# MTL870: Meta Learning Major (Semester II, 2021-22)

Sweta Mahajan(AIZ218356)

April 10, 2022

## 1 Question1

### 1.1 CAVIA: Fast Context Adaptation via Meta-Learning

CAVIA is a gradient based meta learning algorithm which is an extension of the Model Agnostic Meta Learning (MAML) algorithm. It has two sets of parameters. In the inner loop only the task related parameters $\phi$ are updated which allows it to optimise the task specific parameters efficiently which is called the adaptation phase. In the outer loop, across-task parameters are updated which are meta-trained and are shared across tasks which is called the update phase.

### 1.2 Adaptive Risk Minimisation: Learning to Adapt to Domain Shift

The authors deploy meta-learning technique to deal with domain shift or train time and test time distribution shifts. Although meta learning is primarily for few shot recognition problems, the author here are extending meta-learning algorithms to carry out test time adaptation to tackle distribution shift. The complete module is divided into two sub-modules:

Algorithm:

- Initialize parameters $\theta$ and $\phi$
- The first submodule is the adaptation model $h(\theta, x; \phi) : \Phi \times \mathscr{X}^K \rightarrow \Theta$, which is parametrized by $\phi$. This model tries to adapt the $\theta$ using the unlabelled data to produce the adapted parameters $\theta^{'}$. This is equivalent to the adaptation phase in the meta learning. Hence this is called the meta-learner.

- The second module is $g(x;\theta) : \mathscr{X} \rightarrow \mathscr{Y}$, which is parametrized by $\theta \in \Theta$. It uses the adapted parameters from the network $h$ and predicts label $y$ given $x$. Then the prediction error is backpropagated to jointly update the parameters $\theta$ and $\phi$. This is the update phase and hence called the learner.

- Parametrs $\theta$ can be interpreted as task embedding which decides how the predictive model behaves.

## 1.3 Conditional Neural Processes

CNP has two neural networks. The first one takes in the $(X_train, y_train)$ and outputs average class embeddings. This network performs the adaptation phase and called the meta-learner. The second one takes these embeddings and concatenates the test data to predict its label. This step is the update phase, also called the learner.

## 1.4 Analysis

Similarities:

- These are all models to approach meta learning in a particular way. Instead of assimilating the context dependent information and context independent information in the same set of parameters as in MAML, they divide the parameters to deal with context dependent and independent information separately. CNP and ARM do it by deploying two sets of neural networks whereas CAVIA can deploy it in a single neural network.

Differences:

- CNP and CAVIA are few shot whereas ARM is zero shot.

- CNP and CAVIA do adaptation using labelled data but ARM does adaptation using unlabelled data.

- ARM and CNP are model based whereas CAVIA is optimization based model.

- CNP and CAVIA are u sing meta learning to improve accuracy in classification tasks on unknown but related tasks whereas ARM tries to deploy meta learning to tackle distribution shifts.

# 2 Question 2

## 2.1 Concept Learning with Energy Based Models (CLEBM)

Here the aim of the paper is two fold, identify and generate experiences as concepts. They define an energy function $E(x,a,w)$, where 'x' is entities of the concept , 'a' is the attention mask, 'w' is the embedding vector of the concept. For example, suppose an image consists of multiple dots, 4 of which form a square. In this, concept is the square, 'x' is the co-ordinates of all the dots in the image, 'a' is the attention mask over the dots specifying which dots form the square, 'w' is the embedding of the square (similar to the latent representation in autoencoder). The energy function $E(x,a,w) = 0$ when the entities of the concept 'x' under attention mask 'a' satisfies concept 'w' i.e, the dots as specified by the attention mask form a square.

Example: Identification of a Square concept:

- IDENTIFICATION: Given a " entities of a concept x " and concept embedding of the square 'w', we have to find the attention mask 'a' which tells which dots form the square. Since $E(x,a,w)$ is a differentiable function of 'x','w' and 'a', we find the optimal value by running steps of gradient descent wrt 'a'.
- GENERATION: Given a noisy version of of 'x', say $x_0$, we try to generate the denoised version of it. Again we do so by using gradient descent of the energy function $E(x,a,w)$ wrt 'x'.

## 2.2 Meta-Learning Deep Energy Based Memory Models (MLDEBMM)

The aim of this paper is to retrieve denoised version of a noisy input. It defines an energy function $E(x,\theta)$ which should take minimum values at valid input images 'x' and high values at noisy inputs. The method of updating the parameter $\theta$ is called 'writing' and the process of converting the noisy input to denoised one is called 'reading'. Conventionally, autoencoders take lot of time for 'reading' i.e., parameter update. The aim of this paper is to carry out fast parameters updates. So, they formulate this as a meta-learning problem where they try to find a good initialisation to the model parameter $\theta$ using meta-learning techniques.

## 2.3 Analysis

Similarities:

- Both the paper model the energy as a differentiable function of the inputs and the parameters. To find the optimal values of the parameter where the energy is minimum, it uses gradient descent. These are both variations of the gradient descent based energy models.
- There are many facets to both the papers. But one common problem they are dealing is to produce the original image given the distorted version of it.

Differences:

- The CLEBM model works on symbolic forms whereas MLDEBMM works on images.
- The CLEBM model uses conventional gradient descent to find the denoised image whereas MLDEBMM deploys it as a meta-learning problem.

# 3 Question3

- All the four algorithms are model based meta-RL.
- For optimization, in MAML, the reward function to be continuous whereas other three models can work with sparse rewards.
- Since MAML can not work with sparse rewards, it does not do well on scenarios which get rewards at the end.
- For RL2 the reward function should be bounded whereas for other it is not the case.
- At each step the agent gets the information of rewards along with the state-action pair in the three models except MAML.

# 4 Question 4

## 4.1 Application 1: Semi-supervised Meta-learning with Disentanglement for Domain-generalised Medical Image Segmentation[3]

Medical mage segmentation is crucial for disease diagnosis. The model performance decreases due to domain variation i.e., scanners and scan acquisition settings which affect image characteristics like brightness and contrast and variation in patient population which affects the underlying anatomy and pathology due to various cultural and ethnic factors. Variations in images due to scanners and scan

acquisition settings are considered as coming from one domain and the model should be insensitive to this as these do not imply any pathological changes in the body. Variation in patient population is considered as another domain and we want our model to be sensitive to that as it might indicate presence of abnormality in the body. The image is passed through three neural networks N1, N2, N3. N1 encodes a common representation 's' and N2 encodes the domain specific information 'd' and N3 gives feature representation 'Z'. Then a decoder combines these 3 representations to reconstruct the image. The decodes does it in a disentangled way while encouraging Z to encode anatomical information which is helpful for segmentation and encourage 's and d' to encode the domain specific information.

## 4.2 Disentangling by Factorising [2]

Learning latent representation of images which are semantically meaningful is very much crucial for tasks such as supervised learning, reinforcement learning, transfer learning and zero-shot learning. Learning these features in a disentangled way is crucial for synthetic image formation or photo shopping or film making. The authors propose Factor VAE to do the same. It augments the VAE loss with another loss called Total Correlation i.e.,$(KL(q(z)||q_bar(z)))$. The VAE loss trie to bring down the reconstruction error loss and the total correlation term tries to disentangle the hidden factors. Factor VAE implements a discriminator neural network to evaluate the Total Correlation term using the density ratio trick. The authors have trained the VAE and discriminator jointly to achieve the task.

## 4.3 Disentangled Feature Representation for Few-shot Image Classification[1]

Few-shot image classification is challenging. The model has to learn discriminative feature representation which can be generalised across tasks. In practice, many of the excursive features like style, domain and background are irrelevant of the class. The subtle traits are essential to differentiate objects of fine grained class but are hardly preserved in the learned embedding as they are dominated by the style and domain information. The paper selectively extracts class specific information for each task using Disentangled Feature Representation framework (DFR) while maintaining model generalisation. DFR has two branches: classification branch which extracts class specific information and variational branch which encodes class-irrelevant information. Together these two representations are able reconstruct the image. The two set of features are disentangled.

# 5 Question5

- The important question is what is the size of the model and if it has the capacity to model the given task at hand. If the task was complex and a small network is used to model it, then the model would underfit the data and the training error would be large but in the question it is given that even when the sample size is 2 million the training loss goes to zero. From this we infer that the size of the network is sufficient.
- Next, lets try to see if the task at hand is complex or not. For a complex task, if the training sample is very small, then the model parameters take unusually high positive and high negative values to fit the training data. But here we see when the number of samples is 200, the training error goes to 0 and the weights are bounded between -1 and 1. Together it implies that the task is simple.
- It is given that the train and test time data distribution is same and since this is neither a case of overfitting or underfitting, in all the three cases the test loss will be near zero.

# 6 Question6

## 6.1 Part-A

I would suggest going ahead with Adaptive Risk Minimization model. We can consider each geographic location as a separate distribution. Since we have the purchase history of past customers, we have the $(X_train, y_train)$. Here $X_train$ represents the product name and $y_train$ represents the class label i.e., 1 representing the customer bought the product and 0 if not. At test time, we have $X_test$ which is the unlabelled product and we want to estimate whether the customer from the particular geographic location will buy the product or not. Similar to ARM, we can utilize the unlabelled test data to to test time adaptation and prediction.

## 6.2 Part-B

I would suggest to use Neural Algorithmic Reasoning model to extract the relevant information. Since, there is a particular way that the ingredients are listed on a product. The nutritional facts are always printed in a table format below the heading "Nutritional Information" and the ingredients are written below the heading "Ingredients". For NAR, we need to have certain invariances across the

images and these headlines offer us that. Hence with the help of few shot images and noisy clone, I believe we will be able to extract the relevant information.

### 6.3 Part-C

I would suggest to perform it under the aegis of Meta-RL. A fixed budget is given and we want to maximise customer acquisition. In RL, we have the reward function which is bounded from below and above. So, in case of no customer acquisition, the fixed budget is the lower bound and we can only acquise finite number of customers (since the population is finite), the reward function is also upper bounded. So, this is the realm of RL. Since we have to adapt to rapidly changing customer behaviour over time, we have to resort to meta-RL. We can use RL2 as our model since it performs better than MAML RL.

## References


[1] CHENG, H., WANG, Y., LI, H., KOT, A. C., AND WEN, B. Disentangled feature representation for few-shot image classification. *arXiv preprint arXiv:2109.12548* (2021).

[2] KIM, H., AND MNIH, A. Disentangling by factorising, 2018.

[3] LIU, X., THERMOS, S., O'NEIL, A., AND TSAFTARIS, S. A. Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2021), Springer, pp. 307–317.