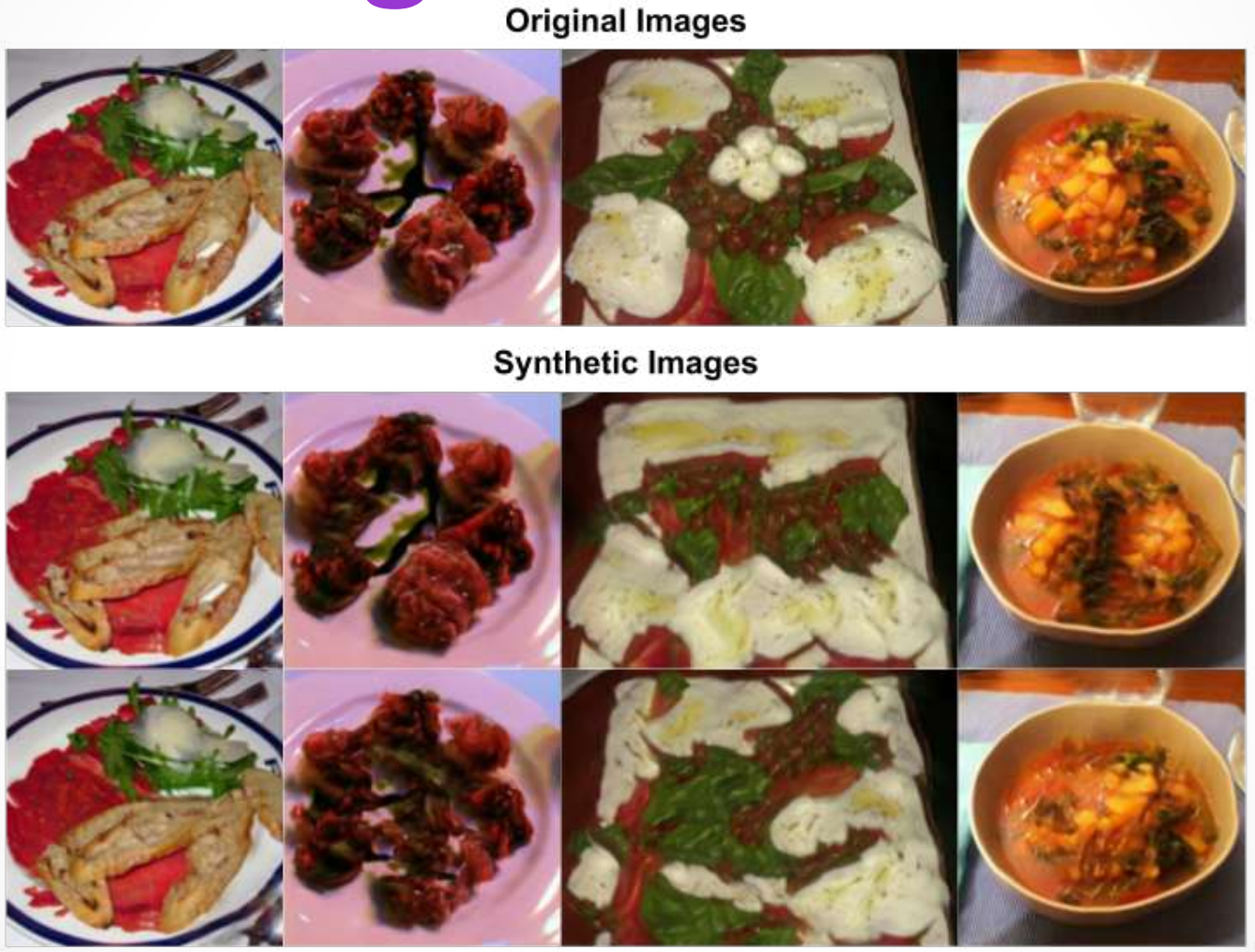
input: Labeled Data L_0, Synthetic Data $S_i \leftarrow \emptyset$, Lower Threshold T_l, Upper Threshold T_u, Generator G;

while *we are not satisfied with the performance* **do**
  Train the classifier M_i on $\{L_0 \cup S_i\}$;
  for $x_j \in L_0$ **do**
    Calculate the $EU(x_j)$;
    if $T_l < EU(x_j) < T_u$ **then**
      Create the synthetic image x_j^* with G;
      Update $S_i \leftarrow S_i \cup \{x_j^*\}$;
    end
  end
end

E. Aguilar et al., Uncertainty-aware Data Augmentation for Food Recognition. ICPR (2020)

Validation

Use Uncertainty for Data Augmentation



Synthetic image generated on the selected images from the training set

Use Uncertainty for Data Augmentation

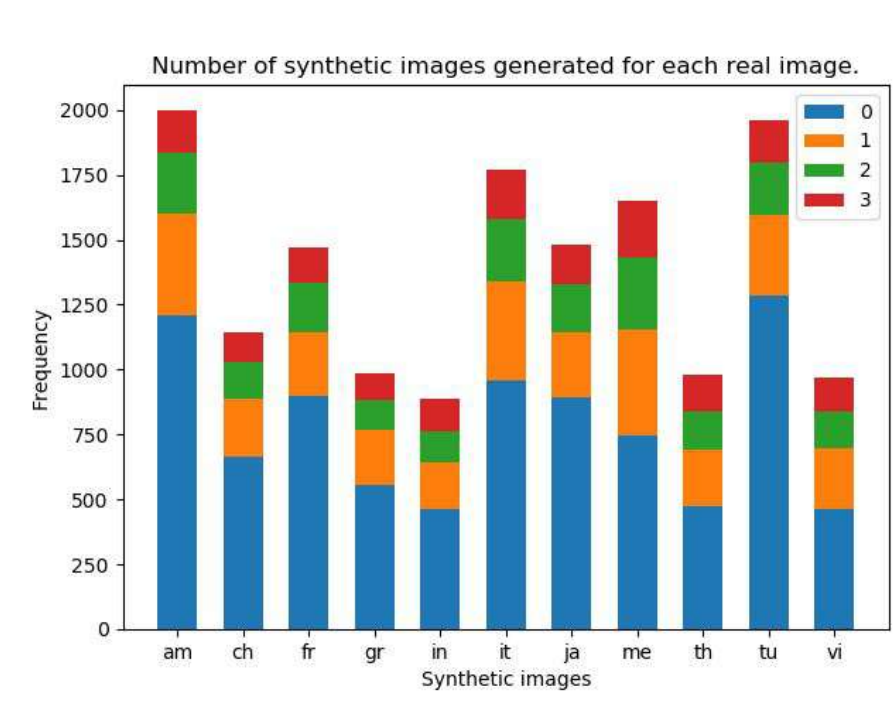
Model	*Acc*	*NEU*
ResNet50	61.00%	30.22%
ResNet50+DA	65.02%	33.55%
ResNet50+DA+A	64.65%	36.53%
Proposed method	65.54%	33,51%

Results on UECFOOD-256 in terms of Acc and NEU for the models trained with different data augmentation techniques.

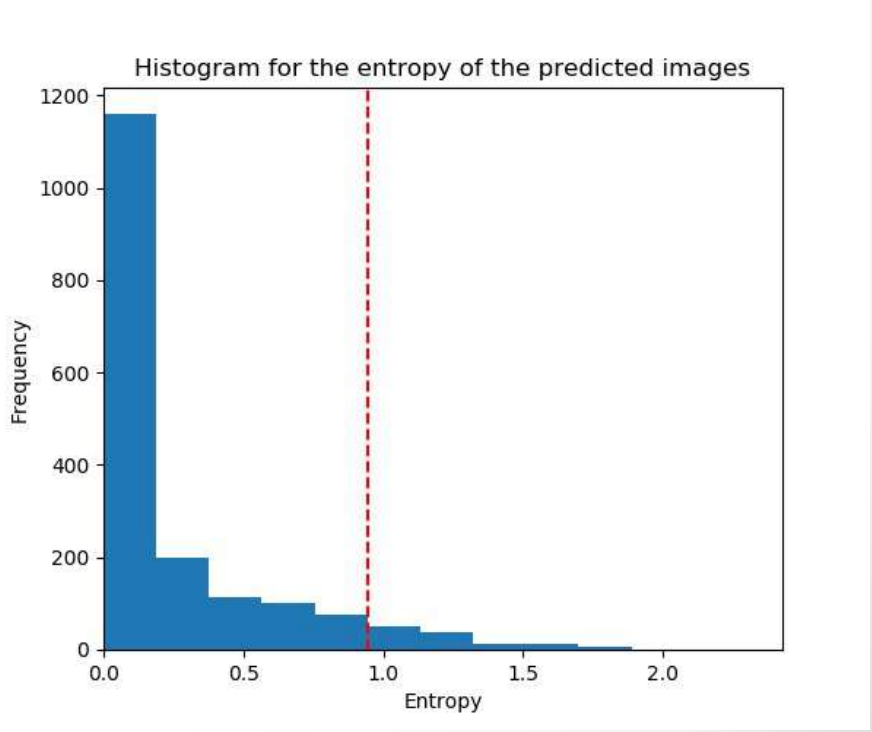
Results on Food-101 in terms of Acc and NEU for the models trained with different data augmentation techniques.

Model	*Acc*	*NEU*
ResNet50	77.66%	19.85%
ResNet50+DA	82.65%	27.35%
ResNet50+DA+A	82.54%	29.45%
Proposed method	82.82%	26.25%

Use Uncertainty for Data Augmentation



Number of synthetic images generated after the third training cycle.



Histogram for the entropy of the predicted images

Results

TABLE II

RESULTS OBTAINED ON THE TEST SETS IN TERMS OF R_{micro}.

Method	American	Chinese	French	Greek	Indian	Italian	Japanese	Mexican	Thai	Turkish	Vietnamese
DenseNet169_DO (S_1)	87,12%	91,83%	93,56%	91,10%	93,68%	86,26%	93,62%	83,07%	84,95%	93,81%	90,67%
DenseNet169_DO (S_2)	88,89%	92,61%	94,24%	93,19%	93,68%	86,49%	93,62%	85,62%	84,95%	93,58%	91,19%
DenseNet169_DO (S_3)	**89,39%**	93,00%	93,22%	92,67%	**93,68%**	87,68%	**94,33%**	86,26%	**86,02%**	**94,03%**	**91,71%**
DenseNet169_DO (S_4)	88,89%	**94,16%**	**94,92%**	**93,72%**	93,16%	**87,91%**	**94,33%**	**86,90%**	**86,02%**	**94,03%**	91,19%

TABLE III

RESULTS OBTAINED ON THE TEST SETS IN TERMS OF R_{macro}.

Method	American	Chinese	French	Greek	Indian	Italian	Japanese	Mexican	Thai	Turkish	Vietnamese
DenseNet169_DO (S_1)	86,93%	91,84%	94,02%	89,60%	93,53%	87,64%	92,75%	82,31%	82,69%	93,68%	90,21%
DenseNet169_DO (S_2)	88,77%	92,08%	93,96%	92,42%	93,51%	87,19%	93,19%	85,55%	83,84%	93,55%	91,31%
DenseNet169_DO (S_3)	**89,26%**	92,69%	93,24%	91,59%	**93,77%**	88,83%	93,81%	86,73%	83,99%	93,96%	**91,50%**
DenseNet169_DO (S_4)	88,98%	**92,89%**	**95,78%**	**92,99%**	92,79%	**89,42%**	**94,04%**	**86,97%**	**84,08%**	**94,00%**	91,18%

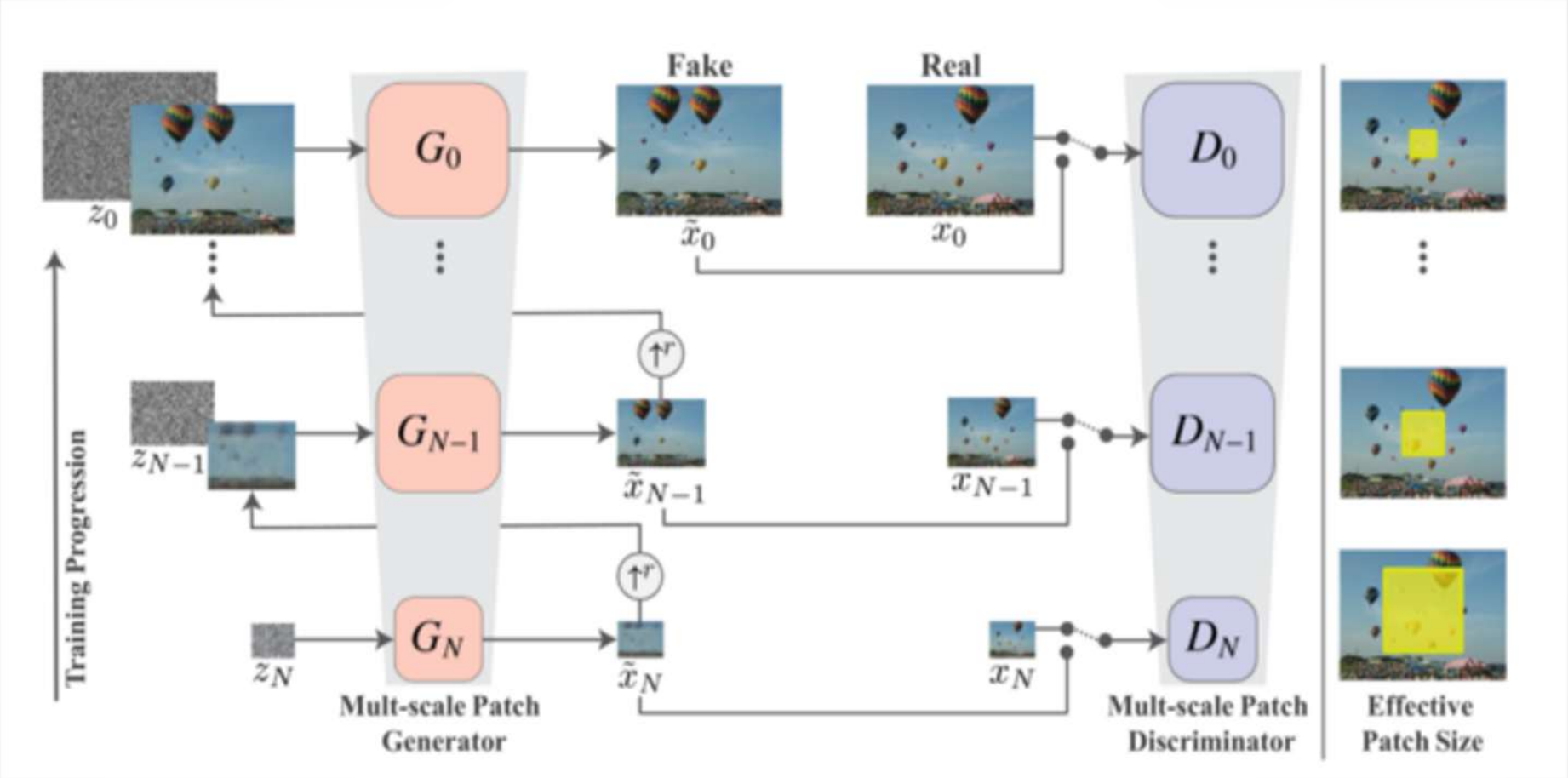
S1: Training with real images without applying any type of data augmentation.
S2: Training with standard online data augmentation, such as random crops and horizontal flips, applied on real images only.
S3: Training with standard online data augmentation, applied on a dataset consisting of both synthetic images (one for each real image) and real images.
S4: Training with standard online data augmentation, applied on a dataset generated by our UDA method.

Conclusions (II)

- Uncertainty modeling is a hot topic with many open questions and challenges but interesting application to DL
- Advantages conclude better estimates of uncertainty, automatic ways of learning structure and avoiding overfitting
- A new framework is proposed for data augmentation based on epistemic uncertainty applying GANs
- A high impact of food analysis is expected from point of view of:
  - Scientific questions, but also
  - Real world applications, specially important for the society.


Thank you!
Petia I. Radeva
petia.ivanova@ub.edu

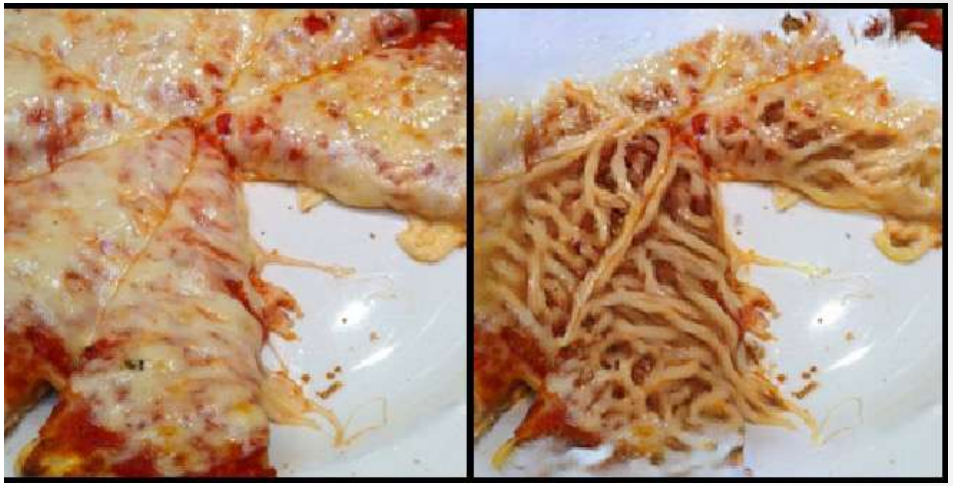
SINGAN



Pyramid of GANs, where both training and inference are done in a coarse-to-fine fashion.

At each scale, $\mathbf{G}_n$ learns to generate image samples in which all the overlapping patches cannot be distinguished from the patches in the down-sampled training image.

Using AttentionGan



Chen, Xinyuan, et al. "Attention-GAN for object transfiguration in wild images." Proceedings of the European Conference on Computer Vision (ECCV). 2018.